Annals of Computer Science and Information Systems
Volume 3

Position Papers of the 2014 Federated Conference on Computer Science and Information Systems

September 7–10, 2014. Warsaw, Poland

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki
(eds.)

Annals of Computer Science and Information Systems, Volume 3

Position Papers of the 2014 Federated Conference on Computer Science and Information Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki
(eds.)

2014, Warszawa,
Polskie Towarzystwo
Informatyczne

Annals of Computer Science and Information Systems, Volume 3

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems

USB: ISBN 978-83-60810-59-0
WEB: ISBN 978-83-60810-57-6

ISSN 2300-5963
DOI 10.15439/978-83-60810-57-6

Al. Solidarności 82A m. 5
01-003 Warsaw
Poland

Contact: secretariat@fedcsis.org
`http://annals-csis.org/`

Cover:
Jana Waleria Denisiuk,
Elbląg, Poland

Also in this series:

Volume 1: Position Papers of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS), **WEB: ISBN 978-83-60810-55-2, USB: ISBN 978-83-60810-56-9**
Volume 2: Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, **WEB: ISBN 978-83-60810-58-3, USB: ISBN 978-83-60810-57-6, ART: ISBN 978-83-60810-61-3**
Volume 4: Proceedings of the E2LP Workshop, **WEB: ISBN 978-83-60810-64-4, USB: ISBN 978-83-60810-63-7**

DEAR Reader, it is our pleasure to present to you Position Papers of the 2014 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Warsaw, Poland, on September 7–10, 2014. This is the second year when position papers have been introduced as a separate category of contributions. They represent emerging research papers and challenge papers. The former present preliminary research results from work-in-progress based on sound scientific approach but presenting work not completely validated as yet. The latter propose and describe research challenges in theory or practice of computer science and information systems.

FedCSIS 2014 was organized by the Polish Information Processing Society (Mazowsze Chapter), Warsaw University of Technology, Wrocław University of Economics and Systems Research Institute Polish Academy of Sciences. It was organized in technical cooperation with: IEEE Computer Society, IEEE Region 8, Computer Society Chapter Poland, Gdańsk Computer Society Chapter, Poland, Polish Chapter of the IEEE Computational Intelligence Society (CIS), ACM Special Interest Group on Applied Computing, International Federation for Information Processing (IFIP), European Alliance for Innovation (EAI), Łódź ACM Chapter, Informatics Europe, Asociación de Técnicos de Informática, Committee of the Computer Science of the Polish Academy of Sciences, Polish Society for Business Informatics, Polish Chamber of Information Technology and Telecommunications, Polish Chamber of Commerce for High Technology, Mazovia Cluster ICT and Eastern Cluster ICT Poland. Furthermore, the 9th International Symposium Advances in Artificial Intelligence and Applications (AAIA'14) was organized in technical cooperation with: International Rough Set Society, International Fuzzy Systems Association and Polish Neural Networks Society.

The following FedCSIS 2014 events included position papers in their program:

- **AAIA'14—9th International Symposium Advances in Artificial Intelligence and Applications**
  - AIMA'14 – 4th International Workshop on Artificial Intelligence in Medical Applications
  - CEIM'14 – 1st Complex Events and Information Modelling
  - WCO'14 – 7th Workshop on Computational Optimization
- **CSS—Computer Science & Systems**
  - CANA'14 – 7th Computer Aspects of Numerical Algorithms
  - ScoDiS-LaSCoG'14 – 2rd Workshop on Scalable Computing in Distributed Systems and 7th Workshop on Large Scale Computations on Grids
- **ECRM—Education, Curricula & Research Methods**
  - ISEC'14 – Information Systems Education & Curricula Workshop
- **iNetSApp—Innovative Network Systems and Applications**
  - EAIS'14 – Emerging Aspects in Information Security
  - SoFAST-WS'14 – 3rd International Symposium on Frontiers in Network Applications, Network Systems and Web Services
  - WSN'14 – 3rd International Conference on Wireless Sensor Networks
- **IT4MBS—Information Technology for Management, Business & Society**
  - ABICT'14 – 5th International Workshop on Advances in Business ICT
  - AITM'14 – 12th Conference on Advanced Information Technologies for Management
  - ISM'14 – 9th Conference on Information Systems Management
  - IT4L'14 – 3rd Workshop on Information Technologies for Logistics
  - KAM&AI4KM'14 – 20th Conference on Knowledge Acquisition and Management and 2nd Workshop on Artificial Intelligence for Knowledge Management
- **SSD&A—Software Systems Development & Applications**
  - MDASD'14 – 3rd Workshop on Model Driven Approaches in System Development

Each of these events had its own Organizing and Program Committee. We would like to express our warmest gratitude to members of all of them for their hard work attracting and later refereeing 430 submissions.

FedCSIS 2014 was organized under the auspices of Prof. Lena Kolarska-Bobińska, Minister of Science and Higher Education, Dr Rafał Trzaskowski, Minister of Administration and Digitization, Prof. Michał Kleiber, President of the Polish Academy of Sciences, Major General Wiesław Leśniakiewicz, Chief Commandant of the State Fire Service, Prof. Hanna Gronkiewicz-Waltz, Mayor of the Capital City of Warsaw, Prof. Jan Szmidt, Rector of Warsaw University of Technology, Gen. Prof. Zygmunt Mierczyk, Rector of Military Technical Academy, and Prof. Andrzej Gospodarowicz, Rector of Wroclaw University of Economics.

FedCSIS was sponsored by Ministry of Science and Higher Eduction, Intel, Orange Polska S.A. and Samsung.

Maria Ganzha, *Co-Chair of the FedCSIS Conference Series*
Systems Research Institute Polish Academy of Sciences, Warsaw, Poland, and Gdańsk University, Gdańsk, Poland
Leszek Maciaszek, *Co-Chair of the FedCSIS Conference Series*
Wrocław University of Economics, Wrocław, Poland and Macquarie University, Sydney, Australia
Marcin Paprzycki, *Co-Chair of the FedCSIS Conference Series*
Systems Research Institute Polish Academy of Sciences, and Warsaw and Management Academy, Warsaw, Poland

Annals of Computer Science and Information Systems, Volume 3

Position Papers of the 2014 Federated Conference on Computer Science and Information Systems (FedCSIS)

September 7–10, 2014. Warsaw, Poland

TABLE OF CONTENTS

9TH International Symposium Advances in Artificial Intelligence and Applications

Call For Papers 1

Full Rough Sets 3
Ray-Ming Chen

Approaches to the Intelligent Subject Search 13
Vladimir Ivanov, Boris Palyukh, Alexander Sotnikov

Solving Graph Coloring Problem with Parallel Evolutionary Algorithms in a Mesh Model 21
Zbigniew Kokosiński, Piotr Domagała

YUV vs RGB—Choosing a Color Space for Human-Machine Interaction 29
Michal Podpora, Grzegorz Paweł Korbaś, Aleksandra Kawala-Janik

4TH International Workshop on Artificial Intelligence in Medical Applications

Call For Papers 35

Fuzzy Decision Tree Based Classification of Psychometric Data 37
Krzysztof Pancerz, Vitaly Levashenko, Elena Zaitseva, Jerzy Gomula

1ST Complex Events and Information Modelling

Call For Papers 43

Evaluation of a Heat Release Rate based on Massively Generated Simulations and Machine Learning Approach 45
Mateusz Fliszkiewicz, Adam Krasuski, Karol Krenski

7TH Workshop on Computational Optimization

Call For Papers 53

A probabilistic traveling salesman problem: a survey 55
Abir Henchiri, Monia Bellalouna, Walid Khaznaji

Transdimensional sequential Monte Carlo for hidden Markov models using variational Bayes - SMCVB 61
Clare McGrory, Daniel Ahfock

Computer Science & Systems

Call For Papers 67

7TH Computer Aspects of Numerical Algorithms

Call For Papers 69

Attractor Selection mechanism Simulink model 71
Dmitrijs Finaskins, Gunars Lauks

Parallel implementation of linear repetitive processes identification using subspace algorithms 79
Dominik Kujawa

Pollutant's Dispersion in the Atmosphere. Parallel Models and Applications 85
Marcin Majer, Michal Podpora

2ND Workshop on Scalable Computing in Distributed Systems and 7TH Workshop on Large Scale Computations on Grids

Call For Papers 91

A self-stabilizing algorithm for locating the center of Cartesian product of K2 and maximal outerplanar graphs 93
Halina Bielak, Michał Pańczyk

Identity Providers-as-a-Service built as Cloud-of-Clouds: challenges and opportunities 101
Diego Kreutz, Eduardo Feitosa

3RD Information Systems Education & Curricula Workshop

Call For Papers 109

Proprietary versus Open Source Software in Support of Learning in Computer Science 111
Antoine Melki

Innovative Network Systems and Applications

1ST Workshop on Emerging Aspects in Information Security

Call For Papers 117

ICT Security Risk Management: Economic Perspectives 119
Gebhard Geiger

Achieving Software Security for Measuring Instruments under Legal Control 123
Daniel Peters, Ulrich Grottker, Florian Thiel, Michael Peter, Jean-Pierre Seifert

Measuring Security: A Challenge for the Generation 131
Janusz Zalewski, Steven Drager, William McKeever, Andrew J. Kornecki

3RD International Symposium on Frontiers in Network Applications, Network Systems and Web Services

Call For Papers 141

SDN Architecture Impact on Network Security 143
Krzysztof Cabaj, Jacek Wytrębowicz, Sławomir Kukliński, Paweł Radziszewski, Khoa Truong Dinh

The Monte Carlo analysis of the media access time distribution in 802.11n MAC layer 149
Iwona Dolińska, Antoni Masiukiewicz, Grzegorz Rządkowski

Non-Destructive Testing in Complexes Cabling Networks using Time Domain Reflectometry and Particle Swarm Optimization 157
Hamza Boudjefdjouf, Rabia Mehasni, Houssem Bouchekara, Antonio Orlandi, Francesco De Paulis

Dynamic Autonomic Network Management: Evaluating the Architectural Challenges of Autonomic Management for Mobile Ubiquitous Access 165
Clifford C. L. Sibanda, Olabisi E. Falowo

Send It Safe – A Novel Application for Secure Key Exchange Using Telecommunications Open Middleware APIs 171
Piotr Wawrzyniak, Łukasz Wronkowski, Damian Kuniszewski, Adam Cackowski, Paweł Czapliński, Karol Szymański

3RD International Conference on Wireless Sensor Networks

Call For Papers 175

A Design of Application Based Wireless Sensor Node 177
Sham P Nayse, Mohammad Atique

The concept of authentication in WSNs using TPM 183
Janusz Furtak, Jan Chudzikiewicz

Toward implementing an efficient gateway for wireless sensor networks 191
Gundars Miezitis, Romans Taranovs

Information Technology for Management, Business & Society

Call For Papers 197

5TH International Workshop on Advances in Business ICT

Call For Papers 199

Enhancement of the ValueSec Risk Management Model 201
Andrzej Bialas

12TH Conference on Advanced Information Technologies for Management

Call For Papers 209

Service Design and Distributed System Reliability in Intelligence Information System Based on Service-Oriented Architecture 211
Jugoslav Achkoski, Vladimir Trajkovik

A Fractal Measure for Comparing the Work Effort of Human and Artificial Agents Performing Management Functions 219
Matthew Gladden

9$^{\text{TH}}$ Conference on Information Systems Management

Call For Papers **227**

Functionality Assessment and Requirements of the IT System in an Education Institution **229**
Ljijana Ruzic-Dimitrijevic, Milorad Stevic

3$^{\text{RD}}$ Workshop on Information Technologies for Logistics

Call For Papers **235**

Cloud-based Cooperation of Logistics Service Providers in Logistics Cluster Organisations **237**
Jan Oberländer, Bogdan Franczyk

Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management **245**
Silva Robak, Bogdan Franczyk, Marcin Robak

20$^{\text{TH}}$ 20th Conference on Knowledge Acquisition and Management & 2$^{\text{ND}}$ Workshop on Artificial Intelligence for Knowledge Management

Call For Papers **251**

Managing Intellectual Capital in Knowledge Economy **253**
Eunika Mercier-Laurent

Music Information Retrieval as a Key Framework to Explore Legal Issues Linked to Personal Data Computation **259**
Francis Rousseaux, Pierre Sorel, Marc Danger

Tool dilemmas of innovation **265**
Halina Tańska, Jolanta Sala

Information Models and Methods of the University's Scientific Knowledge Life Cycle Support **271**
Gulnaz Zhomartkyzy, Tatiana Balova, Marek Milosz

Software Systems Development & Applications

Call For Papers **279**

3$^{\text{RD}}$ Workshop on Model Driven Approaches in System Development

Call For Papers **281**

Function-oriented Inclusive Design **283**
Markus Modzelewski, Michael Lawo

Author Index **291**

9th International Symposium Advances in Artificial Intelligence and Applications

THE AAIA'14 will bring researchers, developers, practitioners, and users to present their latest research, results, and ideas in all areas of artificial intelligence. We hope that theory and successful applications presented at the AAIA'14 will be of interest to researchers and practitioners who want to know about both theoretical advances and latest applied developments in Artificial Intelligence. As such AAIA'14 will provide a forum for the exchange of ideas between theoreticians and practitioners to address the important issues.

Topics

Papers related to theories, methodologies, and applications in science and technology in this theme are especially solicited. Topics covering industrial issues/applications and academic research are included, but not limited to:

- Knowledge Management
- Decision Support Systems
- Approximate Reasoning
- Fuzzy Modeling and Control
- Data Mining
- Web Mining
- Machine Learning
- Combining Multiple Knowledge Sources in an Integrated Intelligent System
- Neural Networks
- Evolutionary Computation
- Nature Inspired Methods
- Natural Language Processing
- Image Processing and Interpreting
- Applications in Bioinformatics
- Hybrid Intelligent Systems
- Granular Computing
- Architectures of Intelligent Systems
- Robotics
- Real-world Applications of Intelligent Systems
- Rough Sets

Professor Zdzislaw Pawlak Best Paper Awards

We are proud to announce that we will continue the tradition started during the AAIA'06 Symposium and award two "Professor Zdzislaw Pawlak Best Paper Awards" for contributions which are outstanding in their scientific quality. The two award categories are:

- Best Student Paper - for graduate or PhD students. Papers qualifying for this award must be marked as "Student full paper" to be eligible for consideration.
- Best Paper Award for the authors of the best paper appearing at the Symposium.

Candidates for the awards can come from AAiA and all workshops organized within its framework (i.e. AIMA, ASIR, CEIM, TAIE, WCO)

In addition to a certificate, each award carries a prize of 300 EUR provided by the Mazowsze Chapter of the Polish Information Processing Society.

IFSA Award for Young Scientist

During the Advances in Artificial Intelligence and Applications (AAIA) Symposium, the International Fuzzy Systems Association (IFSA) Best Paper Award for Young Scientist, will be presented.

Candidates for the awards can come from AAiA and all workshops organized within its framework (i.e. AIMA, ASIR, CEIM, TAIE, WCO)

Founding Chairs

Kwaśnicka, Halina, Wroclaw University of Technology, Poland

Markowska-Kaczmar, Urszula, Wroclaw University of Technology, Poland

Event Chairs

Krawczyk, Bartosz, Wroclaw University of Technology, Poland

Slezak, Dominik, University of Warsaw & Infobright Inc., Poland

Program Committee

Bartkowiak, Anna, Wroclaw University, Poland

Bazan, Jan, University of Rzeszów, Poland

Bodyanskiy, Yevgeniy, Kharkiv National University of Radio Electronics, Ukraine

Budnik, Mateusz, University of Grenoble, France

Błaszczyński, Jerzy, Poznan University of Technology, Poland

Cyganek, Boguslaw, AGH University of Science and Technology, Poland

Czarnowski, Ireneusz, Gdynia Maritime University, Poland

Herrera, Francisco, University of Granada, Spain

Hippe, Zdzislaw, University of Information Technology and Management in Rzeszow, Poland

Jaromczyk, Jerzy W., University of Kentucky, United States

Korbicz, Józef, University of Zielona Gora, Poland

Kwaśnicka, Halina, Wroclaw University of Technology

Marek, Victor, University of Kentucky, United States

Markowska-Kaczmar, Urszula, Wroclaw University of Technology

Mercier-Laurent, Eunika, IAE Lyon3, France

Miroslaw, Lukasz, University of Applied Science Rapperswil & Wroclaw University of Technology, Switzerland

Myszkowski, Pawel, Wroclaw University of Technology, Poland

Ngan, Ben C. K., The Pennsylvania State University, United States

Nguyen, Hung Son, University of Warsaw, Poland

Porta, Marco, University of Pavia, Italy

Ramanna, Sheela, University of Winnipeg, Canada

Ras, Zbigniew, University of North Carolina at Charlotte, United States

Sas, Jerzy, Wroclaw University of Technology, Poland

Snasel, Vaclav, VSB -Technical University of Ostrava, Czech Republic

Szczech, Izabela, Poznan University of Technology, Poland

Szczuka, Marcin, The University of Warsaw, Poland

Szpakowicz, Stan, University of Ottawa, Canada

Szwed, Piotr, AGH University of Science and Technology

Tsay, Li-Shiang, North Carolina A&T State University, United States

Unold, Olgierd, Wroclaw University of Technology, Poland

Wozniak, Michal, Wroclaw University of Technology, Poland

Wysocki, Marian, Rzeszow University of Technology, Poland

Zaharie, Daniela, West University of Timisoara, Romania

Zighed, Djamel Abdelkader, University of Lyon, Lyon 2, France

Ziolko, Bartosz, AGH University of Science and Technology, Poland

DOI: 10.15439/2014F401
ACSIS, Vol. 3

Full Rough Sets

Ray-Ming Chen
Department of Computer Science
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martenstraße 3, 91058 Erlangen, Germany
ray.chen@fau.de

***Abstract*—I will analyze Pawlak's rough sets and extend the usual setting of characterization via an equivalence relation to any arbitrary relation, or a full relation. I extract the specification of rough sets by two operators: probable operator and sure operator. This would provide a general framework to analyze full rough sets and to compare them with Pawlak's rough sets. It also facilitates all the computation and further expansion. This paper extends Pawlak's approximation approach to any arbitrary relations and explicitly defines several computational operators to study the closeness regarding the characterization of operations on sets.**

I. Introduction

IN PAWLAK'S paper "Rough Sets", he studied the properties of an approximation space $A = (U, R)$, where U is a universe and R is an equivalence relation. He then in Section 2.3 (pp. 344-345), listed 30 identities of the properties of the approximation space and these become a foundation of his theory. These identities specify the properties of set operations and best upper approximation $\overline{Apr}$ and best lower approximation $\underline{Apr}$. These identities also have some practical applications. For example, to calculate the best upper and lower approximation of the $X \cup Y \subseteq U$, one only needs to calculate the respective best approximations of X and Y. This would save some resources and calculations if it is put into practical applications. When I read Pawlak's "Rough Sets", the first question comes to my mind is: Could I extend this equivalent relation to any arbitrary relation? Then the second question follows: if I extend it to any arbitrary relation, would these identities hold? Then it goes to the third question: under which circumstances could these identities hold? And then the last question: is there any relationship between different set operations under these approximations?

Such extension is also briefly mentioned in several occasions, for example, a chapter written by Hung Son Nguyen and Andrzej Skowron [7]. Indeed, there are some authors trying to generalize and extend R, in particular, Y.Y.Yao ([3][4],[5],[6]). In Yao's papers, he specifically defines an extended R, say R^* (a non-equivalence relation), and then studies the properties of best approximations and these generalized rough sets.

Unlike Yao's papers, my intention is not to specify or study any particular extended relation. My intention is to study any arbitrary relations. Another main difference is my research will focus on answering my last two questions, which are rarely mentioned in any other papers, not even in Pawlak's original paper. The results will be presented in Section VI. Another characteristic of my paper is to automate full rough sets approximation, in the sense that a machine or a computer can easily implement the calculations. In order to achieve this, in Section III,I introduce several operators and a characteristic matrix to facilitate the whole computation. Throughout the whole paper, I also compare the differences between Pawlak's rough sets and full rough sets, in particular, Example 18.

II. Background

In the typical characterization of a crisp target set, an equivalence relation is used to partition the universe and then the partition is served as a classifier for the target sets. However, in the real world, one needs to develop a much more complicated classifier for the target sets. To begin with, let me show some cases that would involve some complicated relations to form granule knowledge. In reality, there might exist more than one equivalence relation and no favorite is made or the interpretation of data needs some tolerance or other reasons. Then one has to contrive a much more complicated classifier.

Example 1 (tolerant classifier)**.** An experimenter observed some viruses' infection rates of cows and recorded them as the following table:

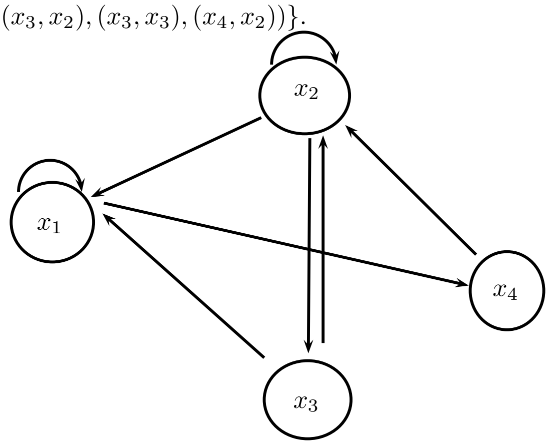
Viruses	Infection Rate (%)	Viruses	Infection Rate (%)
a	8.22	g	40.19
b	50.21	h	11.37
c	32.49	i	20.01
d	18.41	j	10.73
e	29.83	k	19.89
f	1.24	l	9.86

TABLE I
Virus Infection Rate

Considering some degree of tolerance for the measurements in the endpoints, a biologist forms his granule knowledge of these viruses as follows: a type one (or $T1$) virus is a virus whose infection rate lies between 0 and 10.25; a type two (or $T2$), 9.75 and 20.25; a type three (or $T3$), 19.75 and 30.25; a type four (or $T4$), 29.75 and 40.25 and a type five (or $T5$), 39.75 and 50.25. He then forms the classifier $\{\{a, f, l\}, \{d, h, i, j, k, l\}, \{e, i, k\}, \{c, e, g\}, \{b, g\}\}$.

This classifier will be his granule knowledge to characterize target sets. Furthermore, some granule knowledge might be formed subjectively or randomly.

Example 2 (Relational Tree)**.** Suppose the universe $U = \{x_1, x_2, x_3, x_4\}$ and the following relation or its directed diagram:
$\mathcal{R} := \{(x_1, x_1), (x_1, x_4), (x_2, x_1), (x_2, x_2), (x_2, x_3), (x_3, x_1), (x_3, x_2), (x_3, x_3), (x_4, x_2))\}$.



If the granule knowledge is formed via this relation, then the classifier would be $\{\{x_1, x_4\}, \{x_1, x_2, x_3\}, \{x_1, x_2\}, \{x_2\}\}$. This generalized relation is also studied in Section 2.1 of Yao's paper [3] "On Generalizing Rough Set Theory ".

In addition, there are many other cases that would involve non-equivalence relations. For example, some data is missing in forming an equivalent relation. In Section III, full rough sets will be introduced. In Section VI, I will study the properties of full rough sets.

III. Preliminary

First of all, let us recall how Pawlak defines his rough sets. Let $A = (U, R)$ be an approximation space. He then defines the $R-$lower approximation of $X \subseteq U$, $\underline{Apr_A}(X) := \underset{x \in U}{\cup} \{R(x) : R(x) \subseteq X\}$ and the $R-$upper approximation of $X \subseteq U$, $\overline{Apr_A}(X) := \underset{x \in U}{\cup} \{R(x) : R(x) \cap X \neq \emptyset\}$, where $R \subseteq U \times U$ is an indiscernibility relation and $R(x)$ is the equivalence class of R determined by the element $x \in U$. In order to name it, we call the partition (or equivalence classes) U/R a classifier in the sense that any subset $X \subseteq U$ is classified by a pair of rough sets via this partition. Now the first step for me to extract this definition and generalize it to any arbitrary relation is to use a functional classifier. Let U be a finite universe. In this paper, I will assume there exists a bijective function to specify the elements in U. Let $|U|^{\downarrow} \equiv \{1, 2, ..., |U|\}$.

Definition 1. For any function from U to $\mathcal{P}(U)$, which is chosen as a classifier, is called a functional classifier. Let $\mathcal{K} : U \to \mathcal{P}(U)$ be a functional classifier.

It is worth mentioning that this definition also appears in Y.Y.Yao's paper [3] in the name of a successor neighborhood $R_s(x) = \{y | y \in U, xRy\}$, where $R \subseteq$ is any arbitrary binary relation. But we differ in the way forming upper and lower best approximation. Here I reiterate his definitions: $\underline{apr_R}(A) = \{x | R_s(x) \subseteq A\}$ and $\overline{apr_R}(A) = \{x | R_s(x) \cap A \neq \emptyset\}$. Now I introduce some definitions and notations.

Inspired by Lotfi A. Zadeh's approach (in [8]) using characteristic functions to represent fuzzy sets, I introduce characteristic matrix to represent a functional classifier to facilitate the computations. Each functional classifier $\mathcal{K}$ can then be represented by a matrix $\mathcal{K}^*$ in $2^{U \times U}$ via its characteristic values. We name $\mathcal{K}^*$ a characteristic classifier.

Example 3. Suppose $U = \{u_1, u_2, u_3\}$ and $\mathcal{K}(u_1) = \{u_1, u_2, u_3\}$, $\mathcal{K}(u_2) = \{u_1, u_3\}$ and $\mathcal{K}(u_3) = \{u_2\}$. Then one has the following matrix-operation form:

$$\mathcal{K} = \begin{bmatrix} \mathcal{K}(u_1) \\ \mathcal{K}(u_2) \\ \mathcal{K}(u_3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

Thus $\mathcal{K}$ can be represented by its characteristic matrix

$$\mathcal{K}^* = \begin{bmatrix} \mathcal{K}^*(1) \\ \mathcal{K}^*(2) \\ \mathcal{K}^*(3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

where $\mathcal{K}^*(j)$ is the characteristic function of $\mathcal{K}(u_j)$.

By reading at each row vector, one knows immediately the member of that class. This representation would facilitate our computation largely and it could also be programmed to automate the approximation process. In addition, each subset $X \subseteq U$ is also represented by its characteristic function $\rho \in 2^U$, for example, a subset $X = \{u_1, u_4\}$ in a universe U, with $|U| = 5$ would be represented by its characteristic function [10010]. Since we restrict K to be a characteristic classifier, and thus K^* is always a square matrix with a dimension $|U|$. Of course, one can further lift such restriction and extend K to a relational classifier, i.e., for some element $u \in U$, there associate more than two granule classes $K_1(u)$ and $K_2(u)$. In this paper, we will not go into such extension.

Definition 2. For any arbitrary subset $\rho \in 2^U$, define $\rho(n)$ to be the n'th element of U; for any arbitrary characteristic classifier $T \in 2^{U \times U}$, define $T(n)$ be the n'th row of T and $T(m, n)$ be the element in m'th row and n'th column.

Let $\mathcal{K}^*(|U|^{\downarrow}) = \{\mathcal{K}^*(j) : j \in |U|^{\downarrow}\}$, the set of all the granule knowledge (classes). Take Example 3 for example. Then one has $\mathcal{K}^*(|U|^{\downarrow}) = \{\mathcal{K}^*(1), \mathcal{K}^*(2), \mathcal{K}^*(3)\} = \{\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}.\}$. We can further define some set union and intersection operations in the form of characteristic functions.

Definition 3. Define $\cup^* : 2^U \times 2^U \to 2^U$ by $(\rho \cup^* \tau)(n) := \sup\{\rho(n), \tau(n)\}$.

Example 4. $\begin{bmatrix} 1 & 0 & 0 & 1 & 1 \end{bmatrix} \cup^* \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \end{bmatrix}$.

Definition 4. Define $\cap^* : 2^U \times 2^U \to 2^U$ by $(\rho \cap^* \tau)(n) := \inf\{\rho(n), \tau(n)\}$.

Example 5. $\begin{bmatrix} 1 & 0 & 0 & 1 & 1 \end{bmatrix} \cap^* \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}$.

This kind of definitions is common for defining the set operation, for example, in Zadeh's paper [8]. We can further define the subset operation in the form of characteristic functions.

Definition 5. For all $\rho, \tau \in 2^U$, define $\rho \subseteq^* \tau$ iff $\forall n \in |U|^{\downarrow}[\rho(n) \leq \tau(n)]$.

We also use the notations $\vec{0} : |U|^{\downarrow} \rightarrow \{0,1\}$ with $\vec{0}(n) = 0$ for $\forall n \in |U|^{\downarrow}$ and $\vec{1} : |U|^{\downarrow} \rightarrow \{0,1\}$ with $\vec{1}(n) = 1$ for $\forall n \in |U|^{\downarrow}$. Let $\rho \in 2^U$ be a target set. Now we start to define upper and lower best approximations. In correspondence to our setting, here I name them full upper bound and full lower bound, respectively. The definition of these bounds will be presented in the form of operations over a characteristic classifier. Now we can define the full rough sets as follows:

Definition 6 (Full Upper Bound)**.** Define the full upper bound of ρ as $\rho^+ = \cup^* \{\tau \in \mathcal{K}^*(|U|^{\downarrow}) : \tau \cap^* \rho \neq \vec{0}\}$.

Definition 7 (Full Lower Bound)**.** Define the full lower bound of ρ as $\rho^- = \cup^* \{\tau \in \mathcal{K}^*(|U|^{\downarrow}) : \tau \cap^* \rho \neq \vec{0}, \tau \subseteq^* \rho\}$.

$\cup^* S$ denotes the $\cup^*$ operation over all the elements in set S. (ρ^-, ρ^+) are called full rough sets of the characteristic target set ρ via the characteristic classifier K^*. These full rough sets will be further characterized in Lemma 1. These definitions indeed are the characteristic version of the following usual rough set definitions (except now we consider arbitrary relation). Let U be an arbitrary universe. Let $X \subseteq U$ be a target set. Let $R \subseteq U \times U$ be an arbitrary relation. Let $R(x) = \{y \in U : xRy\}$. Let the functional classifier $\mathcal{K} = \{R(x) : x \in U\}$.

Definition 8 (Full Upper Bound)**.** Define the full upper bound of X as $X^+ = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \emptyset\}$.

Definition 9 (Full Lower Bound)**.** Define the full lower bound of X as $X^- = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \emptyset, R(x) \subseteq X\}$.

Here I change the usual setting of best lower approximation a bit by adding a clause $R(x) \cap X \neq \emptyset$. This change has no impact on the current theory. It is just a condition for set specification when one intends for further extension. We use the notation $\rho^{-1}\{1\} = \{n \in |U|^{\downarrow} : \rho(n) = 1\}$, for all $\rho \in 2^U$. $\rho^{-1}\{1\}$ shows the elements in ρ. Now I define the difference of two sets in the form of characteristic functions.

Definition 10. Define $-^* : 2^U \times 2^U \rightarrow 2^U$ by $(\rho -^* \tau)(n) = inf\{\rho(n), 1 - inf\{\rho(n), \tau(n)\}\}$.

Example 6. $\begin{bmatrix}1 & 0 & 0 & 1 & 0\end{bmatrix} -^* \begin{bmatrix}0 & 1 & 0 & 1 & 1\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}$.

There are many ways to define a characteristic-function version of a difference set operation. For any set $X \in \mathcal{P}(U)$, let $X^* \in 2^U$ denote its characteristic function. In order to construct a computational approach for upper or lower best approximations, I extract the definitions and redefine them in the form of characteristic functions. To test whether two sets (characteristic functions) intersects or not, I define the following intersection indicator $\bullet$.

Definition 11 (Intersection Indicator:$\bullet$)**.** Define $\bullet : 2^U \times 2^U \rightarrow \{0,1\}$ by $\rho \bullet \tau := sup\{\tau(n) : n \in \rho^{-1}\{1\}\}$.

The computational idea for this definition is: to see whether τ and ρ has common elements or not, one finds the order numbers of ρ and then checks whether these numbers are shared by τ. If yes, then they intersects with each other; if not, there is no intersection between them, i.e., the intersection is $\vec{0}$ (the characteristic function for the empty set). In sum, if $\rho \bullet \tau = 1$, it means ρ intersects τ; if $\rho \bullet \tau = 0$, it means ρ and τ are disjointed.

Example 7. $\begin{bmatrix}1 & 1 & 0 & 1 & 1\end{bmatrix} \bullet \begin{bmatrix}0 & 1 & 1 & 1 & 0\end{bmatrix} = 1$.

Now we almost go into the main body of my approach. Recall a usual setting of a best approximation, for example, $\underline{Apr_A}(X) := \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\}$. In order to put it into a computational content, I separate it into two parts: specification test and value assigned. In this case, the specification to be tested is: $R(x) \subseteq X$. So one tests whether $R(x) \subseteq X$ or not. If yes, then one assigns one value for the $\underline{Apr_A}(X)$ and if no, one assigns the other value. There are two main reasons for such approach: firstly, to automate the computation and secondly, to leave a room for further definition of the usual $\cup$ approximation-indeed one would be given much more freedom to assign the approximation values.

In the next claim, I will convert the usual specification of rough sets into the context of characteristic functions and classifiers.

Claim 1. Let $Y \in \mathcal{P}(U)$ be arbitrary. Let $\mathcal{K}$ be an arbitrary classifier. Then

1) $Y \cap \mathcal{K}(u_j) \neq \emptyset$ iff $Y^* \cap^* \mathcal{K}^*(j) \neq \vec{0}$ iff $Y^* \bullet \mathcal{K}^*(j) = 1$.
2) $\mathcal{K}(u_j) \subseteq Y$ iff $\mathcal{K}^*(j) \subseteq^* Y^*$ iff $(\vec{1} -^* Y^*) \bullet \mathcal{K}^*(j) = 0$.
3) $Y \cap \mathcal{K}(u_j) \neq \emptyset$ and $\mathcal{K}(u_j) \subseteq Y$ iff $Y^* \cap^* \mathcal{K}^*(j) \neq \vec{0}$ and $\mathcal{K}^*(j) \subseteq^* Y^*$ iff $inf\{Y^* \bullet \mathcal{K}^*(j), 1 - (\vec{1} -^* Y^*) \bullet \mathcal{K}^*(j)\} = 1$ iff $Y^* \bullet \mathcal{K}^*(j) - (\vec{1} -^* Y^*) \bullet \mathcal{K}^*(j) = 1$

Proof. (1) and (2) follow immediately from the definitions. Here I show (3): $inf\{Y^* \bullet \mathcal{K}^*(j), 1 - (\vec{1} -^* Y^*) \bullet \mathcal{K}^*(j)\} = 1$ iff $Y^* \bullet \mathcal{K}^*(j) = 1 - (\vec{1} -^* Y^*) \bullet \mathcal{K}^*(j)\} = 1$ iff $Y^* \bullet \mathcal{K}^*(j) = 1$ and $(\vec{1} -^* Y^*) \bullet \mathcal{K}^*(j) = 0$ iff $Y^* \cap^* \mathcal{K}^*(j) \neq \vec{0}$ and $\mathcal{K}^*(j) \subseteq^* Y^*$. □

This claim indeed forms a cornerstone which converts the usual rough sets formation into a computational characteristic target sets and characteristic classifier. From this claim, we characterize the set specification $Y \cap \mathcal{K}(u_j) \neq \emptyset$ with a probable operator po defined in Definition 13. We can also characterize the set specification $Y \cap \mathcal{K}(u_j) \neq \emptyset$ and $\mathcal{K}(u_j) \subseteq Y$ with a sure operator so defined in Definition 14. Since we have already defined an intersection indicator, we could now further define an intersection operator $\odot$ to compute and test the intersection between a target set (a characteristic function) and a functional classifier (a characteristic classifier).

Definition 12 (Intersection Operator $\odot$)**.** Define $\odot : 2^U \times 2^{U \times U} \rightarrow 2^U$ by $(\rho \odot T)(n) := \rho \bullet T(n)$.

Example 8. $\begin{bmatrix}1 & 0 & 1 & 0 & 0\end{bmatrix} \odot \begin{bmatrix}1 & 0 & 1 & 0 & 0\\ 0 & 1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1\\ 0 & 1 & 0 & 0 & 0\\ 1 & 1 & 1 & 0 & 1\end{bmatrix} = \begin{bmatrix}1 & 1 & 0 & 0 & 1\end{bmatrix}$.

This operator shows whether the target set intersects with the classifier or not. With this operator, one can now easily visualize the intersection between target sets and its classifier.

Now we separate the set specifications from the set comprehension via two operators: specification operators and value operators. Specification operators consist of two operators: probable operator (po) and sure operator (so).

Definition 13. Define $po : 2^U \times 2^{U\times U} \to 2^U$ by $po(\rho, T) := \rho \odot T$

Definition 14. $so : 2^U \times 2^{U\times U} \to 2^U$ by $so(\rho, T) = \rho \odot T -^* (\vec{1} -^* \rho) \odot T$.

Example 9. $so(\begin{bmatrix}1 & 0 & 0 & 1 & 1\end{bmatrix}, \begin{bmatrix}1 & 0 & 0 & 0 & 1\\ 0 & 1 & 1 & 0 & 0\\ 1 & 0 & 0 & 1 & 0\\ 0 & 1 & 1 & 1 & 0\\ 1 & 0 & 1 & 0 & 1\end{bmatrix}) = \begin{bmatrix}1 & 0 & 1 & 1 & 1\end{bmatrix} -^* \begin{bmatrix}0 & 1 & 0 & 1 & 1\end{bmatrix} = \begin{bmatrix}1 & 0 & 1 & 0 & 0\end{bmatrix}$.

Definition 15. We say (ρ, T) is a crisp specification iff $po(\rho, T) = so(\rho, T)$ and (ρ, T) is a rough specification iff $po(\rho, T) \neq so(\rho, T)$.

Though I separate specification and value assignment, in this paper, I will not change the usual setting of $\cup$-value assignment. The following definitions maintain the usual $\cup$-value assignment in the computational context. In the following definition, we start to specify the form of a classifier.

Definition 16. Define $sup^* : 2^{U\times U} \to 2^U$ by $(sup^*(T))(n) := sup\{T(n, m) : m \in |U|^{\downarrow}\}$.

Example 10. $sup^*(\begin{bmatrix}0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 1 & 1\end{bmatrix}) = \begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}$.

Example 11. $sup^*(\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 1 & 0 & 1\\ 1 & 0 & 1 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}) = \begin{bmatrix}0 & 1 & 1 & 0\end{bmatrix}$.

We say that a characteristic classifier T is degenerated if and only if there exists an element 0 in $sup^*(T)$ and non-degenerated iff all the elements are 1 in $sup^*(T)$. For example, T in Example 10 is non-degenerated, while in Example 11 is degenerated. For a general setting, one would like to specify the exact form of T, for example, T is formed via equivalence relation or other relations. However, I study a more general relation, i.e., a non-generated T.

In most of the settings and derivations, the universe U itself is taken as closed under set operations. In order to study the detailed properties, I generalize it to study any subsets of U that are closed under set operations. On the one hand, it helps us to understand that some properties not only belong to the universe per se, but it also applies to its subsets; on the other hand, it also helps us classify a set of target sets that are closed under set operations.

Definition 17. $S \subseteq 2^U$ is $\cup^*$-closed iff for all $\rho, \tau \in S$, one has $\rho \cup^* \tau \in S$.

Definition 18. $S \subseteq 2^U$ is $\cap^*$-closed iff for all $\rho, \tau \in S$, one has $\rho \cap^* \tau \in S$.

Definition 19. $S \subseteq 2^U$ is c^*-closed iff for all $\rho \in S$, one has $\vec{1} -^* \rho \in S$.

Example 12. $S = \{\begin{bmatrix}1 & 0 & 1\end{bmatrix}, \begin{bmatrix}1 & 0 & 0\end{bmatrix}, \begin{bmatrix}1 & 1 & 1\end{bmatrix}\}$ is $\cup^*$-closed and $\cap^*$-closed, but not c^*-closed.

IV. Basic Identity

Let $\rho, \tau, \eta, \theta \in 2^U$ be arbitrary. Then we have the following basic identities of set operations in the form of characteristic functions. These identities will be applied in Section VI without an explicit mention.

Claim 2.
1) $\rho \cup^* (\tau -^* \eta) = (\rho \cup^* \tau) -^* (\eta -^* \rho)$.
2) $\rho -^* (\tau -^* \eta) = (\rho \cap^* \eta) \cup^* (\rho -^* \tau)$.
3) $(\rho -^* \tau) -^* \eta = \rho -^* (\tau \cup^* \eta)$.
4) $\rho \cap^* (\tau -^* \eta) = (\rho \cap^* \tau) -^* \eta = \tau \cap^* (\rho -^* \eta)$.
5) $\rho -^* (\tau \cup^* \eta) = (\rho -^* \tau) \cap^* (\rho -^* \eta)$.
6) $\rho -^* (\tau \cap^* \eta) = (\rho -^* \tau) \cup^* (\rho -^* \eta)$.
7) $(\rho \cup^* \tau) -^* \eta = (\rho -^* \eta) \cup^* (\tau -^* \eta)$.
8) $(\rho -^* \tau) \cup^* \eta = (\rho \cup^* \eta) -^* (\tau -^* \eta)$.

Proof. All these identities follow immediately from the basic set operations in terms of characteristic functions. Take (6) for example. It comes directly from the identity: $A - (B \cap C) = (A - B) \cup (A - C)$ for any arbitrary sets A, B and C. □

V. Property

First of all, a computational characterization of a target set is derived. Secondly, the properties of all the operators defined will be studied. In the next lemma, I will show how to characterize full rough sets defined in Definition 6 and 7 by the following lemma.

Lemma 1. Given a target set $\rho \in 2^U$ and a characteristic classifier $T \in 2^{U\times U}$, one gets its full upper bound (i.e., best upper approximation) $\rho^+ = po(\rho, T) \odot T^t$ and its full lower bound (i.e., best lower approximation) as $\rho^- = so(\rho, T) \odot T^t$, where T^t denotes the transpose of the matrix T.

Proof. Suppose $po(\rho, T) = f$. Then the full upper bound $\rho^{+} = \cup^{*}\{T(n) : n \in f^{-1}\{1\}\} = f \odot T^{t}$. Similarly, let $so(\rho, T) = g$. Then the full lower bound $\rho^{-} = \cup^{*}\{T(n) : n \in g^{-1}\{1\}\} = g \odot T^{t}$. □

This lemma shows directly how to compute best upper approximation and best lower approximation. Indeed one only needs to find out the $po(\rho, T)$ for each target set ρ to get the final result. Henceforth, in this paper, I will focus on the study of the properties of specification operators po and so.

Definition 20 (Full Upper Bound)**.** Define $ub : 2^{U} \times 2^{U\times U} \to 2^{U}$ to be $ub(\rho, T) := po(\rho, T) \odot T^{t}$, where T^{t} denotes the transpose of the matrix of T.

Definition 21 (Full Lower Bound)**.** Define $lb : 2^{U} \times 2^{U\times U} \to 2^{U}$ to be $lb(\rho, T) := so(\rho, T) \odot T^{t}$, where T^{t} denotes the transpose of the matrix of T.

Example 13. Let the target set be $X = \{x_2, x_3\}$. Let T be the characteristic classifier induced by the classifier in Example 2. Let us use this formalism to compute its full rough sets. To begin with, let me compute the usual setting first via Definition 8 and 9. Then one has the best lower bound $X^{-} = \{x_2\}$ and the best upper bound $X^{+} = \{x_1, x_2, x_3\}$. Now let us look into the characteristic version. The $\mathcal{R}$-induced characteristic classifier is $T = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$ and the characteristic target set is $X^{*} = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}$. $so(X^{*}, T) = X^{*} \odot T -^{*} (\vec{1} -^{*} X^{*}) \odot T = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix} -^{*} \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$. Now we have $lb(X^{*}, T) = so(X^{*}, T) \odot T^{t} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$; similarly, one has $po(X^{*}, T) = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}$ and thus $ub(X^{*}, T) = po(X^{*}, T) \odot T^{t} = \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}$.

Image that the size of the universe is two million and the size of the target set is one million. Are we still using our naked eyes to classify this target set? Basically, this characteristic version provides a systematic way to assign full rough sets to a target set and it also provides a theoretical framework for better understanding of full rough sets.

Once one gets the values of $po(\rho, T)$ and $so(\rho, T)$, he gets the values of upper bound and lower bound straightaway. Henceforth, we will focus on studying the properties of po and so in this paper. To further understand the above-mentioned operators, I derive some of their properties here. Some of the properties would be applied later. There operators also play vital roles in gaining our main results in Section VI. Let $\rho, \rho', \tau, \tau', \eta \in 2^{U}$ and $T \in 2^{U\times U}$ be arbitrary.

A. Operator $\bullet$

Claim 3. If $\rho \subseteq^{*} \tau$ and $\rho' \subseteq^{*} \tau'$, then $\rho \bullet \rho' \leq \tau \bullet \tau'$.

Proof. $\rho \bullet \rho' = sup\{\rho'(n) : n \in \rho^{-1}\{1\}\} \leq sup\{\rho'(n) : n \in \tau^{-1}\{1\}\} \leq sup\{\tau'(n) : n \in \tau^{-1}\{1\}\} = \tau \bullet \tau'$. □

Claim 4. $\rho \bullet \tau = \tau \bullet \rho$

Proof. $\rho \bullet \tau = 1$ iff there exists $n \in |U|^{\downarrow}$ such that $\rho(n) = \tau(n) = 1$. □

Claim 5. $(\rho \cup^{*} \tau) \bullet \eta = sup\{\rho \bullet \eta, \tau \bullet \eta\}$.

Proof. $(\rho \cup^{*} \tau) \bullet \eta = sup\{\eta(n) : n \in (\rho \cup^{*} \tau)^{-1}\{1\} = sup\{\eta(n) : n \in \rho^{-1}\{1\} \cup \tau^{-1}\{1\}\} = sup\{sup\{\eta(n) : n \in \rho^{-1}\{1\}\}, sup\{\eta(n) : n \in \tau^{-1}\{1\}\}\}$. □

B. Operator $\odot$

Let $\rho, \rho', \tau \in 2^{U}$ and $T \in 2^{U\times U}$ be arbitrary.

Claim 6. $(\rho \cup^{*} \tau) \odot T = (\rho \odot T) \cup^{*} (\tau \odot T)$, i.e., $po(\rho \cup^{*} \tau, T) = po(\rho, T) \cup^{*} po(\tau, T)$.

Proof. $((\rho \cup^{*} \tau) \odot T)(n) = (\rho \cup^{*} \tau) \bullet T(n) = sup\{\rho \bullet T(n), \tau \bullet T(n)\}$, by Claim 5. □

We can then apply this identity to show other special cases.

Example 14. Property (6) and (15) at Section 2.3 in Pawlak's paper [1] are the direct instances of this claim (via Lemma 1). Take property (6) for example: $ub(X^{*} \cup^{*} Y^{*}, T) = po(X^{*} \cup^{*} Y^{*}, T) \odot T^{t} = (po(X^{*}, T) \cup^{*} po(Y^{*}, T)) \odot T^{t} = (po(X^{*}, T) \odot T^{t}) \cup^{*} (po(X^{*}, T) \odot) T^{t} = ub(X^{*}, T) \cup^{*} ub(Y^{*}, T)$, where T is the characteristic classifier induced by the equivalence relation.

Remark 1. Generally speaking, $(\rho \cap^{*} \tau) \odot T \neq (\rho \odot T) \cap^{*} (\tau \odot T)$, for example, $\rho = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}, \tau = \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}, T = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$. Then $(\rho \cap^{*} \tau) \odot T = \vec{0} \neq \rho \odot T \cap^{*} \tau \odot T = \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}$.

In the following, I will show what kind of T is eligible to keep hold of $\rho \odot T \cap^{*} \tau \odot T = (\rho \cap^{*} \tau) \odot T$.

Definition 22. Define $sum^{*} : 2^{U\times U} \to (|U|^{\downarrow})^{U}$ by $(sum^{*}(T))(n) := \sum_{j \in |U|^{\downarrow}} T(n, j)$.

Example 15. $sum^{*}(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}) = \begin{bmatrix} 2 & 1 & 3 \end{bmatrix}$.

Claim 7. $sum^{*}(T) \subseteq^{*} \vec{1}$ iff for all $\rho, \tau \in 2^{U}[\rho \odot T \cap^{*} \tau \odot T = (\rho \cap^{*} \tau) \odot T]$.

Proof. Let $\rho \cap^{*} \tau = \eta$. Then $\rho \odot T \cap^{*} \tau \odot T = [(\eta \cup^{*} (\rho -^{*} \eta)) \odot T] \cap^{*} [(\eta \cup^{*} (\tau -^{*} \eta)) \odot T] = \eta \odot T \cup^{*} [(\rho -^{*} \eta) \odot T \cap^{*} (\tau -^{*} \eta) \odot T] = \eta \odot T$. On the other hand, suppose $sum^{*}(T) \not\subseteq^{*} \vec{1}$, i.e., there exist $i, j, n \in |U|^{\downarrow}$ such that $i \neq j$ and $T(n, i) = T(n, j) = 1$. Now define $\rho(k) = 1$ if $k = i$ and $\rho(k) = 0$, otherwise and define $\tau(k) = 1$ if $k = j$ and $\tau(k) = 0$, otherwise for all $k \in |U|^{\downarrow}$. Then one has $\rho \odot T \cap^{*} \tau \odot T \neq (\rho \cap^{*} \tau) \odot T = \vec{0}$. □

Corollary 1. $\forall \rho, \tau \in 2^U [\rho \odot T \cap^* \tau \odot T = (\rho \cap^* \tau) \odot T]$ iff $\forall \rho \in 2^U [\rho \odot T \cap^* (\vec{1} -^* \rho) \odot T = \vec{0}]$.

Proof. Since $sum^*(T) \subseteq^* \vec{1}$ iff $\forall \rho \in 2^U [\rho \odot T \cap^* (\vec{1} -^* \rho) \odot T = \vec{0}]$, by Claim 7, the result follows. □

Claim 8. If $\rho \subseteq^* \tau$, then $\rho \odot T \subseteq^* \tau \odot T$.

Proof. It follows immediately from the definition and Claim 3. □

Example 16. Property (10) at Section 2.3 in Pawlak's paper [1] is a direct instance of this claim (via Lemma 1).

C. Operator $-^*$

Claim 9. $(\rho -^* \tau)^{-1}\{1\} = \rho^{-1}\{1\} - \tau^{-1}\{1\}$.

Proof. $n \in (\rho -^* \tau)^{-1}\{1\}$ iff $\rho(n) = 1$ and $\tau(n) = 0$ iff $n \in \rho^{-1}\{1\} - \tau^{-1}\{1\}$. □

D. Specification operators: po and so

Let $\rho, \rho', \tau \in 2^U, T \in 2^{U \times U}$ and $S \subseteq 2^U$ be arbitrary.

Claim 10. (ρ, T) is a crisp specification (defined at Definition 15) iff $(\rho \odot T) \cap^* ((\vec{1} -^* \rho) \odot T) = \vec{0}$.

Proof. $\rho \odot T = \rho \odot T -^* (\vec{1} -^* \rho) \odot T$ iff $\rho \odot T \cap^* (\vec{1} -^* \rho) \odot T = \vec{0}$. □

Claim 11. $so(\rho, T) \cap^* so(\rho', T) = (\rho \odot T \cap^* \rho' \odot T) -^* (\vec{1} -^* \rho \cap^* \rho') \odot T$.

Proof. $so(\rho, T) \cap^* so(\rho', T)[\rho \odot T -^* (\vec{1} -^* \rho) \odot T] \cap^* \rho' \odot T -^* (\vec{1} -^* \rho') \odot T] = [\rho \odot T \cap^* (\rho' \odot T -^* (\vec{1} -^* \rho') \odot T)] -^* (\vec{1} -^* \rho) \odot T = [(\rho \odot T \cap^* \rho' \odot T -^* (\vec{1} -^* \rho') \odot T)] -^* (\vec{1} -^* \rho) \odot T = \rho \odot T \cap^* \rho' \odot T -^* [(\vec{1} -^* \rho) \odot T \cup^* (\vec{1} -^* \rho') \odot T] = \rho \odot T \cap^* \rho' \odot T -^* [((\vec{1} -^* \rho) \cup^* (\vec{1} -^* \rho')) \odot T]$, by Claim 6 $= \rho \odot T \cap^* \rho' \odot T -^* (\vec{1} -^* \rho \cap^* \rho') \odot T$. □

Claim 12. po, so are all $\subseteq^*$-non-decreasing functions.

Proof. Let ρ, τ be arbitrary such that $\rho \subseteq^* \tau$. Then one has $\rho \bullet T(x) = sup\{T(x)(j) : j \in \rho^{-1}(1) \leq sup\{T(x)(j) : j \in \tau^{-1}(1)\} = \tau \bullet T(x)$; on the other hand,$so(\rho, T) = \rho \odot T -^* (\vec{1} -^* \rho) \odot T \subseteq^* \rho \odot T -^* (\vec{1} -^* \tau) \odot T \subseteq^* \tau \odot T -^* (\vec{1} -^* \tau) \odot T = so(\tau, T)$ □

Corollary 2. (Monotonicity of Full Upper/Lower Bound) If $\rho \subseteq^* \tau$, then $lb(\rho, T) \subseteq^* lb(\tau, T)$ and $ub(\rho, T) \subseteq^* ub(\tau, T)$.

Proof. The results follow immediately from Definition 6 and 7, Claim 12, Lemma 1 and Claim 8. □

Example 17. Property (10) and (11) in Pawlak's paper are such instances.

VI. Main Results

From Lemma 1, we know it is sufficient to study po and so in order to have best upper and lower approximations. In this section, I will present my main results in order to answer my last two questions in the section Introduction, i.e., the relations between upper and lower best approximation and the closeness of different set operations under various characteristic classifiers. We also try to find out what kind of imposition is needed to preserve all these properties. Before we go further, let us define some notions first.

Definition 23. $\{po(\rho, T) : \rho \in S\}$ is c^*-closed iff $\forall \rho \in S, \exists \eta \in S$ such that $\vec{1} -^* po(\rho, T) = po(\eta, T)$.

Definition 24. $\{po(\rho, T) : \rho \in S\}$ is isomorphically c^*-closed iff $\forall \rho \in S, \vec{1} -^* po(\rho, T) = po(\vec{1} -^* \rho, T)$.

Isomorphically c^*-closed is a strong relation for the complement operation. It is more informative than c^*-closed. For example, if this property holds, then one only needs to compute $po(\rho, T)$ to in order to have a result for $po(\vec{1} -^* \rho, T)$. This would save a lot of resources in the process of computation and classification. It is the same for the operation so as well.

Definition 25. $\{so(\rho, T) : \rho \in S\}$ is c^*-closed iff $\forall \rho \in S, \exists \eta \in S$ such that $\vec{1} -^* so(\rho, T) = so(\eta, T)$.

Definition 26. $\{so(\rho, T) : \rho \in S\}$ is isomorphically c^*-closed iff $\forall \rho \in S, \vec{1} -^* so(\rho, T) = so(\vec{1} -^* \rho, T)$

Now let us define some weak relations and strong relations of the set operations $\cup^*$ and $\cap^*$ for both po and so. Again, the strong relations, i.e., isomorphically-closed operations are more informative than the weak ones.

Definition 27. $\{po(\rho, T) : \rho \in S\}$ is $\cup^*$-closed iff $\forall \rho, \tau \in S, \exists \eta \in S$ such that $po(\rho, T) \cup^* po(\tau, T) = po(\eta, T)$

Definition 28. $\{po(\rho, T) : \rho \in S\}$ is isomorphically $\cup^*$-closed iff $\forall \rho, \tau \in S, po(\rho, T) \cup^* po(\tau, T) = po(\rho \cup^* \tau, T)$

When isomorphically $\cup^*$-closed holds, one only needs to compute $po(\rho, T)$ and $po(\tau, T)$ in order to compute $po(\rho \cup^* \tau, T)$. In the real computation and classification, this would save a lot of resources and time. The whole argument holds for so too.

Definition 29. $\{so(\rho, T) : \rho \in S\}$ is $\cap^*$-closed iff $\forall \rho, \tau \in S, \exists \eta \in S$ such that $so(\rho, T) \cap^* so(\tau, T) = so(\eta, T)$

Definition 30. $\{so(\rho, T) : \rho \in S\}$ is isomorphically $\cap^*$-closed iff $\forall \rho, \tau \in S, so(\rho, T) \cap^* so(\tau, T) = so(\rho \cap^* \tau, T)$

Definition 31. $\{po(\rho, T) : \rho \in S\}$ is $\cap^*$-closed iff $\forall \rho, \tau \in S, \exists \eta \in S$ such that $po(\rho, T) \cap^* po(\tau, T) = po(\eta, T)$

Definition 32. $\{po(\rho, T) : \rho \in S\}$ is isomorphically $\cap^*$-closed iff $\forall \rho, \tau \in S, po(\rho, T) \cap^* po(\tau, T) = po(\rho \cap^* \tau, T)$

Definition 33. $\{so(\rho, T) : \rho \in S\}$ is $\cup^*$-closed iff $\forall \rho, \tau \in S, \exists \eta \in S$ such that $so(\rho, T) \cup^* so(\tau, T) = so(\eta, T)$

Definition 34. $\{so(\rho,T) : \rho \in S\}$ is isomorphically $\cup^*$-closed iff $\forall \rho,\tau \in S, so(\rho,T) \cup^* so(\tau,T) = so(\rho \cup^* \tau, T)$.

Now let us start the derivations. Let $T \in 2^{U \times U}$ be arbitrary. Recall that sup^* is already defined in Definition 16.

A. Properties of unrestricted T

The following result is the most general condition for a characteristic classifier T as there is no special restriction imposed on T, i.e., T could be degenerated or non-degenerated.

Theorem 1 ($\cup^*$-Closeness of $\{po(\rho,T) : \rho \in S\}$)**.** If $S \subseteq 2^U$ is $\cup^*-$closed, then $\{po(\tau,T) : \tau \in S\}$ is also $\cup^*-$closed.

Proof. Let $\rho,\tau \in S$ be arbitrary. Then, by Claim 6, $\rho \odot T \cup^* \tau \odot T = (\rho \cup^* \tau) \odot T$. □

This specifies a sufficient condition that as long as a universe is closed under $\cup^*$ operation, po will be always closed under $\cup^*$ operation. An application is when one intends to design a classifying system via full rough sets and expects it is closed under union operation, then the universe that guarantees this is a universe that is closed under $\cup^*$ operation too.

B. Properties of T with $sup^*(T) = \vec{1}$

Now we look into the non-degenerated characteristic classifier T. A non-degenerated characteristic classifier T is a very weak requirement of T that every element in the universe U has some relation with others (at least itself). In other words, $T(n) \neq \vec{0}$ for all $n \in |U|^{\downarrow}$. Let $\rho,\tau \in 2^U$ be arbitrary. Let $T \in 2^{U \times U}$ be arbitrary such that $sup^*(T) = \vec{1}$.

Claim 13. $\vec{1} = po(\rho,T) \cup^* po(\vec{1} -^* \rho, T)$ iff $sup^*(T) = \vec{1}$.

Proof. By Claim 6, $po(\rho,T) \cup^* po(\vec{1} -^* \rho, T) = \vec{1} \odot T$. Hence $\vec{1} = \vec{1} \odot T$ iff $\forall n \in |U|^{\downarrow}(T(n) \neq \vec{0})$, i.e., $sup^*(T) = \vec{1}$. □

Claim 14. If $sup^*(T) = \vec{1}$, then $\vec{1} -^* po(\rho,T) = so(\vec{1} -^* \rho, T)$.

Proof. By Claim 13, it follows $\vec{1} -^* po(\rho,T) = [po(\rho,T) \cup^* po(\vec{1} -^* \rho,T)] -^* po(\rho,T) = po(\vec{1} -^* \rho, T) -^* po(\rho,T)$. □

This is an important result as it specifies the relation between operators po and so. One can even apply it to all the special cases.

Example 18. Property (2), (3), (8) and (9) at Section 2.3 in Pawlak's paper [1] are the direct instances of this claim. With the help of other claims, property (14), (15) and (16) could also be demonstrated (via Lemma 1, Claim 14, Claim 6, Claim 13). Take property (14) for example: $ub(X^*,T) \cup^* lb(\vec{1} -^* X^*, T) = po(X^*,T) \odot T^t \cup^* so(\vec{1} -^* X^*) \odot T^t = po(X^*,T) \odot T^t \cup^* po(\vec{1} -^* X^*, T) \odot T^t = po(\vec{1},T) \odot T^t = \vec{1}$, where T is the characteristic classifier induced by the equivalence relation.

Though the main purpose for this paper is to find out the closeness of various set operations under the specification operators po and so, it is worth mentioning that some of the identities in Pawlak's paper would fail in full rough sets. For example, Property (18) holds if T is the equivalence-relation-induced classifier by the fact that $T^t = T$ and $(\rho \odot T) \odot T^t = (\rho \odot T) \odot T = \rho \odot T$: (by Claim 14)
$ub(X^*,T) \cap^* lb(\vec{1} -^* X^*, T)$
$= (X^* \odot T) \odot T^t \cap^* (\vec{1} -^* X^* \odot T) \odot T^t$
$= (X^* \odot T) \odot T \cap^* (\vec{1} -^* X^* \odot T) \odot T$
$= (X^* \odot T) \cap^* (\vec{1} -^* X^* \odot T) = \vec{0}$.
Incidentally, the first part of Property (4): $\overline{Apr}(\overline{Apr}(X)) = \overline{Apr}(X)$ is also an instance of the fact $T^t = T$ and $(\rho \odot T) \odot T^t = (\rho \odot T) \odot T = \rho \odot T$ when T is the equivalence-relation-induced classifier. However, in a full relation (i.e., no restriction on the relation), this identity might fail. For example $X^* = \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}$ and $T = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$. Then $ub(X^*,T) = \begin{bmatrix} 1 & 0 & 1 & 1 \end{bmatrix}$ and $lb(\vec{1} -^* X^*, T) = \begin{bmatrix} 0 & 1 & 0 & 1 \end{bmatrix}$. Hence $ub(X^*,T) \cap^* lb(\vec{1} -^* X^*, T) \neq \vec{0}$.

If T is induced by a full relation, generally speaking, $(\vec{1} -^* po(\rho,T)) \odot T^t \neq \vec{1} -^* po(\rho,T)$. However, if it is restricted to an equivalence relation, $(\vec{1} -^* po(\rho,T)) \odot T^t = \vec{1} -^* po(\rho,T))$ and this , together with Claim 14 (which yiels $so(\rho,T) \odot T = so(\rho,T)$, if T is an equivalence-relation-induced classifier), directly leads to validity of Property (20) and (21) in Pawlak's paper. Moreover, generally speaking, $(\rho \odot T) \odot T^t \neq \rho \odot T$; but, imposing a restriction with an equivalence relation leads $(\rho \odot T) \odot T^t = \rho \odot T$. Hence Property (5), (7), and the De Morgan's laws: (22)-(29) in Pawlak's paper are such instances. Take (5) for example:
$lb(lb(X^*,T),T) = so(so(X^*,T) \odot T^t, T) \odot T^t$
$= so(so(X^*,T),T) = so(\vec{1} -^* po(\vec{1} -^* X^*, T), T)$
$= \vec{1} -^* po(po(\vec{1} -^* X^*, T), T) = \vec{1} -^* po(\vec{1} -^* X^*, T)$
$= so(X^*,T) = so(X^*,T) \odot T^t = lb(X^*,T)$.
Take (7) for example:
$lb(X^* \cap^* Y^*, T) = so(X^* \cap^* Y^*, T) \odot T^t = so(X^* \cap^* Y^*, T)$
$= \vec{1} -^* po(\vec{1} -^* (X^* \cap^* Y^*), T)$
$= \vec{1} -^* po((\vec{1} -^* X^*) \cup^* (\vec{1} -^* Y^*), T)$
$= \vec{1} -^* [po(\vec{1} -^* X^*, T) \cup^* po(\vec{1} -^* Y^*, T)]$
$= [\vec{1} -^* po(\vec{1} -^* X^*, T)] \cap^* [\vec{1} -^* po(\vec{1} -^* Y^*, T)]$
$= so(X^*,T) \cap^* so(Y^*,T)$
$= so(X^*,T) \odot T^t \cap^* so(Y^*,T) \odot T^t$
$= lb(X^*,T) \cap^* lb(Y^*,T)$.

Theorem 2 (Closeness: $(\{po(\rho,T) : \rho \in S\}, c^*)$ and $(\{so(\rho,T) : \rho \in S\}, c^*)$)**.** If $sup^*(T) = \vec{1}$ and $S \subseteq 2^U$ is c^*-closed, then $\{po(\rho,T) : \rho \in S\}$ is c^*- closed iff $\{so(\rho,T) : \rho \in S\}$ is c^*- closed.

Proof. Let $\rho \in S$ be arbitrary. By Claim 14, we have $\vec{1} -^* \rho \odot T = so(\vec{1} -^* \rho, T) = \vec{1} -^* [\vec{1} -^* so(\vec{1} -^* \rho, T)] = \vec{1} -^* so(\eta,T) = (\vec{1} -^* \eta) \odot T$ for some $\eta \in S$; similarly, $\vec{1} -^* so(\rho,T) = \vec{1} -^* [\vec{1} -^* po(\vec{1} -^* \rho, T)] = \vec{1} -^* po(\delta,T) = so(\vec{1} -^* \delta, T)$ for some $\delta \in S$. □

This theorem says that for any non-degenerated characteristic classifier and a c^*-closed universe (or subset), the operators po and so behaves identically with respect to the closeness of c^* operation. This is a good feature, since it simplifies some derivations when one tries to find out the c^*-closeness. He then only need to check one of the two operators.

Theorem 3. If $sup^*(T) = \vec{1}$ and $S \subseteq 2^U$ is $c*$-closed, then $\{so(\rho, T) : \rho \in S\}$ is isomorphically c^*-closed iff $\{po(\rho, T) : \rho \in S\}$ is isomorphically c^*-closed.

Proof. Let $\rho \in S$ be arbitrary. By Claim 14, $\vec{1} -^* po(\rho, T) = so(\vec{1} -^* \rho, T) = \vec{1} -^* so(\rho, T) = po(\vec{1} -^* \rho, T)$; similarly, $\vec{1} -^* so(\rho, T) = po(\vec{1} -^* \rho, T) = \vec{1} -^* po(\rho, T) = so(\vec{1} -^* \rho, T)$. □

Theorem 4 (Closeness:$(\{so(\rho, T) : \rho \in S\}, \cap^*)$ and $(\{po(\rho, T : \rho \in S)\}, \cup^*)$)**.** If $sup^*(T) = \vec{1}$ and $S \subseteq 2^U$ is $c*-$closed, then $\{so(\tau, T) : \tau \in S\}$ is $\cap^*-$close iff $\{po(\tau, T) : \tau \in S\}$ is $\cup^*$-closed.

Proof. Let $\rho, \tau \in S$ be arbitrary. By Claim 14, $po(\rho, T) \cup^* po(\tau, T) = \vec{1} -^* [so(\vec{1} -^* \rho, T) \cap^* so(\vec{1} -^* \tau, T)] = \vec{1} -^* so(\eta, T) = po(\vec{1} -^* \eta, T)$ for some $\eta \in S$. Similarly, $so(\rho, T) \cap^* so(\tau, T) = \vec{1} -^* [po(\vec{1} -^* \rho, T) \cup^* po(\vec{1} -^* \tau, T)] = \vec{1} -^* po(\delta, T) = so(\vec{1} -^* \delta, T)$ for some $\delta \in S$. □

Theorem 5. If $sup^*(T) = \vec{1}$ and $S \subseteq 2^U$ is $c*$-closed, then $\{so(\rho, T) : \rho \in S\}$ is isomorphically $\cap^*$-closed iff $\{po(\rho, T) : \rho \in S\}$ is isomorphically $\cup^*$-closed.

Proof. Let $\rho, \tau \in S$ be arbitrary. By Claim 14, $po(\rho, T) \cup^* po(\tau, T) = \vec{1} -^* [so(\vec{1} -^* \rho, T) \cap^* so(\vec{1} -^* \tau, T)] = \vec{1} -^* so((\vec{1} -^* \rho) \cap^* (\vec{1} -^* \tau), T) = \vec{1} -^* so(\vec{1} -^* \rho \cup^* \tau, T) = po(\rho \cup^* \tau, T)$. Similarly, $so(\rho, T) \cap^* so(\tau, T) = \vec{1} -^* [po(\vec{1} -^* \rho, T) \cup^* po(\vec{1} -^* \tau, T)] = \vec{1} -^* po(\vec{1} -^* \rho \cap^* \tau, T) = so(\rho \cap^* \tau, T)$. □

Corollary 3. If $sup^*(T) = \vec{1}$ and $S \subseteq 2^U$ is $\cup^*-$closed, $\cap^*-$closed and c^*-closed, then $\{so(\rho, T) : \rho \in S\}$ is isomorphically $\cap^*-$closed.

Proof. From Claim 6, $\{po(\rho, T) : \rho \in S\}$ is isomorphically $\cup^*$-closed and thus by this theorem, the result follows. □

Theorem 4 and 5 show the relation between $(\{po(\rho, T) : \rho \in S\}, \cup^*)$ and $(\{so(\rho, T) : \rho \in S\}, \cap^*)$. This again simplifies our derivations or computations since we only have to check or compute either one of the two operators. It is also useful when one tries to find a universe that will keep the relations hold, he only needs to consider one operator.

Claim 15. $so(\rho, T) \cup^* so(\tau, T) = \vec{1} -^* [po(\vec{1} -^* \rho, T) \cap^* po(\vec{1} -^* \tau, T)]$.

Proof. By Claim 14, it follows $so(\rho, T) \cup^* so(\tau, T) = so(\vec{1} - (\vec{1} - \rho), T) \cup^* so(\vec{1} - (\vec{1} - \tau), T) = [\vec{1} -^* po(\vec{1} -^* \rho, T)] \cup^* [\vec{1} -^* po(\vec{1} -^* \tau, T)] = \vec{1} -^* [po(\vec{1} -^* \rho, T) \cap^* po(\vec{1} -^* \tau, T)]$. □

Example 19. Property (22) at Section 2.3 in Pawlak's paper [1] is a direct instance of this claim (via Lemma 1).

Claim 16. $po(\rho, T) \cap^* po(\tau, T) = \vec{1} -^* [so(\vec{1} -^* \rho, T) \cup^* so(\vec{1} -^* \tau, T)]$.

Proof. By Claim 14, it follows $po(\rho, T) \cap^* po(\tau, T) = [\vec{1} -^* so(\vec{1} -^* \rho, T)] \cap^* [\vec{1} -^* so(\vec{1} -^* \tau, T)] = \vec{1} -^* [so(\vec{1} -^* \rho, T) \cup^* so(\vec{1} -^* \tau, T)]$. □

Theorem 6 (Closeness:$(\{so(\rho, T) : \rho \in S\}, \cup^*)$ and $(\{po(\rho, T) : \rho \in S\}, \cap^*)$)**.** If $sup^*(T) = \vec{1}$ and $S \subseteq 2^U$ is $c*-$closed, then $\{so(\rho, T) : \rho \in S\}$ is $\cup^*-$closed iff $\{po(\rho, T) : \rho \in S\}$ is $\cap^*$-closed.

Proof. Let $\rho, \tau \in S$ be arbitrary. By Claim 16, $po(\rho, T) \cap^* po(\tau, T) = \vec{1} -^* [so(\vec{1} -^* \rho, T) \cup^* so(\vec{1} -^* \tau, T)] = \vec{1} -^* so(\eta, T) = po(\vec{1} -^* \eta, T)$ for some $\eta \in S$, by Claim 14; similarly, by Claim 15, $so(\rho, T) \cup^* so(\tau, T) = \vec{1} -^* [po(\vec{1} -^* \rho, T) \cap^* po(\vec{1} -^* \tau, T)] = \vec{1} -^* po(\delta, T) = so(\vec{1} -^* \delta, T)$ for some $\delta \in S$, by Claim 14. □

Theorem 6 and the following 7 show the relation between $(\{po(\rho, T) : \rho \in S\}, \cap^*)$ and $(\{so(\rho, T) : \rho \in S\}, \cup^*)$. As analyzed previously, this again simplifies our computations since we only have to check one of the two operators $p0$ and so. It is also useful when one tries to construct or design a universe that will keep these relations hold, he only has to consider one operator instead.

Theorem 7. If $sup^*(T) = \vec{1}$ and $S \subseteq 2^U$ is $c*$-closed, then $\{so(\rho, T) : \rho \in S\}$ is isomorphically $\cup^*$-closed iff $\{po(\rho, T) : \rho \in S\}$ is isomorphically $\cap^*$-closed.

Proof. Let $\rho, \tau \in S$ be arbitrary. By Claim 14, 15 and 16, it follows $po(\rho, T) \cap^* po(\tau, T) = \vec{1} -^* [so(\vec{1} -^* \rho, T) \cup^* so(\vec{1} -^* \tau, T)] = \vec{1} -^* so((\vec{1} -^* \rho) \cup^* (\vec{1} -^* \tau), T) = \vec{1} -^* so(\vec{1} -^* \rho \cap^* \tau, T) = po(\rho \cap^* \tau, T)$. Similarly, $so(\rho, T) \cup^* so(\tau, T) = \vec{1} -^* [po(\vec{1} -^* \rho, T) \cap^* po(\vec{1} -^* \tau, T)] = \vec{1} -^* po(\vec{1} -^* \rho \cup^* \tau, T) = so(\rho \cup^* \tau, T)$. □

VII. Conclusions and Future Work

I provide an approach to analyze full rough sets, i.e., the usual rough sets, but with an arbitrary binary relation. The framework for this generalization offers a different perspective to look into Pawlak's rough sets. In this paper, I have analyzed and extended Pawlak's approach for characterization and defined several operators to accommodate such extension. I define the set specifications by two operators: probable operator and sure operator. These operators serve as logical conditions. I then study the properties of these logical operations. These operators also offer some computational advantages. I also study the properties of the closeness of set operations regarding full rough sets and compare their differences with Pawlak's rough sets. With this approach, one would be able to construct more complicated classifiers based on arbitrary relations. This could also enable one to define a classifier based on his own design or comprehension. For further research, one can extend the usual $\cup$-operation when defining a rough set. Since we have already separated the set specifications from such operation and known the properties of these set specifications, the introduction of other alternative of $\cup$-operation becomes feasible.

References

[1] Zdzislaw Pawlak, "Rough Sets," International Journal of Computer and Information Sciences, Vol. 11, No. 5, 1982.

[2] Zdzislaw Pawlak, "Rough Sets: theoretical aspects of reasoning about data," Springer Science+Business Media Dordrecht, 1991.

[3] Y.Y. Yao, "On generalizing rough set theory", Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Proceedings of the 9th International Conference (RSFDGrC 2003), LNAI 2639, pp. 44-51, 2003.

[4] Y,Y, Yao and S.K.M. Wong: http://www2.cs.uregina.ca/ yyao/PAPERS/relation.pdf.

[5] Y.Y. Yao, “Generalized rough set models,” Rough Sets in Knowledge Discovery, Polkowski, L. and Skowron, A. (Eds.), Physica-Verlag, Heidelberg, pp. 286-318, 1998.

[6] Y.Y. Yao and T.Y. Lin, “Generalization of rough sets using modal logic,”, Intelligent Automation and Soft Computing, An International Journal, Vol. 2, No. 2, pp. 103-120, 1996.

[7] Hung Son Nguyen and Andrzej Skowron, “Rough Sets: From Rudiments to Challenges,” Rough Sets and Intelligent Systems Professor Zdzisaw Pawlak in Memoriam, Vol. 1, Chapter 3, Springer-Verlag Berlin Heidelberg 2013.

[8] Lotfi A. Zaden, “Fuzzy Sets”, Information and Control 8, 338-353, 1965.

DOI: 10.15439/2014F70
ACSIS, Vol. 3

Approaches to the Intelligent Subject Search

Vladimir K. Ivanov,
Boris V. Palyukh
Tver State Technical University,
Afanasy Nikitin st. 22, Tver,
Russia
Email: {mtivk,
pboris}@tstu.tver.ru

Alexander N. Sotnikov
Joint Supercomputer Centre of
Russian Academy of Science
Leninsky Prospect 32a, Moscow,
Russia
Email: asotnikov@jscc.ru

***Abstract*—This article presents main results of the pilot study of approaches to the subject information search based on automated semantic processing of mass scientific and technical data. The authors focus on technology of building and qualification of search queries with the following filtering and ranking of search data. Software architecture, specific features of subject search and research results application are considered.**

I. Introduction

NEW efficient scientific knowledge search and synthesis methods (in particular, breakthrough technologies and innovative ideas in economics, science, education) are one of the top research and development targets in the field of information technology. The project Intelligent Distributed Information Management System for Innovations in Science and Education powered by the Russian Foundation of Basic Research is to solve this problem. This article presents main results of the pilot study of approaches to the subject information search based on automated semantic processing of mass scientific and technical data.

II. Specific Features of Subject Search

The major features of subject search tasks which determine the approaches are:

• the required information is often located at the junction of adjacent areas, hence, there is some complexity in the exact wording of the search query.

• along with the information on proper innovation it is desirable to obtain information on applications, risks, specific features, users, authors, producers.

• there is a necessity of available alternatives and different criteria mixing for selecting the most effective practices.

• the information on innovations is fragmentized and heterogeneous; primarily sector-specific character.

In contrast to the search for specific information (facts) on particular aspects of the required content, it is rather difficult to solve a sophisticated problem of searching coordinated information on a target subject. For example, it is required to find the economic performance of mine Raspadskaya JSCo for the first half of 2013. If we use this phrase as a search query, it is possible to get a relevant answer in the first ten search results of Google. But how can one find the information to analyze scientific, technical, economic and social factors affecting the innovative technical, technological, or financial mechanisms of coal-mining in the eastern regions of Russia?

To solve such search problems users have to employ lots of key concept combinations, clarify them in the course of en-route search on the Web or specialized stores such as patent databases (DB). It is not obvious that for this purpose any reasonable method would be used without fail. Eventually, a large amount of search results would be at the disposal of a user (tens and hundreds of documents), with the found information being more or less relevant to queries. As is quite common, there would be no opportunity to go into details of all the result data. So, the following questions can arise:

• How can one simultaneously assess the relevance of documents found by different queries? Is the relevance of documents determined correctly?

• Is the data ranking in a certain search system correct from the perspective of a user? Do all the results available for direct assessment meet the user's expectations?

• Are all the results that meet the user's expectations available for direct assessment? Are all the required data (e.g. innovative solutions) found at all?

• How one can filter documents extrinsic to the searched subject?

• Is it possible to find any effective solutions relevant in other application fields, but would be successfully used as an innovation in this domain.

• Is it possible to give a visual assessment to lots of found innovative solutions together with linked objects?

There are no clear-cut ways of solving these problems within trivial solutions. Obviously, we need efficient methods of creating and populating the computer-assisted collections of advanced technologies and ideas which would contain not only their descriptions, but selected, classified and associated data. These data can be used to analyze retrospective and prospects of specific innovations, to search current and likely trends. The project in question is an attempt to offer a number of such innovative approaches.

This work was supported by the Russian Foundation of Basic Research (contract No NK13 -07- 00342)

III. Our approach and other studies

Currently, the R&D management assurance is of great importance (see, for example, the US National Trends and International Linkages in [1]). In this context, the automated semantic processing of large arrays of scientific and technical information, for sure, is used to search for breakthrough technologies and other innovative concepts, in the same manner as it is done in the prior art solutions illumin8 [2], NetBase [3], Orbit [4], Kalypso [5], as well as in large data stores such as CORDIS [6]. The intensive application of these and other similar tools make our attention focused on IT-related issues [7], [8].

The brief overview of publications, which are instrumental in pinning down the goal of the project under discussion, is given below. Our current development solutions mainly deal with the problem of data filtration [9] in terms of a content-based approach. In this respect it is important to note some interesting and topical visions of users and developers of RDF [10], semantic web-services control [11], as well as development of the document vector space model which is fundamental to most information retrieval problems, including rating of documents, data filtering, classification and clusterization of documents [12].

Taxonomy of web searches is also of great interest to us [13]. It is worth mentioning the pattern-searching procedure like the one recently described in [14] Information Filtering by Multiple Examples (IFME) which allows users to identify their information needs as a set of relevant documents, not keywords. The use of additional relevance assessment sources can help in work with lots of short texts (for example, Twitter). The use of alternate algorithms focused on solving the top-N recommendation task [15] seems to be useful too.

Another important field of our work is associated with effective query formulation. It is subject to optimal use of the combination of information sources in order to create an extended search set [16], [17], [18]. The next essential study to be mentioned is [19]. It offers models and infrastructure for complex searches.

Note that our paper significantly expands and details [20].

IV. Problem Definition

Thus, a project goal can be summarized as: the exploration of new approaches to innovative solution search methods in the database of a data center and its population with Internet data mining results adapted to visual assessment of selected, classified and associated data. We see three key tasks to attain the goal:

- To develop the technology of building and qualification of search queries with the following filtering and ranking of search data.
- To set up methods of cluster analysis to text documents and multimedia objects in order to use them for tagging the links between search results.
- To create a store of innovative solutions for educational and scientific purposes.

V. Software Architecture

When developing a general software architecture based on mechanisms of direct automated search of innovative solutions the authors determined view layers, those of services, business logic, data access as well as crosscutting concerns (the UML notations and artifacts were applied). In the behavioral model of a system (Fig. 1), in a particular session, we can distinguish two periods of user's activation: query formulation (first step) and visualization of the results including the options of the requested and innovative solutions and linked objects (final step). Interim steps are hidden, offline run and implement the algorithm of interaction between the system components without active participation of a user.

The main functional components:

- Search module. It involves executing a search query in the Internet search systems and the custom directory of innovative solutions; basic search (query by attributes and full-texts), location, data retrieval and summarizing.
- Query qualification module. Selection and ranking of search results: filtration, subject control, qualification of search query.
- Classification module. Classification of search results: selection of a method, cluster analysis of text documents and multimedia objects, data qualification. As a result we obtain a subset of semantically linked data.
- Link identification module. Link start-up: qualitative classification assessment, selection of the best results, interpretation of results; generating the descriptions of solutions with innovative potential in a given subject segment or for a specified object (article, technology, product).
- Visualization module. It involves mapping of search results, procedures of data processing, classification results including semantic links between objects.
- Data warehouse (DW) management module involves storage and updating of data search and processing results, parameters, and intermediate data; registry of innovative scientific, technological and educational problems. DW is built on the basis of a vector space model, includes document database access libraries and a data indexer.
- Service module. It involves monitoring and analysis of user access to information resources.

It is particularly remarkable that the developed original object model is oriented to work with any text objects related to the subject of the processing: queries, search results, text documents. Over 30 entity classes specify a document processing environment, a set of documents, methods of calculating the package document similarity measures as well as search functions in document package, types of reports, a collection of document words, lemmatization, a document structure and its specific parts.

The detailed architectural solutions are described in [21].

VI. General Search Algorithm

One of the elements of the presented above architecture is a generalized heuristic algorithm for filtering and rating the search results, which is based on available search engines; the algorithm is supposed to provide a background for

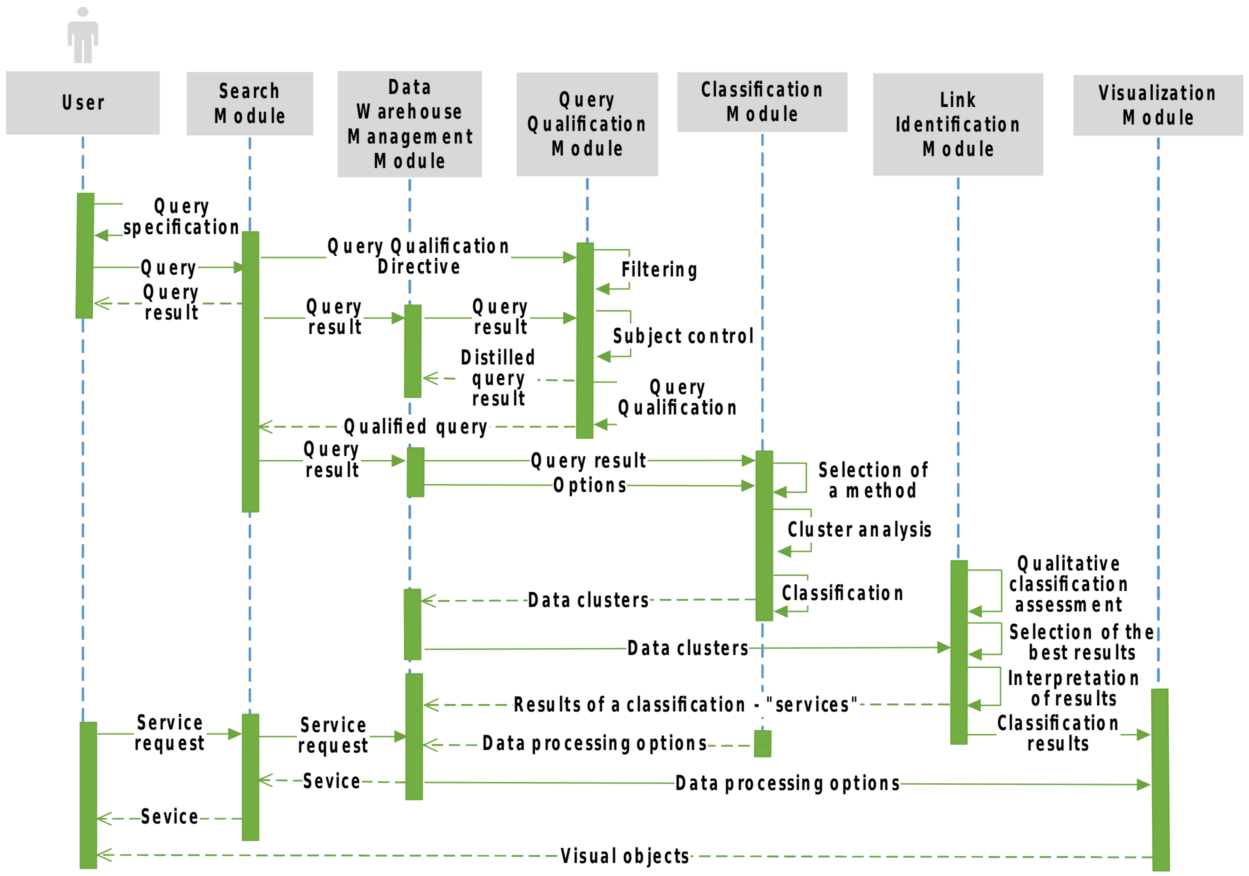


Fig 1. Behaviour of software components

search modules and inquiry qualification, as well as for retrieval schedule and search procedures in general.

The algorithm under consideration uses search results of known search engines being in service; it is invariant to them; with various degree of automation; it uses the search engine rating results.

The algorithm instruments a multistep process of sequential filtrations of search results and the analysis of semantic similarity of the found object content to adaptively generated reference texts (*k*-patterns). Ranking quality of the filtered search results made as per algorithm was estimated by *DCG* metric (see below). The ways of generating effective *k*-patterns were investigated as well.

Let us briefly run through the algorithm operation (Fig. 2). The description of a generalized request Q_o includes the initial set of key concepts of the target document subject.

The generation of the set Q of search queries $q \in Q$ and $|Q| = N$ is automated with an adaptive genetic algorithm searching for an effective total pertinence of the resulting document sampling under given evolutionary process depth constraints (see below).

The execution of queries q_{ij} is accompanied with filtering search results R_{qs} rated by a search engine and generating total results R . Filtering provides for the exclusion of some documents which subject area is formally pertinent but should not be the subject of the search for some reasons. It is done by hand or with a classifier which learning set is updated during the analysis of found texts.

The examples of documents being filtered are tutorials, student's papers, training programs, tests and notes, site promotion materials, company's sites, shopping sites, social networking sites; blogs; advertisements; virus-infected resources; nonexistent resources. The generation of *k*-patterns or reference texts is done simultaneously. They are used for calculating document similarity measures (P_{ka} is a text combination based on the first positions of rated search results, P_{kc} is the most pertinent result, P_{kb} is the text constructed from authority dictionary entries and P_{kd} is a text constructed from Q_o). Further the model of document vector space is used, i.e. each document d (the search results from R and *k*-patterns) is interpreted as vector $\bar{v}(d) = (w_{1,d}, w_{2,d}, w_{N,d})$, where $w_{t,d}$ is determined with a common metric $tf_{t,d} * idf_{t,d}$. A matrix $M_{N,x4}$ of document semantic similarity from set R with common *k*-patterns is generated and the rating of docu-

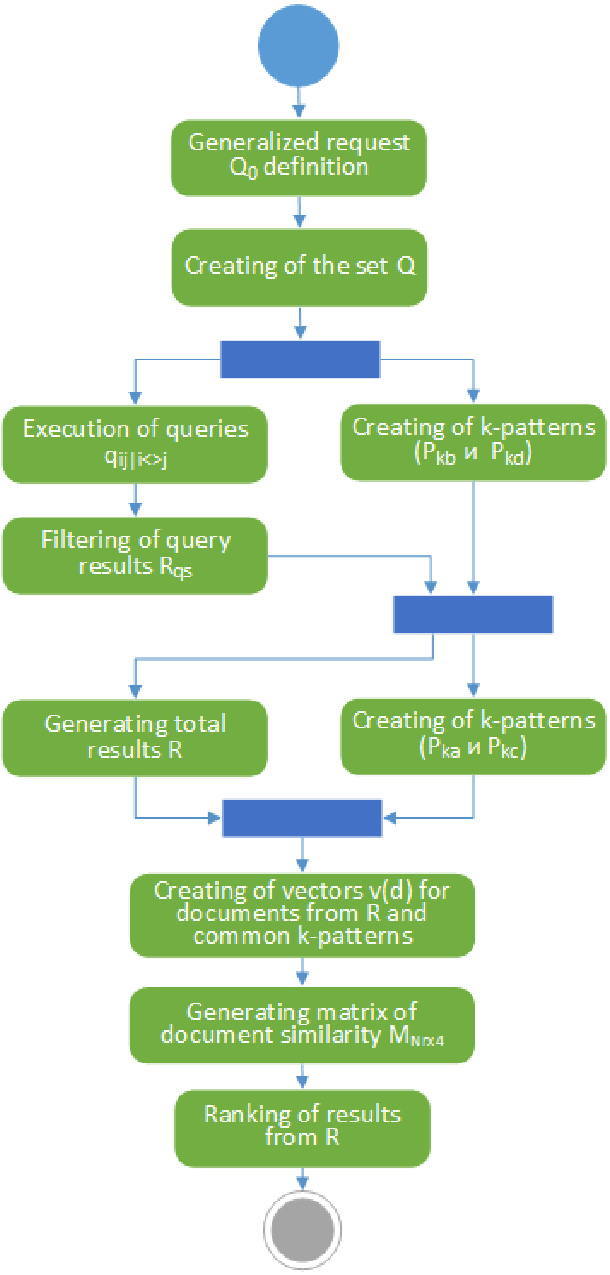


Fig 2. General pattern of a generalized heuristic algorithm for filtering and rating the search results

ments from *R* in accordance with their similarity $Similarity(d_1 d_1)$ to *k*-patterns is done.

The algorithm and some results of its use in patent searching are described in detail in [22]. To assess the algorithm quality, *DCG* metric [23] was used. For documents arranged by semantic similarity to $Similarity(d_1 d_1)$, the values were calculated for every *k*-pattern:

$$DCG=\sum_{p=1}^{n} 2^{gr(p)}-1/\log_2(2+p) \quad ,$$

where $gr(p)$ is mean expert relevance assessment given to the document located on *p* position in the list of results,

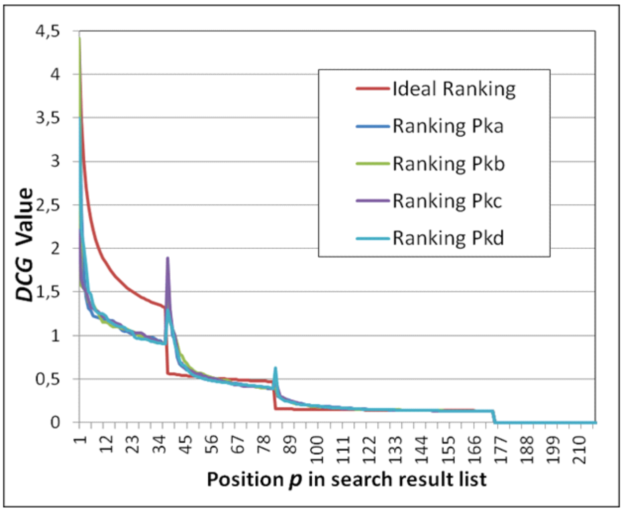


Fig 3. *DCG* values of ideal ranging and various *k*-patterns.

$gr \in [0,3]10$ with 3 standing for "relevant", 0 – "irrelevant", 1 and 2 – "partially relevant" ("relevant (+)" or "relevant"); $1/\log_2(2+p)$ – document position discount (the documents at the head of the list are of greater importance).

Fig. 3 shows the ratio of *DCG* metrics of ideal ranging and various *k*-patterns. Good agreement of metric values is observed in various patterns. At the same time there are reserves for more exact labeling of documents by relevance groups *gr*. The algorithm under consideration labels 10-15% of documents as a group with value *gr* which is different from ideal ranging (see peaks of breakpoints). Then normalized values *NDCG=DCG/Z* were calculated for every *k*-pattern, with *Z* being equal to the greatest possible *DCG* value in case of ideal ranging according to the expert assessment. Indicator *NDCG* assumes values from 0 to 1. The ratio of *NDCG* values for generalized query and various *k*-patterns is presented in Fig. 4. It is evident that algorithm shows the best results under *k*-pattern P_{ka} (combination of texts from the first positions of the ranked search results).

Note that the project provides for the usage of Internet search engine work results. The proposed search algorithms will be added to authoritative decisions – classical approaches to search result ranking (HITS, PageRank, BrowseRank, MatrixNet) which are based on the combination of document semantic pertinence and authority as well as user’s behaviour and experience.

VII. Generation of Search Queries

The project proposes and investigates the approach to search result generation based on a genetic algorithm (Fig. 5). The approach is used to specify a semantic kernel of a document desired set and generate sets of effective queries. The problem definition provides for the organization of an evolutionary process generating a stable and effective query population forming a relative search image of a document. A target set of search results is to be formed by such document addresses which are (a) in the first positions of a ranked list constructed by a search engine; (b) present in the result lists of multiple queries; (c) semantically similar to reference texts generated during evolutionary queries; (d) adequate to the environment given to a crawler by a user profile.

The original population from N search queries may be a set of $Q=\{q_i\}$, $|Q|=N$, $N<|Q_0|/2$, $q_i=(k_1,k_2,\dots,k_m)$, where $(k_1,k_2,\dots,k_m)$ is a random combination of key concepts of a search image Q_0. The value of an objective function must determine the query quality (population individual fitness). For each *i*-th query result the value may be calculated as $w_i(f,p,s,a)$, where f is determined by a result position in a ranked result list made by a search engine; p is determined by entering the result in the result lists of most queries; s is determined by a semantic similarity to *k*-patterns formed adaptively during the algorithm execution; a is determined by a user profile as an environment factor (values f,p,s,a are normalized for the range from 0 to 1). The value of a target function for each query is calculated as an averaged weight of query results, where w_i is a weight of each result calculated after executing all queries; P is a number of document addresses seen as the query result. The value of a objective function is interpreted as the capability of a search query to generate the results to be in the next population generation.

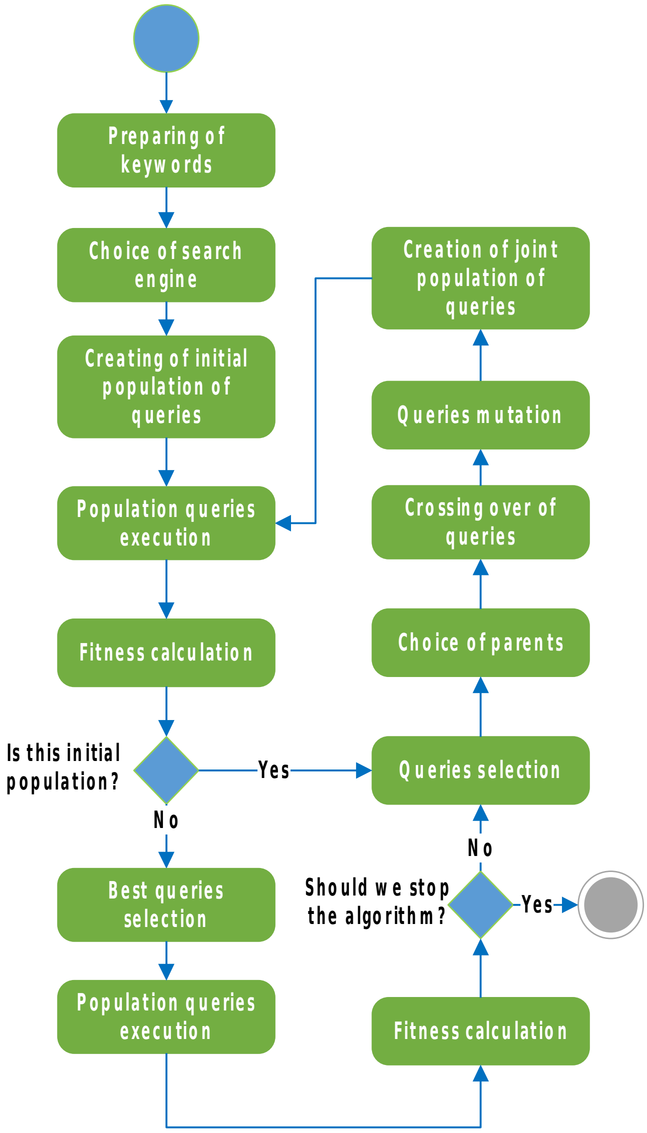

Preparing of keywords
Choice of search engine
Creating of initial population of queries
Population queries execution
Fitness calculation
Is this initial population?
Yes
No
Best queries selection
Population queries execution
Creation of joint population of queries
Queries mutation
Crossing over of queries
Choice of parents
Queries selection
No
Should we stop the algorithm?
Yes
Fitness calculation

Fig 5. Genetic algorithm for search result generation.

To choose parent couples the method of genotype outbreeding is proposed. It can provide for the most complete participation of all current queries in generating the next query population (the first parent individual is chosen randomly and the second individual is the "farthest" from the first one, the distance can be calculated as $\hat{w}=\bar{w}_1-\bar{w}_1$). The evolutionary operator of crossover is done with discrete recombination which corresponds to the exchange of key words (genes) between queries. The peculiarity of the proposed implementation is that the key word of a parent query is not substituted for the other parent query key word but its synonym. It allows generating considerably more child queries, with properties (semantics) of parent queries being preserved.

The essence of the most adequate mutation operation of the approach under study is the probabilistic change of a key query word (gene) chosen randomly. Because if the number of key words in a query $q_i=(k_1,k_2,\dots,k_m)$ is fixed, then it is not possible to use such mutation operators as a new gene addition, new gene insertion, gene deletion. Otherwise, we can doing it. Besides, there is no sense in gene place exchange in the context of executing search queries.

To generate a new population an elite selection denying the loss of best solutions is used. An intermediate population is generated. It includes both the parents and their children. N with the best values of a objective function $\hat{w}$ is chosen from all the population members. they will go in the next population. Generally, the condition of terminating the algorithm is considered to be population stability. For example, when a mean-square deviation of fitness function $\bar{W}$ reaches some threshold specified by an algorithm parameter. The genetic algorithm is described in detail in [24].

Some results of preliminary experimental studies of algorithm are briefly described below. The developers used original software support, search engine Bing and the following initial values of key parameters: *N*=15, *m*=3, *P*=10|50; number of search results returned after ranking all the results - 50. Weights of arguments *ƒ*, *p* and *s* in search ranking were taken as equal to each other. The science of calculation of fitness function for groups of results is average value; the algorithm exit strategy is given number of passes. Terms from the document corpus (students' papers) were used in the origin collection .

Fig. 6 shows the plots of $\bar{W}$ against population number, with *P*=10 and $|Q_0|=50$. Local maxima $\bar{W}$ and points of relative stabilization $\bar{W}$ (the 6-7th population) can be observed.

Fig. 7 shows plots of $\bar{W}$ against number of keywords of every generated query *m* (shown as numbers beside

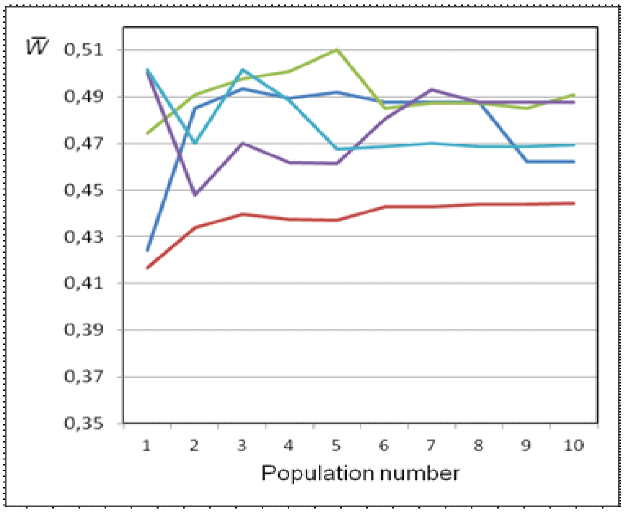


Fig 6. The plots of fitness function $\bar{W}$ against population number, $|Q_0|=50, m=3$.

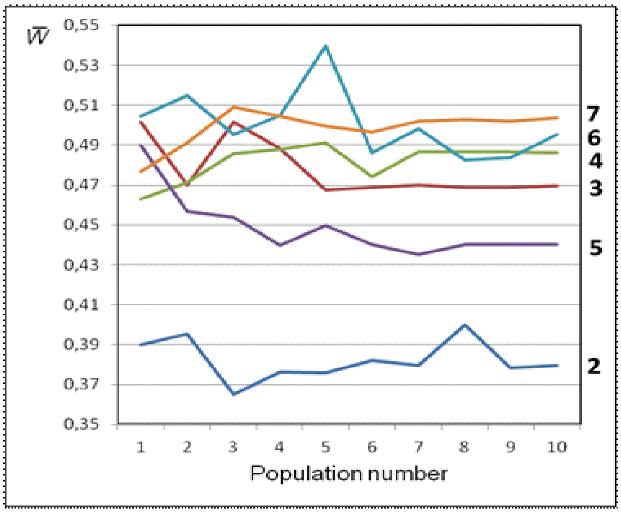


Fig 7. The plots of fitness function $\bar{W}$ against number of keywords of generated queries, $|Q_0|=50, m=2...7$

plots). It is seen that the increase of *m* leads, in the large, to the population quality improvement.

Fig. 8 shows the influence of fitness function arguments on its value. The greatest influence belongs to *ƒ*, the lowest one – to *p*.

VIII. Data Warehouse

The possibilities of Data Warehouse (DW) generation with realizing a document vector space model [25] to use it as a base of a data-centre information support are researched in the project. A software platform Document Text Analyzer (DTA) for semantic document analysis (their metric similarity computation) is developed within DW.

The prototype of the DW was tested successfully when associated technologies of the integral electronic document quality assessment and document pertinence in different contexts analysis were employed [26]. In particular, the debugging of software shell and interface of the TSTU specialized electronic teaching pack database, data centre warehouse components, was done (Fig. 9). The database is used to test and apply the project research results. The pioneering technology of the students' work uniqueness assessment (course and design-graphic papers, semester tasks, reports, essays, tests) is put to use.

The methods of semantic text comparison are used here. They are the computation of key concept weights and the construction of document vectors and not the known approaches (e.g. shingling) based on detecting direct adoptions in the text.

The research of some approaches to data centre different information systematization should be noted. As a result, a multistep algorithm of alternative search in an information catalogue with a target step number to be a base of a desired solution selection is developed [27].

IX. Application Areas

The list of application areas of the approaches under discussion in the paper, the research results and technologies is given below:

- A competitive analysis and competitive intelligence. A survey of commercial, scientific and technical, social information sources in a target field. A search of business valuable information. A client information acquisition (in CRM systems). A characterization of new fields and directions in business planning. A search of sector innovation decision descriptions.
- Educational technologies. An analysis of students' paper works (graduation, course papers), theses. A selection and expert examination of teaching materials (books, articles, papers, essays, surveys, etc., including web-resources). Scientometrical analytical services.
- The work of competition committees and sponsoring agencies. An expert examination in venture and other investment funds, the work of councils and groups of experts. An analysis of applications, information cards, competition documentation, expert examination rules and conditions. Normalizing and metrological control of technical documentation. An analysis of project design documentation, standards, norms, rules, regulations, manuals.
- Patenting, novelty expert examination. Materials selection for patent investigations. A documentation analysis of intellectual property objects, license contracts. Technological development forecasting.
- A content analysis of document texts in sociological surveys.
- Staff recruitment at enterprises and in organisations. An analysis of applicants' resumes vacancy descriptions.
- Rubrication of personal digital documents. PC text document (files) classification and grouping.

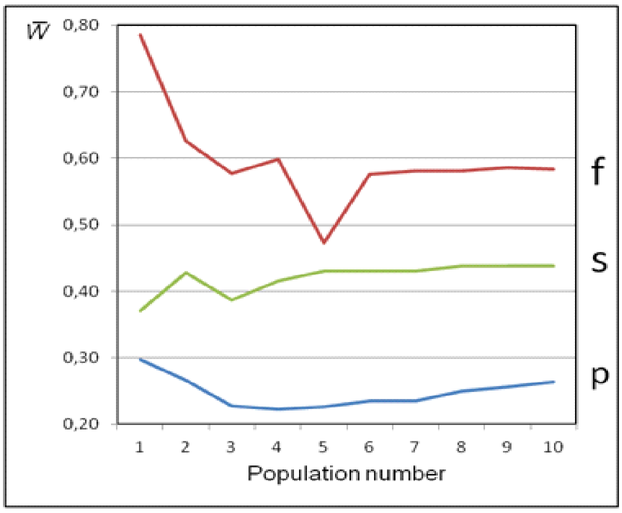
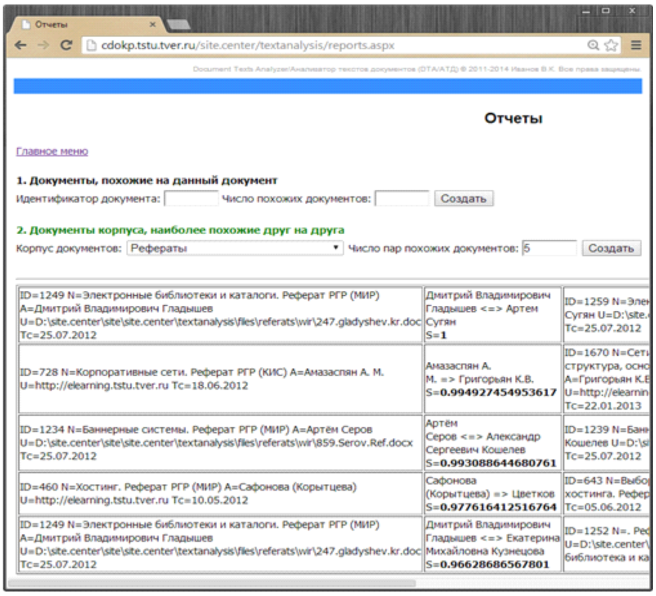


Fig 8. The influence of fitness function arguments on $\bar{W}$ value.

It should be noted that the project made some patent research which aim was to find analogs of the system designed and establish its novelty. At the moment of the research result report preparation any data of direct project analogs or its components realized are not discovered. The search of the Federal Institute of Industrial Property's document database did not show any matches of the project results with technologies recorded in official publications of the titles of protection.

X. Conclusion

One of the R&D management reference models include a competitive analysis and technological development forecasting based on scientometrical analytical services and semantical systems of business valuable information search. A relatively new world trend is evident: an effective use of global knowledge dataflow. With all the differences the major search pattern is selecting materials on demand, highlighting key concepts in the desired area and grouping materials respectively, filtering and semantic result processing, generating analytical reports. In this sense, the project tasks the results of which were discussed in the article are timely and urgent, and on the appropriate level of the problem interpretation.



Document Texts Analyzer/Анализатор текстов документов (DTA/АТД) © 2011-2014 Иванов В.К. Все права защищены.

Отчеты

Главное меню

1. Документы, похожие на данный документ

Идентификатор документа: Число похожих документов:

2. Документы корпуса, наиболее похожие друг на друга

Корпус документов: Рефераты Число пар похожих документов: 5

ID=1249 N=Электронные библиотеки и каталоги. Реферат РГР (МИР) A=Дмитрий Владимирович Гладышев U=D:\site.center\site\site.center\textanalysis\files\referats\wir\247.gladyshev.kr.doc Tc=25.07.2012	Дмитрий Владимирович Гладышев <=> Артем Сугян S=**1**	ID=1259 N=Элек Сугян U=D:\site. Tc=25.07.2012
ID=728 N=Корпоративные сети. Реферат РГР (КИС) A=Амазаспян А. М. U=http://elearning.tstu.tver.ru Tc=18.06.2012	Амазаспян А. М. => Григорьян К.В. S=**0.994927454953617**	ID=1670 N=Сети структура, осно A=Григорьян К.Е U=http://elearnin Tc=22.01.2013
ID=1234 N=Баннерные системы. Реферат РГР (МИР) A=Артём Серов U=D:\site.center\site\site.center\textanalysis\files\referats\wir\859.Serov.Ref.docx Tc=25.07.2012	Артём Серов <=> Александр Сергеевич Кошелев S=**0.993088644680761**	ID=1239 N=Бан Кошелев U=D:\si Tc=25.07.2012
ID=460 N=Хостинг. Реферат РГР (МИР) A=Сафонова (Корытцева) U=http://elearning.tstu.tver.ru Tc=10.05.2012	Сафонова (Корытцева) => Цветков S=**0.977616412516764**	ID=643 N=Выбор хостинга. Рефер Tc=05.06.2012
ID=1249 N=Электронные библиотеки и каталоги. Реферат РГР (МИР) A=Дмитрий Владимирович Гладышев U=D:\site.center\site\site.center\textanalysis\files\referats\wir\247.gladyshev.kr.doc Tc=25.07.2012	Дмитрий Владимирович Гладышев <=> Екатерина Михайловна Кузнецова S=**0.96628686567801**	ID=1252 N=. Ре U=D:\site.center\ библиотека и ка

Fig 9. The appearance of one of the reports of experimental software platform DTA

References

[1] National Science Board. 2010. Science and Engineering Indicators 2010. Chapter 4. Research and Development: National Trends and International Linkages. Arlington, VA: National Science Foundation (NSB 10-01), 66 p.

[2] Illumin8. A powerful research tool for innovation & product development. URL: http://www.illumin8.com.

[3] NetBase Social Media Management System (SMMS). URL: http://www.netbase.com.

[4] Questel Intellectual Property Portal. URL: http://www.orbit.com.

[5] R&D Management Framework. URL: http://kalypso.com/rd.

[6] Community Research and Development Information Service (CORDIS). URL: http://cordis.europa.eu.

[7] Nambisan S. (ed.). Information Technologies and Product Development, Annals of Information Systems, Vol. 5, 218 p. DOI 10.1007/978-1-4419-1081-3_1.

[8] Song L.Z., Song M. The Role of Information Technologies in Enhancing R&D–Marketing Integration: An Empirical Investigation. Journal of Product Innovation Management, Vol. 27, Issue 3, pp. 382–401, May 2010. DOI: 10.1111/j.1540-5885.2010.00723.x.

[9] Hanani U., Shapira B., Shoval P. Information Filtering: Overview of Issues. Research and Systems, User Modeling and User-Adapted Interaction II: pp. 203-259. 2001.

[10] Fensel D., Patel-Schneider P.F., Layering the Semantic Web: Problems and Directions. ISWC'02 Proceedings of the First International Semantic Web Conference on The Semantic Web, pp. 16-29, 2002, DOI: 10.1007/3-540-48005-6_4.

[11] Dong H., Hussain F.Kh., Chang E. Semantic Web Service matchmakers: state of the art and challenges. Concurrency and Computation: Practice and Experience, Vol. 25, Issue 7, pp. 961–988, May 2013, DOI: 10.1002/cpe.2886

[12] Manning C. D., Raghavan P., Schütze H.. Introduction to information retrieval. Cambridge University Press, Cambridge, England, 2008, 482 p.

[13] Broder A. A taxonomy of web search. ACM SIGIR Forum Vol. 36, Issue 2, Fall 2002, pp. 3 – 10, DOI: 10.1145/792550.792552.

[14] Zhu M., Xu C., Wu Y.-F.B. IFME: information filtering by multiple examples with under-sampling in a digital library environment. JCDL'13 Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, pp. 107-110, DOI: 10.1145/2467696.2467736.

[15] Cremonesi P., Koren Y., Turrin R. Performance of recommender algorithms on top-n recommendation tasks. RecSys'10 Proceedings of the fourth ACM conference on Recommender systems, pp. 39-46, DOI: 10.1145/1864708.1864721.

[16] Wu J., Ilyas I., Weddell G. A Study of Ontology-based Query Expansion, Cheriton School of Computer Science, University of Waterloo, Technical Report CS-2011-04, February 09, 2011, p. 38

[17] Bendersky M., Metzler D., Croft W. B. Effective query formulation with multiple information sources. WSDM'12 Proceedings of the fifth ACM international conference on Web search and data mining, pp. 443-452, DOI: 10.1145/2124295.2124349.

[18] Sevcech, J., Bielikova, M., Query Construction for Related Document Search Based on User Annotations. Proceedings of the Second International Federated Conference on Computer Science and Information Systems (FedCSIS), Krakow, 8-11 September, 2013, pp. 279-286.

[19] Ageev M., Guo O., Lagun D., Agichtein E. Find it if you can: a game for modeling different types of web search success using interaction data. SIGIR'11 Proceedings of the 34th international ACM SIGIR

conference on Research and development in Information Retrieval, pp. 345-354, DOI: 10.1145/2009916.2009965.

[20] Ivanov, V.K., Palyukh, B.V., Sotnikov, A.N. Intelligent subject search support in science and education. Innovative Information Technologies : Materials of the III International scientific-practical conference. Part 2. Innovative Information Technologies in Science / Ed. S.U. Uvaysov. - pp. 34-40. - M., 2014.

[21] Ivanov V.K., Palyukh B.V., Sotnikov A.N. Arkhitektura intellektual'noy sistemy informatsionnoy podderzhki innovatsiy v nauke i obrazovanii // Programmnyye produkty i sistemy. – Tver', 2013. – № 4. – pp. 197-202.

[22] Ivanov V.K., Vinogradova N.V. Evristicheskiy algoritm fil'tratsii i semanticheskogo ranzhirovaniya rezul'tatov poiska dokumentov // Vestnik Tverskogo gosudarstvennogo universiteta: Seriya "Prikladnaya matematika" / № 41. –Tver', 2013. – № 3. – pp. 97-107.

[23] Järvelin K., Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS), Vol. 20, issue 4, October 2002, 422-446, DOI: 10.1145/582415.582418.

[24] Ivanov V.K. Osnovnyye shagi geneticheskogo algoritma fil'tratsii rezul'tatov tematicheskogo poiska dokumentov: stat'ya // Innovatsii v nauke. – Novosibirsk, 2013. –№ 25. – P. 8-15.

[25] Salton G., Wong A., Yang C.S. A Vector Space Model for Automatic Indexing. Communications of the ACM. - 1975. - Vol. 18, nr. 11. - pp. 613–620.

[26] Ivanov V.K., Mironov V.I. Osobennosti analiza skhodstva dokumentov v razlichnykh kontekstakh zaimstvovaniya pri podgotovke tekstovykh materialov // Otsenka kachestva vysshego professional'nogo obrazovaniya s uchetom trebovaniy FGOS i professional'nykh standartov: materialy dokladov – Tver', 2013. – pp. 19-26.

[27] Paliukh, B., Egereva, I. Multistep algorithm of alternatives search in an information catalogue // 10th International Conference on Interactive Systems: Problems of Human-Computer Interaction. – Collection of scientific papers. – Ulyanovsk : USTU, 2013. – pp.. 129-132.

Position papers of the 2014 Federated Conference on Computer Science and Information Systems pp. 21–28

DOI: 10.15439/2014F391
ACSIS, Vol. 3

Solving Graph Coloring Problem with Parallel Evolutionary Algorithms in a Mesh Model

Zbigniew Kokosiński and Piotr Domagała
Cracow University of Technology
Faculty of Electrical & Computer Engineering
ul. Warszawska 24, 31-155 Kraków, Poland
Email: zk@pk.edu.pl

***Abstract*—In this paper a parallel evolutionary algorithm (PEA) for coloring graph vertices is investigated. In the algorithm we apply a diffusion model of parallelism (DM). Evolutionary computations are performed in a regular mesh with either a constant size global population or a constant subpopulation in a single node. The performance of the PEA-DM is verified by computer experiments on standard DIMACS graph coloring instances. For recombination well known crossover and mutation operators are chosen. Selection mechanisms include standard roulette and tournament. The results obtained by PEA-DM are compared with a classical evolutionary algorithm. It is possible to define dimensions of the rectangular mesh and two types of additional local connections: boundary enclosures (cyclic mesh) and diagonal links. The problem of optimal selection of the mesh configuration as well as global population and subpopulation sizes is adressed.**

I. Introduction

GRAPH k–colorability problem belongs to the class of NP–hard combinatorial problems [14], [20]. This decision problem is defined for an undirected graph $G = (V, E)$ and positive integer $k \leq |V|$: is there an assignment of available k colors to graph vertices, providing that adjacent vertices receive different colors ? With additional assumptions many variants of the coloring problem can be defined such as equitable coloring, sum coloring, contrast coloring, harmonious coloring, circular coloring, consecutive coloring, list coloring etc. [18], [27]. In optimization version of the basic problem called GCP, a conflict–free coloring with minimum number of colors is searched. Intensive research conducted in this area resulted in a large number of exact and approximate algorithms, heuristics and metaheuristics [5], [26], [31], [36], [37]. However, the reported results are often difficult to compare due to specific assumptions, different algorithms and their implementation details, tuning of parameters, computing platforms, test data sets etc. GCP was the subject of Second DIMACS Implementation Challenge held in 1993 [19] and Computational Symposium on Graph Coloring and Generalizations in 2002. A collection of graph coloring instances in DIMACS format and summary of results are available at [38], [39], [40].

Evolutionary algorithm (EA) is a metaheuristic often used for GCP [10], [11], [12], [13], [21], [30], [32], [35]. Parallel versions of evolutionary algorithms (PEAs) were reviewed in [1]. One popular model is master–slave in which master processor assignes portions of computations to slave processors [6]. Another approach is based on co-evolution of a number of populations that exchange the genetic information during the evolutionary process according to a communication pattern [2], [3], [9]. That approach includes migration and diffusion models od PEAs.

PEA were applied to many hard optimization problems (e.g. [1], [28]). Parallel metaheuristics for GCP and related problems were recently published in [23], [24], [25], [31].

The main purpose of the master–slave model is speeding up processing of one global population by parallelization of computations. In the migration model of PEA the model of interactions between co–evolving populations can affect the quality of the solution [23]. If speedup is not essential one can simulate and test migration–based PEA with the help of a sequential program. An alternative parallel model is the diffusion model, where evolution takes place among neighbouring subpopulations. There are many possible neighbourhood patterns. Among them 2–dimentional mesh with subpopulations placed in its nodes is a typical option because it corresponds to a model of computations very popular in parallel procesing, i.e. 2D array of processors.

In this article some results of experiments with PEA in diffusion model for graph coloring problem are described. In the paper two recombination operators for coloring chromosomes are used: CEX (Conflict Elimination Crossover) [23] that reduces the number of color conflicts with the help of selective copy operations and GPX (Greedy Partition Crossover) [13]. They both are problem–specific crossovers, designed particularly for GCP and passed a series of experimental verification in EA environment [15], [22].

In experimental part of the paper, widely accepted DIMACS benchmark graphs were used. The obtained results show adventages and limitations of PEA in the diffusion mesh model for hard optimization problems like GCP.

II. Graph Coloring Problem

Let us define formally the optimization problem GCP.

For given graph $G(V, E)$, where : V — set of graph vertices, $|V| = n$, and E — set of graph edges, $|E| = m$, the optimization problem GCP is formulated as follows: find the minimum positive integer k, $k \leq n$, and a function $c : V \longrightarrow \{1, \ldots, k\}$, such that $c(u) \neq c(v)$ whenever

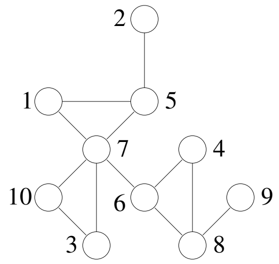


Fig. 1. Exemplary graph G(V,E)

$(u, v) \in E$. The obtained value of k is refered to as graph chromatic number $\chi(G)$.

Similarly, graph edge coloring problem for given graph $G(V, E)$ can be defined. One can find solution to mimimum edge coloring by solving vertex coloring problem for edge graph $G_e(V_e, E_e)$ associated with the given graph $G(V, E)$ [18], [26]. An exemplary graph $G(V, E)$ with ten vertices is shown in Fig.1.

In graph coloring problem *k*–colorings of graph vertices are encoded in chromosomes representing set partitions with exactly *k* blocks. There are two equivalent notations for vertex colorings that are commonly used in algorithm design.

In *assignment* representation available colors are assigned to an ordered sequence of graph vertices. Thus, the vector c=<c[1],c[2], . . . ,c[n]>. represents a vertex coloring. For the graph in Fig.1, an optimal 3–coloring is denoted by vector c=<1,2,3,2,3,1,2,3,2,1>.

In *partition* representation a vertex coloring is a unique sequence of partition blocks in Hutchinson representation [16]. Each block of partition p does correspond to a single color. Elements inside each block are ordered in increasing lexicographic order, and all blocks are ordered increasingly according to the value of their first element. For our graph the same optimal 3–coloring is denoted by partition p={1,6,10}{2,4,7,9}{3,5,8}.

III. Models of Parallel Evolutionary Algorithms

There are many models of parallelism in evolutionary algorithms: master–slave PEA, migration based PEA, diffusion based PEA, PEA with overlaping subpopulations, population learning algorithm, hybrid models etc. [4], [6], [7], [8], [17], [29], [33], [34].

The above models are characterized by the following criteria:

- number of populations : one, many;
- population types : disjoint, overlaping;
- population topologies : various graph models;
- interaction model : isolation, migration, diffusion;
- recombination, evaluation of individuals, selection : distributed/local, centralized/global;
- synchronization on iteration level: synchronous/asynchronous algorithm.

The most common models of PEA are:

- master-slave : one global population, global evolutionary operations, fitness functions computed by slave processors);
- massively parallel (cellular): static overlapping subpopulations with a local structure, local evolutionary operations and evaluation;
- migration (with island as a submodel): static disjoint subpopulations/islands, local evolutionary operations and migration;
- diffusion (with mesh as a submodel); static disjoint subpopulations/nodes, local evolutionary operations and migration;
- hybrid : combination of one model on the upper level and other model on the lower level (the speedup achieved in hybrid models is equal to product of level speedups).

A. Migration Model of Parallel Evolutionary Algorithm

Migration models of PEAs consist of a finite number of disjoint subpopulations that evolve in parallel on their "islands" and only occasionally exchange evolutionary informations under control of a migration operator. Co–evolving subpopulations are built of individuals of the same type and are ruled by one adaptation function. The selection process is decentralized.

In the model the migration is performed on a regular basis. During the migration phase every island sends its representatives (emigrants) to all other islands and receives the representatives (immigrants) from all co–evolving subpopulations. This topology of migration reflects so called "pure" island model. The migration process is fully characterized by migration size, distance betweeen populations and migration scheme. Migration size determines the emigrant fraction of each population. This parameter is limited by capacity of islands to accept immigrants. The distance between migrations determines how often the migration phase of the algorithm occurs. Three migration schemes may be applied: no migration, migration of randomly selected individuals and migration of best individuals of the subpopulation.

B. Diffusion Model of Parallel Evolutionary Algorithm

The diffusion model of PEA is a fine-grained EA [6] with its global population distributed on a 2D mesh of size $w \times z$, where subpopulations are placed in the mesh nodes (cells). The crossover operation is local in that sense, that the recombined chromosomes are members of subpopulations in the closest neighbourhood, i.e. have distance 1 in the underlying graph G(V, E). Solutions with outstanding fitness function are able to "diffuse" step by step in the graph across the whole global population. The implementation of PEA-DM in diffusion model can be either synchronous or anynchronous on a parallel machine with shared memory. Parent chromosomes can be selected for recombination at random, but it is recommended to choose at random only the first parental population, all next should be selected on the basis of the fitness function of all chromosomes in the neighbourhood .

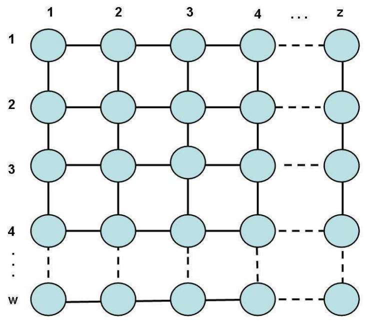


Fig. 2. A simple rectangular mesh of size $w \times z$.

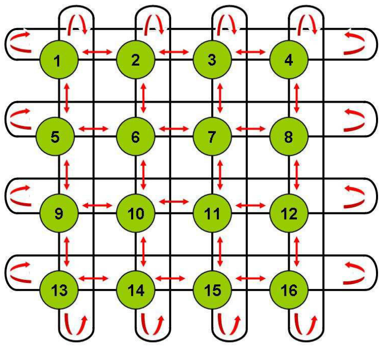


Fig. 3. A cyclic square mesh of size 4×4.

In the simple mesh model presented in Fig. 2 graph edges represent connections between nodes. Boundary nodes have a limited communication ability, because they have smaller vertex degrees then internal nodes of the graph G(V, E). For instance, for the square mesh of size 4×4 the mesh containes 75 % of boundary nodes. Similarly, for the mesh 5×5 there is 64 % of boundary nodes, for the mesh 8×8 there is 43 % of boundary nodes, for the mesh 10×10 there is 36 % of boundary nodes, for the mesh 15×15 there is 25 % of boundary nodes, for the mesh 20×20 there is 19 % of boundary nodes, etc. The lowest degree have the four corner nodes.

It is reasonable to expect that the mesh size and node degree in the graph shall influence the PEA-DM performance. Propagation time of good solutions across the whole network increases with the network size. The simplest way to eliminate node degree irregularities in the ordinary mesh is to add cyclic connections to boundary nodes (boundary enclosures), The resulting cyclic mesh 4×4 is depicted in Fig. 3.

The structure of 2D mesh is very popular since it can be implemented in MIMD computers. The sizes of the mesh can be variable within certain ranges. The number of communication chanels for each node processor is 4. That value is constant and does not depend on the number of processors. Maximum distance between processors in the cyclic mesh is 0,5(w+z).

In order to increase node degree in the cyclic mesh an additional connections are to be added. In Fig. 4 extra diagonal connections are shown. In this way degrees of internal nodes increase from 4 to 8 and degrees of boundary nodes increase from 4 to 6, except the corner nodes which degrees increase only by one, from 4 to 5. It is possible to complete the network connectivity by adding diagonal cycles for boundary nodes. In such architecture, for the regularity of the extended cyclic mesh, we pay in longer iteration time of PEA-DM.

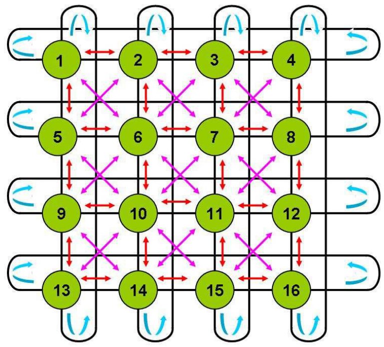


Fig. 4. A cyclic square mesh of size 4×4 with extra diagonal connections.

Algorithm 1 EA for a subpopulation in diffusion model

```
Require: cell position in the mesh, subpopulation size and chromosome parameters
Ensure: best coloring in $P_{t+1}$
  $t \leftarrow 0$; [reset iteration counter]
  initialization of subpopulation $P_t$;
  evaluation of subpopulation $P_t$;
  while not termination condition do
      parents $T_t \leftarrow$ selection from $P_t$ and its neighbourhood;
      offspring population $O_t \leftarrow$ crossover and mutation on $T_t$;
      evaluation of $\{P_t \cup O_t\}$;
      $P_{t+1} \leftarrow$ selection from $\{P_t \cup O_t\}$;
      $best \leftarrow$ best coloring in $P_{t+1}$;
      $t \leftarrow t + 1$;
  end while
```

IV. Recombination Operators

In this section a collection of crossover, mutation and selection operators is introduced that can be used in our PEA-DM. Two recombination operators: CEX, GPX and the mutation operator First Fit were designed especially for GCP. The mutation Transposition is more versatile. Other efficient recombination operators were proposed in [32]. The cost function is adapted for PEA-DM. All examples given in this section refer to the graph instance shown in Fig.1.

A. Conflict Elimination Crossover

In conflict–based crossovers for GCP the assignement representation of colorings is used and the offspring tries to copy conflict–free colors from their parents. The operator CEX (Conflict Elimination Crossover) reveals some similarity to the classical crossover. Each parental chromosome p and r is partitioned into two blocks. The first block consists of conflict–free nodes while the second block is built of the remaining nodes that break the coloring rules.

The last block in both chromosomes is then replaced by corresponding colors taken from the other parent. This recombination scheme provides inheritance of all good properties of one parent and gives the second parent a chance to reduce the number of existing conflicts. However, if a chomosome represents a feasible coloring the recombination mechanism will not work properly. Therefore, the recombination must be combined with an efficient mutation mechanism. As a result two chromosomes s and t are produced. The operator CEX is almost as simple and easy to implement as the classical crossover (see Algorithm 2).

Behaviour of the CEX crossover is shown in Example 1.

Example 1

Two parents represent different 5–colorings of a graph with 10 vertices i.e. sequences p=<5,2,**3**,**4**,1,**4**,2,5,1,**3**>, and r=<1,4,**5**,2,3,3,**5**,4,2,**5**>. Vertices with color confiicts are marked by bold fonts. Thus, the chomosome p has 6 vertices with feasible colors and 4 vertices with color conflicts while the chomosome r has 7 vertices with feasible colors and 3 vertices with color conflicts.

Replacing the vertices with color conflicts by vertices taken from the other parent we obtain the following two chromosomes: s=<5,2,**5**,2,1,3,2,5,1,**5**> and t=<1,4,**3**,2,3,3,2,4,2,**3**>. (see Fig. 5)

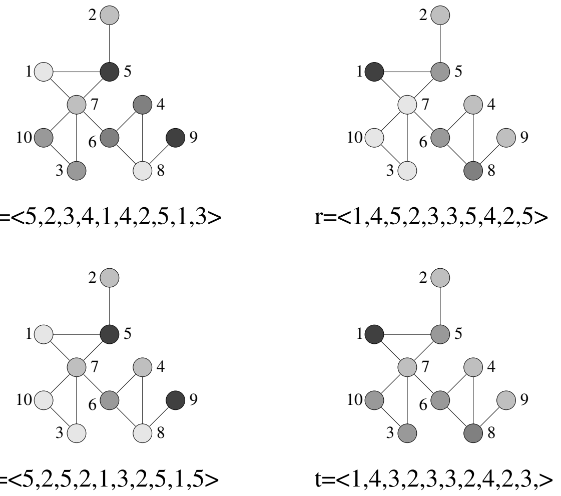
It is observed that obtained chromosomes represent now two different 4–colorings of the given graph (reduction by 1 with respect to initial colorings) and the number of color conflicts is now reduced to 2 in each chromosome.

B. Greedy Partition Crossover

The method called Greedy Partition Crossover (GPX) was designed by Galinier and Hao for recombination of colorings or partial colorings in partition representation [13]. It is assumed that both parents are randomly selected partitions with exactly k blocks that are independent sets. The result is a single offspring (a coloring or partial coloring) that is built successively in a greedy way. In each odd step select the maximum block from the first parent is selected. Then the block is added to the result and all its nodes from the both parents are removed. In each even step the maximum block is selected from the second parent. Then the block is added to the result and all its nodes from the both parents are removed. The procedure is repeated at most k times since in some cases the offspring has less blocks then the parents. This possibility is not considered in the original paper [13]. Finally, unassigned vertices (if they exist) are assigned at random to existing blocks of partition.

Algorithm 2 The crossover operator CEX

```
Require: $V$, $p$, $r$
Ensure: $s$, $t$
    $s \leftarrow r$; $t \leftarrow p$;
    $b \leftarrow x$;
    copy block of conflict-free vertices $V_{cf}^{p}$ from $p$ to $s$;
    copy block of conflict-free vertices $V_{cf}^{r}$ from $r$ to $t$;
```



p=<5,2,3,4,1,4,2,5,1,3> r=<1,4,5,2,3,3,5,4,2,5>

s=<5,2,5,2,1,3,2,5,1,5> t=<1,4,3,2,3,3,2,4,2,3,>

Fig. 5. An illustration of CEX crossover (see Example 1)

The first parent is replaced by the offspring while the second parent is returned to population and can be recombined again in the same generation. GPX crossover is performed with a constant probability.

C. Mutation Operators

Two types of mutation operators described in literature are used. Transposition (T) is a classical type of mutation that exchanges colors of two randomly selected vertices in the assignment representation. The second mutation operation called First Fit (FF) is designed for colorings in partition representation and is well suited for GCP. In First Fit mutation one block of the partition is selected at random and we try to make a conflict–free assignment of its vertices to other blocks using the heuristic First Fit. Vertices with no conflict–free assignment remain in the original block. Thus, as a result of the mutation First Fit the color assignment is partially rearranged and the number of partition blocks is often reduced by one.

D. Selection Operator

Selection process maintains constant size of population selected by means of a fitness function.

In the first phase of EA, when no initial information is available, the quality of a solution is measured by the following cost function:

$$cost(c) = conflicts(c) \cdot colors(c), \tag{1}$$

where:
c – is the current coloring,
$conflicts(c)$ – is the number of conflicts in c,
$colors(c)$ – is the number of colors in the c.

In the second phase of the algorithm, for conflict–free colorings $conflicts(c) = 0$ and $cost(c) = 0$. Therefore, in that case the cost function is computed by the following formula:

$$f(c) = conflicts(c) + colors(c) + p(c), \tag{2}$$

where the penalty function $p(c)$ equals:

$$p(c) = \begin{cases} 2 \cdot (colors(c) - best), & if \quad colors(c) \geq best \\ 0, & if \quad colors(c) < best \end{cases} \tag{3}$$

and:
$best$ – is a number of colors in the best individual so far.

The proportional (roulette) selection can be performed in two phases of the algorithm with the fitness function $1/f(p)$. Alternatively, the tournamet selection can be performed in both phases of the algorithm on randomly selected individuals from subpopulations with the analogous fitness function.

V. Computer experiments

For computer experiments several graph instances were used that are available in the web archives [38], [39]. They are collections of graphs in DIMACS format with known parameters m, n and usually $\chi(G)$.

In our program *PGA-DM for GCP* diffusion models of PEA can be simulated. It is possible to set up most parameters of evolution, monitor evolution process in each node and measure both the number of generations and time of computations. In order to avoid misunderstanding we always report throughout the paper the total execution time of the sequential simulation of the PGA. In the preprocessing phase we converted list of edges representation into adjacency matrix representation. The program generates detailed reports and basic statistics [22]. All computer experiments were performed on a computer with AMD Athlon 2000 processor (1,67 GHz, 512 MB RAM). The performance of the processor was never a critical factor.

The research was focused on the diffusion model of PEA. In the experiments the following aspects of this model were taken into consideration: 1. comparison of PEA-DM versus EA with respect to quality of solution, number of iterations and time of computation. 2. influence of mesh size for constant node population solution; 3. comparison of acyclic versus cyclic square meshes.

In all experiments and for all crossover operators we used constant crossover probability = 0.8 and mutation probability = 0.1.

A. Comparison of PEA-DM with EA

In the first experiment PEA-DM was tested against traditional EA. The jean(80, 254, 10) graph was used. Classical EA was obtained as a special case in the program PEA-DM for GCP with parameters: mesh size = 1×1, population size = 320, initial number of colors = 2, CEX crossover, FF mutation and Tournament selection. In diffusion PEA the mesh size = 8×8 and node subpopulation = 5. Computations were repeated 30 times. Termination condition was either optimal coloring for $\chi(G) = 10$ or the number of iterations = 1000. All results of the comparison were collected in Table I.

In all experiments a conflict–free coloring was reached for the given graph. Optimal colorings were more likely to happen with PEA-DM. Average time of computations for obtaining a conflict–free coloring was lower for EA. Execution times of 1000 iterations in PEA-DM and EA were close to each other: the average execution time for EA was 202,7 [s]; for PEA-DM it was 204,3 [s].

B. PEA-DM with a constant population size in a node

In the second experiment PEA-DM was investigated for various sizes of acclic square mesh, constant population size =5 in all nodes and initial number of colors = 2. Two DIMACS graphs were used for this configuration of the program PEA-DM : huck(74,301,11) and queen6.6(36,290,7). Computations were conducted with the following settings: random initial population, CEX crossover, FF mutation and Tournament selection. All experiments were repeated 10 times. Termination condition was either finding an optimal graph coloring ($\chi(G)$ is known for both graphs) or 5000 iterations completed. The computational results are presented in Table II (columns B, W, A contain the best, the worst, and the average results obtained in 10 experiments, respectively).

For huck graph and mesh size 4×4 only in 5 experiments (5/10) a conflict–free coloring was obtained and only 2 optimal colorings. For the mesh size 6×6 all colorings were conflict–free and half of them was optimal. For bigger mesh sizes a higher percentage of optimal coloring was obtained in a shorter time.

For queen6.6 graph conflict–free colorings were possible starting from the mesh size 6×6 (6/10). For bigger mesh sizes more conflict–free colorings were found with less colors used. However, due to graph difficulty the PEA-DM was not able to find any optimal solution with $\chi(G)$=7 colors. The minimal coflict–free coloring was found with the number of colors = 8.

In conclusion, the efficiency of the PEA-DM with constant population size in a node strongly depends on the size of the mesh. For small meshes small global population (rather tens then hundreds of individuals) can not provide sufficient diversity of solutions. The bigger mesh size implies in general more computations (longer processing time), but usually less iterations and sufficient diversity of population to achieve a satisfactory solution. The optimal mesh size for huck graph is 14×14. The other possibility to improve PEA-DM efficiency is to increase the population size in nodes. Thus, PEAs with small meshes and bigger population sizes in nodes become similar to PEAs in migration models, where the optimal number of populations is moderate. No mesh of optimal size was found for queen6.6 graph. For this graph the PEA efficiency can be improved with bigger subpopulations and higher number of iterations.

C. PEA-DM with a constant global population size

Taking into account results of the previous experiments the constant global populations size was set to 700, approximately. The are minor deviations from that size due to variable mesh

TABLE I
COMPARISON OF EA AND PEA-DM

graph $G(V,E)$	algo–rithm	colo–rings		cost / number of iterations / time			
				min	max	avg.	std. dev.
jean |V|=80 |E|=254 $\chi(G)$=10	EA	30/30 conflict–free	colors	10	13	11,23	0,68
			it	19	345	46,56	57,05
			t[s]	3,8	68,7	9,42	11,35
		2/30 optimal (7%)	colors	10	10	10	0
			it	37	39	38	1,41
			t[s]	7,0	7,7	7,35	0,50
	PEA-DM	30/30 conflict–free	colors	10	15	11,17	0,89
			it	26	471	180,1	216,4
			t[s]	7,0	94,7	36,56	0,68
		4/30 optimal (14%)	colors	10	10	10	0
			it	68	153	97,5	39,64
			t[s]	14,5	32,3	20,3	8,1

TABLE II
PEA-DM WITH CONSTANT POPULATION SIZE IN A NODE = 5

Graph	mesh size	4 × 4			6 × 6			8 × 8		
huck |V|=74 |E|=301 $\chi(G)$=11		B	W	A	B	W	A	B	W	A
	colors	11	13	12	11	13	11,6	11	12	11,4
	it	26	1440	336,8	5	4594	1022,3	3	1748	262,3
	t[s]	1,2	67,5	17,4	0,5	515,6	115,1	2,7	317,3	54
	mesh size	10 × 10			12 × 12			14 × 14		
		B	W	A	B	W	A	B	W	A
	colors	11	11	11	11	11	11	11	11	11
	it	6	74	38	4	75	32,7	8	29	16,6
	t[s]	1,6	20,9	11	1,4	32	14	5,2	16,6	9,6
	mesh size	16 × 16			18 × 18			20 × 20		
		B	W	A	B	W	A	B	W	A
	colors	11	11	11	11	11	11	11	11	11
	it	15	48	23,7	5	37	20,4	8	25	16,4
	t[s]	10,8	34,8	17,5	4,9	34,3	18,8	8,2	28,4	17,3
queen6.6 |V|=36 |E|=290 $\chi(G)$=7	mesh size	4 × 4			6 × 6			8 × 8		
		B	W	A	B	W	A	B	W	A
	colors	-	-	-	8	11	9,8	8	11	9,2
	it	-	-	-	49	4596	1990	10	4507	1529
	t[s]	-	-	-	2,2	185,5	89,5	0,6	319,5	108
	mesh size	10 × 10			12 × 12			14 × 14		
		B	W	A	B	W	A	B	W	A
	colors	8	10	8.9	8	10	8,8	8	9	8,5
	it	60	2933	638	7	3580	694,2	31	1725	762,4
	t[s]	6,2	319	69,2	1	552,7	197,5	6,3	360,2	150,3
	mesh size	16 × 16			18 × 18			20 × 20		
		B	W	A	B	W	A	B	W	A
	colors	8	9	8,3	8	9	8,1	8	8	8
	it	49	936	428,3	59	886	469,6	130	917	627,9
	t[s]	13	250,8	115,6	19,9	299,3	158,5	54,2	380,3	260,9

size. Two DIMACS graphs were used for this configuration of the program PEA-DM : games120(120, 638, 9) and david(87, 406, 11). Computations for games120 graph were conducted with the following settings: random initial population, CEX crossover, FF mutation and Tournament selection. All experiments were repeated 10 times. Termination condition was either finding an optimal graph coloring or 500 iterations completed. Both acyclic and cyclic meshes were considered. The computational results are presented in Table III (columns B, W, A contain the best, the worst, and the average results obtained in 10 experiments, respectively).

For games120 graph the optimal coloring were found for all acyclic mesh sizes. Only for the mesh 4×4 a single solution was not optimal. Changing the mesh size does not influence the computational time of PEA-DM. In average 500 iterations lasted about 450 [s]. Also the time of finding solutions was not influenced.

A variant of PEA-DM with 12×12 mesh and CEX crossover replaced by GPX was also tested for games120 graph. Unfortunately no conflict–free coloring was found, even with 1500 iterations.

For david graph the conflict–free colorings were found for all acyclic mesh sizes, but no optimal colorings were met. The best average results were obtained for 4×4 mesh (141,4 [s])

TABLE III
PEA-DM WITH APPROX. CONSTANT GLOBAL POPULATION SIZE = 700

Graph	population size	4 × 4 × 43=688			6 × 6 × 19=684			8 × 8 × 11= 704		
games120 |V|=120 |E|=638 $\chi(G)$=9	acyclic mesh	B	W	A	B	W	A	B	W	A
	colors	9	10	9,1	9	9	9	9	9	9
	it	25	482	221,5	46	297	149	68	397	211,1
	t[s]	22	433	232,1	35,5	236,6	117,9	61,7	359,7	190,5
	population size	10 × 10 × 7=700			12 × 12 × 5=720					
	acyclic mesh	B	W	A	B	W	A			
	colors	9	10	9,1	9	9	9			
	it	58	405	185	68	380	217			
	t[s]	44,7	314,6	152,2	63	352,5	201,5			
	population size	4 × 4 × 43=688			6 × 6 × 19=684			8 × 8 × 11= 704		
	cyclic mesh	B	W	A	B	W	A	B	W	A
	colors	9	10	9,2	9	9	9	9	9	9
	it	12	329	119,4	19	156	131,7	32	234	120,6
	t[s]	12,9	362,5	120	16,2	131,7	81,3	28,7	232,6	110
	population size	10 × 10 × 7=700			12 × 12 × 5=720					
	cyclic mesh	B	W	A	B	W	A			
	colors	9	9	9	9	9	9			
	it	87	307	167,2	44	306	178,6			
	t[s]	76,8	273	155	40	280	163			
david |V|=87 |E|=406 $\chi(G)$=11	population size	4 × 4 × 43=688			6 × 6 × 19=684			8 × 8 × 11= 704		
	acyclic mesh	B	W	A	B	W	A	B	W	A
	colors	13	15	13,7	13	14	13,9	14	15	14.1
	it	39	493	301,9	123	449	264	151	348	233
	t[s]	27	355	216,6	88,5	327	190,1	103	271	168,7
	population size	10 × 10 × 7=700			12 × 12 × 5=720					
	acyclic mesh	B	W	A	B	W	A			
	colors	14	14	14	14	15	14,4			
	it	121	476	197,7	142	364	225			
	t[s]	85,9	337	141,4	122	288,4	196,5			
	population size	4 × 4 × 43=688			6 × 6 × 19=684			8 × 8 × 11= 704		
	cyclic mesh	B	W	A	B	W	A	B	W	A
	colors	11	16	13,8	13	15	14,2	13	14	13,9
	it	56	400	193,4	94	496	223,3	82	426	205
	t[s]	41	397	160,5	63	317	148,8	64	336	161
	population size	10 × 10 × 7=700			12 × 12 × 5=720					
	cyclic mesh	B	W	A	B	W	A			
	colors	13	15	13,9	13	15	13,8			
	it	90	483	220,4	119	456	263,4			
	t[s]	73,4	392	179,3	105	396	229			

and the worst were received for 12 × 12 mesh (229 [s]).

The above experiments were repeated for cyclic meshes. It results, that in general a smaller number of iterations is sufficient for the same result. Only for david graph and 10×10 mesh the number of iterations increases but the smaller number of colors is received. In other cases the number of colors is similar for acyclic and cyclic meshes. It seems that using PEA-DM with cyclic meshes is advantageous, since the number of iterations decreases and the time of computations decreases too. One exception is david graph and cyclic 10×10 and 12×12 meshes when longer computations lead to colorings with lower number of colors. Computational time for performing the same number of iterations is longer for cyclic meshes.

One can expect that with extra diagonal connections the processing time for one iteration in cyclic meshes will increase, but on the other hand it is quite possible that the computed solutions will be closer to optimal in terms of the required number of colors.

VI. Conclusions

From the above experiments results that efficient computations with diffusion-based PEA in mesh model are obtained in configurations with relatively small cyclic meshes with sufficiently large global population what is very similar result as that obtained in migration-based PEA in island model. For instance, 2 × 2 mesh with cross connections is equivalent to the migration model with four islands, if subpopulations in mesh nodes and on islands are of equal size.

References

[1] Alba, E.: Parallel metaheuristic - a new class of algorithms, John Wiley & Sons, 2005. DOI: 10.1002/0471739383

[2] Alba, E.—Tomasini, M.: Parallelism and evolutionary algorithms, IEEE Trans. Evol. Comput. Vol. 6, No. 5, 2002, pp. 443–462. DOI: 10.1109/TEVC.2002.800880

[3] Bäck, T.: Evolutionary algorithms in theory and practice, Oxford University Press, 1996.

[4] Barbucha, D.—Jędrzejowicz, P.—Ratajczak, E.—Forkiewicz, M.: Population learning algorithm versus evolutionary computation, 15th International Parallel and Distributed Processing Symposium, IPDPS'2001, IEEE Comput. Society, 2001, pp. 1315–1322 (CD-ROM edition) DOI: 10.1109/IPDPS.2001.925108

[5] Bouziri, H.—Jouini, M.: A tabu search approach for the sum coloring problem, Electronic Notes in Discrete Mathematics, Vol. 36, 2012, pp. 915–922. DOI: 10.1016/j.endm.2010.05.116

[6] Cantú-Paz, E.: A survey of parallel genetic algorithms. Calculateurs Paralleles, Reseaux et Systems Repartis Vol. 10, No. 2, Paris, Hermes, 1998, pp. 141–171.

[7] Cantú-Paz, E.: Topologies, migration rates, and multi–population parallel genetic algorithms, Proceedings of the Genetic and Evolutionary Computation Conference, GECCO'99, Morgan Kaufmann, 1999, pp. 91–98.

[8] Cantú-Paz, E.: Efficient and accurate parallel genetic algorithms, Kluwer, 2000.

[9] Cantú-Paz, E.—Goldberg, D. E.: Parallel genetic algorithms: theory and practice. Computer Methods in Applied Mechanics and Engineering, Elsevier, 2000. DOI: DOI: 10.1016/S0045-7825(99)00385-0

[10] Croitoriu, C.—Luchian, H.—Gheorghies, O.—Apetrei A.: A new genetic graph coloring heuristic, Computational Symposium on Graph Coloring and Generalizations COLOR'02. In: Constraint Programming, Proceedings of the International Conference, CP'02, 2002.

[11] Filho, G. R.,—Lorena, L. A. N.: Constructive genetic algorithm and column generation: an application to graph coloring, Proc. Asia Pacific Operarions Research Symposium, APORS'2000, 2000.

[12] Fleurent, C.—Ferland, J. A.: Genetic and hybrid algorithms for graph coloring, Annals of Operations Research Vol. 63, 1996, pp. 437–461. DOI: 10.1007/BF02125407

[13] Galinier, P.—Hao, J-K.: Hybrid evolutionary algorithms for graph coloring, J. Combinatorial Optimization, 1999, pp. 374–397. DOI: 0.1023/A:1009823419804

[14] Garey, R.—Johnson, D. S.: Computers and intractability. A guide to the theory of NP-completeness, Freeman, 1979.

[15] Glass, C. A.—Prügel-Bennett, A.: Genetic algorithm for graph coloring: Exploration of Galinier and Hao's algorithm, J. Combinatorial Optimization, 2003, pp. 229–236. DOI: 10.1023/A:1027312403532

[16] Hutchinson, G.: Partitioning algorithms for finite sets, Comm. ACM No. 6, 1963, pp. 613–614.

[17] Iorio, A.—Li, X.: Parameter control with a co-operative, co-evolutionary genetic algorithm, Parallel Problem Solving from Nature, Proceedings of the International Conference, PPSN'2002, LNCS Vol. 2439, 2002, pp. 247–256. DOI: 10.1007/3-540-45712-7_24

[18] Jensen, T. R.—Toft, B.: Graph coloring problems, Wiley Interscience, 1995.

[19] Johnson, D. S.,—Trick, M. A.: Cliques, coloring and satisfiability: Second DIMACS Implementation Challenge, DIMACS Series in Discr. Math. and Theor. Comp. Sc. Vol. 26, 1996.

[20] Karp, R. M.: Reducibility among combinatorial problems, In: Miller R. E. and Thatcher J. W. (Eds.), Complexity of Computer Computations, Plenum Press, 1972, pp. 85–103.

[21] Khuri, S.—Walters, T.—Sugono, Y.: Grouping genetic algorithm for coloring edges of graph, Proc. 2000 ACM Symposium on Applied Computing, 2000, pp. 422–427. DOI: 10.1145/335603.335880

[22] Kokosiński, Z.: Effects of versatile crossover and mutation operators on evolutionary search in partitions and permutation problems, Intelligent Information Systems: New Trends in Intelligent Information Processing and Web Mining 2005, Proceedings of the International Conference, IIS:IIPWM'2005, Advances in Soft Computing, Springer, 2005, pp. 299–308. DOI: 10.1007/3-540-32392-9_31

[23] Kokosiński, Z.—Kolodziej, M.—Kwarciany, K.: Parallel genetic algorithm for graph coloring problem, Computational Science, Proceedings of the International Conference, ICCS'2004, LNCS Vol. 3036, 2004, pp. 215–222. DOI: 10.1007/978-3-540-24685-5_27

[24] Kokosiński, Z.—Kwarciany, K.: On sum coloring of graphs with parallel genetic algorithms, Adaptive and Natural Computing Algorithms, Proceedings of the International Conference, ICANNGA'2007, LNCS Vol. 4431, 2007, pp. 211-219. DOI: 10.1007/978-3-540-71618-1_24

[25] Kokosiński, Z.: Parallel metaheuristics in graph coloring, Bulletin of the National University "Lviv Politechnic", No. 744 , 2012, pp. 209–214.

[26] Kubale, M.: Introduction to computational complexity and algorithmic graph coloring, GTN, Gdańsk, 1998.

[27] Kubale, M. (Ed.): Graph colorings, American Mathematical Society, 2004. DOI: 10.1090/conm/352

[28] Levine, D.: A parallel genetic algorithm for the set partitioning problem, Argonne Nat. Lab., Argonne, IL, 1996.

[29] Lis, J.: Parallel genetic algorithm with the dynamic control parameter, Evolutionary Computation, ICEC'1996, Proceedings of the International Conference, IEEE Computer Society, 1996, pp.324–329. DOI: 10.1109/ICEC.1996.542383

[30] Lorena, L. A. N.—Filho, G. R.: Constructive genetic algorithm for graph coloring, Proc. Asia Pacific Operarions Research Symposium APORS'97, 1997.

[31] Łukasik, S. Kokosiński Z. Świętoń G.: Parallel simulated annealing algorithm for graph coloring problem, Parallel Processing and Applied Mathematics, Proceedings of the International Conference, PPAM'2007, LNCS Vol. 4967, 2008, pp. 229-238. DOI: 10.1007/978-3-540-68111-3_25

[32] Myszkowski, P.B.: Solving scheduling problems by evolutionary algorithms for graph coloring problem, [in :] Xhafa F., Abraham A.: Metaheuristics for scheduling in industrial and manufacturing applications, Studies in Computational Intelligence, Vol. 128, 2008, pp. 145–167. DOI: 10.1007/978-3-540-78985-7_7

[33] Nowostawski, M.—Poli, R.: Parallel genetic algorithm taxonomy, Knowledge–based Intelligent Information Engineering Systems, KES'99, Proceedings of International Conference, IEEE, 1999, pp. 88–92. DOI: 10.1109/KES.1999.820127

[34] Nowostawski, M.—Poli, R.: Dynamic demes parallel genetic algorithm, Knowledge–based Intelligent Information Engineering Systems, KES'99, Proceedings of International Conference, IEEE, 1999, pp. 93–98. DOI: 10.1109/KES.1999.820128

[35] Szyfelbein, D.: Genetic algorithms for graph coloring. Neural Networks and Their Applications, Proceedings of the Conference, Polish Neural Network Society, 1999, pp. 605–610.

[36] Szyfelbein, D.: Metaheuristics in graph coloring. In: Kubale M. (Ed.): Discrete optimization. Models and methods for graph coloring, WNT, Warszawa, 2002, pp. 26–52 (in Polish).

[37] de Werra, D.: Heuristics for graph coloring, In: Tinhofer, G. et all. (Eds.) Computational graph theory, Springer–Verlag, 1990, pp. 191–208.

[38] COLOR web site.
Available at: http://mat.gsia.cmu.edu/COLOR/instances.html.

[39] DIMACS ftp site.
Available at: ftp://dimacs.rutgers.edu/pub/challenge/graph/benchmarks/.

[40] COLORING3 web site.
Available at: http://mat.gsia.cmu.edu/COLORING03/.

DOI: 10.15439/2014F206
ACSIS, Vol. 3

YUV vs RGB – Choosing a Color Space for Human-Machine Interaction

Michal Podpora
Opole University of Technology
Faculty of Electrical Engineering,
Automatic Control and Informatics
Institute of Control and Computer Engineering
ul. Sosnkowskiego 31, 45-272 Opole, Poland
m.podpora@po.opole.pl

Grzegorz Paweł Korbaś, Aleksandra Kawala-Janik
Opole University of Technology
Faculty of Electrical Engineering,
Automatic Control and Informatics
Institute of Electromechanical Systems
and Industrial Electronics
ul. Prószkowska 76, 45-758 Opole, Poland
{g.korbas, a.kawala-janik}@po.opole.pl

***Abstract*—This paper describes and compares two color spaces – YUV and RGB, taking into account possible human-computer interaction applications. Human perception-oriented properties are compared, including not only file size or bandwidth, but also subjective visibility of artifacts. 1700 tests on a group of 170 people were performed to describe the subjective quality of compressed YUV and RGB images. The paper shows that the use of the YUV color space for a machine vision implementation can give better subjective image quality than the RGB color space. The authors conclude that YUV is better for machine vision implementations than RGB due to the perceptual similarities to the human vision.**

Keywords* – *YUV, RGB, color space, image processing, HCI, HMI, image understanding, computer graphic, machine vision, image compression

I. Introduction

MACHINE Vision and Computer Vision are dominated by the RGB color space, which seems to be the most intuitive programmer's choice, while it is being used by digital image acquisition hardware and in the majority of processing methods and algorithms. Red, green and blue optical filters, combined with Active-Pixel Sensors comprise the simplest and most popular color vision acquisition systems.

This article confronts the RGB color space with YUV. Although YUV was primarily introduced to add the color information to existing monochromatic channel, it turned out that YUV is also in a way similar to human vision – the "black and white" information has more impact on the image for human eye than the color information.

Therefore, it is worth considering which of these color spaces might be better for Human-Machine Interaction systems.

The authors have prepared a simple application to test perceptual capabilities of volunteers. The following chapters describe the research and present the conclusions.

II. RGB and YUV Color Spaces

RGB and YUV color spaces are both based upon the perceptual capabilities of human eye. The RGB color space is plainly based on the acquisition capabilities of cone cells in retina, which are able to react to different wavelengths. Electronic devices usually display three base colors (red, green and blue). Other colors and shades are achieved by combining these three colors using the additive color mixing. The cones' response to a specific wavelength is presented in fig.1. Fig.2 shows the spectrum of two exemplary computer monitors available on the market.

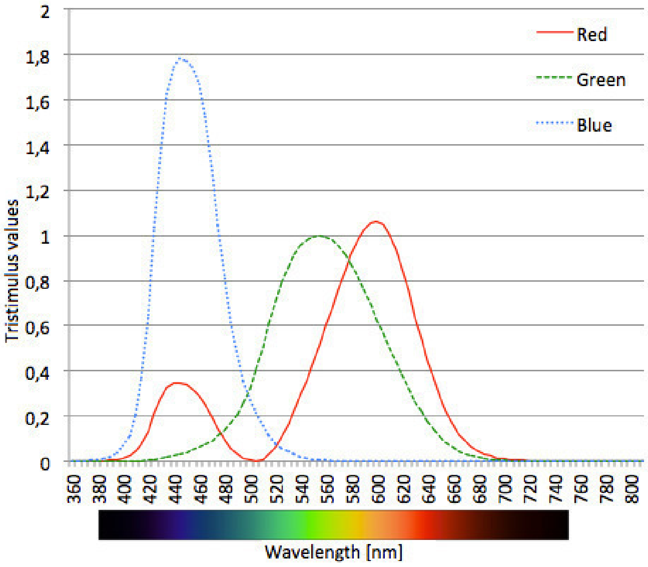


Fig. 1. CIE 1931 Color Matching Functions [4] [8]

Fig.1 (CIE 1931 CMF, corresponding to the acquisition capabilities of cone cells) and fig.2 (spectrograms of popular monitor technologies) are obviously different, but this difference, for a human eye, is perceptually negligible. Nevertheless, these figures show that there is a change in information when using RGB displays. Designers of Machine Vision systems should keep in mind that flattering the spectrum to three values is a huge simplification.

The YUV color space, on the other hand, can also be considered to be similar to human eye's retina, while the main

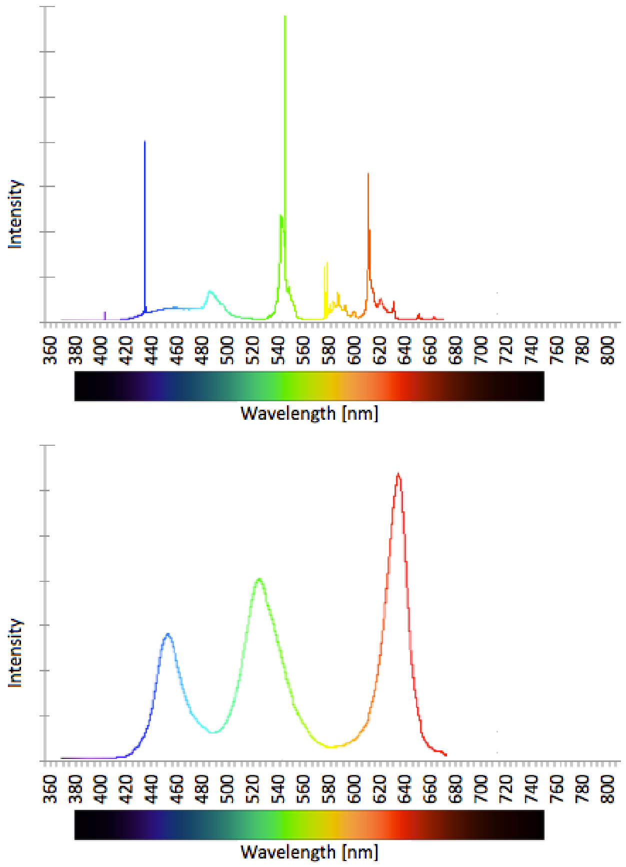


Fig. 2. Spectrograms of a CCFL-backlit LCD monitor (Samsung SyncMaster 913N) and a LED monitor (Samsung SyncMaster XL20), respectively [1]

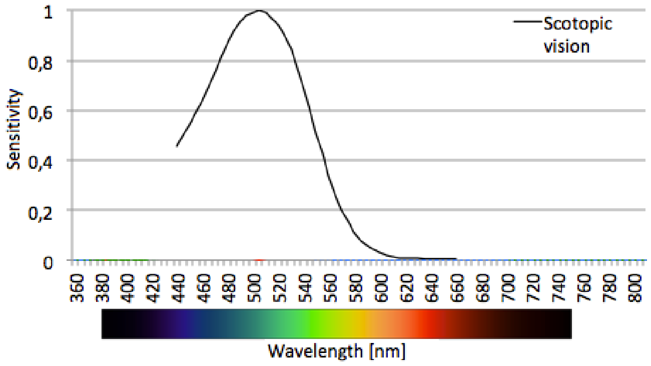


Fig. 3. Scotopic (rod cells) sensitivity [9] [10]

channel – luminance (denoted as Y channel) or "luma" (denoted as Y') describes the intensity of light, just like rod cells in the retina. Rod cells are the primary source of information in the dark, when the cone cells do not have sufficient intensity of light for activation to distinguish colors. However, when the intensity slightly increases, additional information from cone cells become available. In the YUV color space, two additional channels – chrominance components called "U" and "V" – carry the color information (e.g. as blue-luminance and red-luminance, respectively, for a digital signal – in case of YCbCr).

In the YUV color space, the black and white information is separated from the color information. Primarily, YUV was used in analog television standards, when color information was added to the existing luminance channel. To enable backward compatibility for black-and-white transceivers, the chrominance channels were added in a separate subcarrier.

Using a YUV color space, also usually involves loss of information, but for a different reason than in RGB color space. In analog YUV it is popular to use interlacing in chrominance channels (the contrast in luminance channel is more significant information for a human eye than color in chrominance channels). In digital YUV, the signal is usually converted from RGB acquisition hardware, which involves a lossy conversion from RGB to YUV.

Therefore, the YUV color space is also a compromise of perceptually-reasoned loss of information.

A. *RGB formats*

The RGB color space has many various representations, but they all have one in common: three separate color values are stored for three predefined colors: red, green and blue. The colors can be ordered starting from red (RGB) or starting from blue (BGR). If the fourth letter ("A") is present, the fourth channel contains the "alpha" (transparency) value for the pixel. If the name of the format contains any digits, the usually mean the amount of bits for every pixel – e.g. RGB24p (or RGB24bpp) means 24 bits for pixel's color information (i.e. 8 bits for red, green and for blue), and BGRA8888 means 8 bits for subsequent channels (i.e. blue, green, red and alpha, respectively). The formats are well explained [6] in OpenCV-related webpages, regarding converting image between two specific formats (functions like cv_bgr2gray(), cv_rgb2ycrcb() and other).

Some alternative RGB formats are also available in OpenCV (e.g. Bayer pattern) but these are not discussed in this paper.

B. *YUV formats and conversion from RGB*

Historically, the term YUV was used for analog encoding. Nowadays this term is frequently used for analog and digital encoding as well. There are many formulas to convert from RGB to YUV [5] [8]. In this article digital YCbCr defined by ITU-R BT.601 has been used. In this color space Y (denoted as Y') represents "luma" (the weighted sum of gamma-compressed RGB components) while Cb and Cr are blue-difference and red-difference chrominance components. Referring to the mentioned recommendation YCbCr is derived as follows:

$$\begin{bmatrix} Y' \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

where Rd, Gd, Bd represent 8-bit values for red, green and blue color channels.

C. *Popularity of formats*

A trial research has been carried out to check the popularity of the most popular image representation formats. Authors studied the hit count of webpages containing 41 most popular OpenCV format-conversion functions (inter alia: cv_rgb2yuv(), cv_bgr2hsv()). Fig.4 shows the total hit count (number of webpages) for specific image format conversion methods, divided into 4 groups: grayscale, RGB, YUV, HSV. Fig.4 shows also average hit count (number of webpages divided by the number of queried conversion methods names).

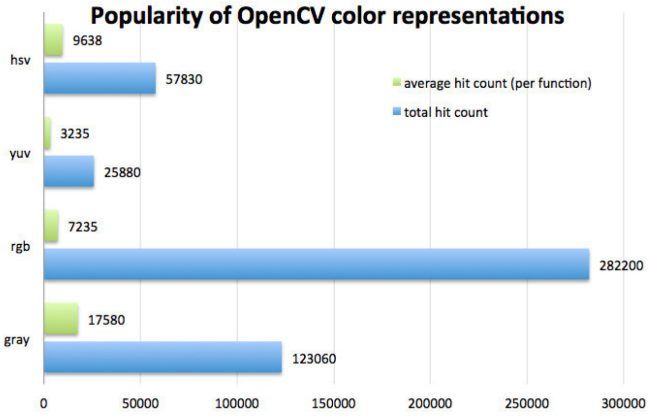


Fig. 4. Popularity of the most popular color representations based on the Google hit count for 41 most popular OpenCV [6] format conversion functions [own work]

Fig.4 clearly indicates that the OpenCV conversion functions for RGB color space have the greatest number of webpages/resources/queries, while the YUV color space is currently much less popular amongst OpenCV programmers.

III. Research incentives and Expectations

One of the authors in [7] has suggested that YUV color space might be more similar to human vision than RGB. However, the research [7] involved only a brief comparison of the color spaces and discussion on possible compression differences. The issue was analyzed in a Machine Vision aspect, no human factor/opinion was taken into account.

In this paper, the authors have decided to investigate if the subjective quality difference is real (i.e. if it is also visible/noticeable to other people), and carried out extensive tests to find out if the YUV representation of an image has perceptually better quality than RGB. It turned out that, indeed, people have noticed the difference in quality.

Technically (and mathematically) compressing RGB (RGB24p) and (YUV888) color spaces images should give comparable quality, while in both of them there are three bytes describing every single pixel. However, the "black and white" detail has more impact on the image for a human eye because of its rather low color sensitivity. Manipulating with red, green or blue value always gives a perceptibly different image, while converting an image to the YUV color space gives possibility to process the luminance and chrominance signals independently. The luminance channel is surely considered the most useful one for image processing in YUV, therefore reducing the chrominance signal quality may pass unnoticed to a human.

The idea of chroma subsampling has been formerly used for image coding in YUV formats, e.g. YUV422, but in every case the output image quality was aimed at a human recipient. If the image is to be analyzed by a Machine Vision system, the RGB color space is nowadays considered to be the most useful form of the visual information. This is not necessarily true. If a robot is supposed to work and to co-exist along with human, it should "see" the world the way we do – with similar inaccuracies. The threshold and the ability of "not recognizing" an object is the key issue of learning and robot's better understanding of it's environment. [7]

Converting an RGB image to the YUV color space is a lossy operation – this might be the reason (according to [7]) compressed YUV image files are often smaller than compressed RGB image files. However, the difference in the file sizes seems not to be proportional to the difference in quality of the images. It is difficult to discuss the change in quality as the quality loss is usually defined as the difference between original image and the compressed image. In that case, the YUV representation of the image should be considered as the worse one (due to additional lossy operation - RGB-YUV conversion). If (after the same lossy compression/conversion) the RGB representation would be of better quality than YUV, it would be natural to try to lower the threshold for compression/conversion of YUV to improve its quality (so that the quality of RGB and YUV would become similar). Surprisingly, the quality of the RGB representation did not seem to be of better quality than YUV. Contrary, it seemed to be of worse quality than YUV. Authors have named it as the subjective quality difference.

To confirm the existence of the subjective quality difference, and to assess it's extent, a simple test application has been developed. A brief description of the application, test images, testing procedures and the discussion of the results are included in the next section (Research Methodology).

A. *The preparation of test images*

The authors expected to find and estimate the subjective quality difference between RGB and YUV images. To ensure the quality of the research, all images have been processed using the same algorithms and settings. The basic image preparation procedure, presented in fig.5, consisted of following steps:

- choosing an interesting image with representative thus unique image attributes (contrast, quality, saturation, edges, gradients, etc.) and satisfying ppi (pixels per inch) resolution,
- cropping the image to 256x256 pixels,
- saving the image as an RGB 24bpp format bmp file.

Both images, visible on the form of the application (fig.6) are created in runtime, basing on the same RGB 24bpp test image.

One of ten predefined test images is loaded and processed in following steps:

- the image is cloned in the application's memory,
- one of the images is converted to YUV888 using algorithm from equation (1),
- both matrices are converted using DWT (discrete wavelet transform) [2] with a specific threshold,
- both images are recovered using IDWT (inverse DWT),
- the YUV image is converted back to RGB,
- both images are displayed on the application's form.

The threshold of the DWT algorithm of one of the images is random (in a predefined range), whereas the threshold of the second image's DWT algorithm is available on the front-end of the application to user as a trackbar - the user is able to modify it's value (and re-run the DWT-IDWT algorithm with the new setting).

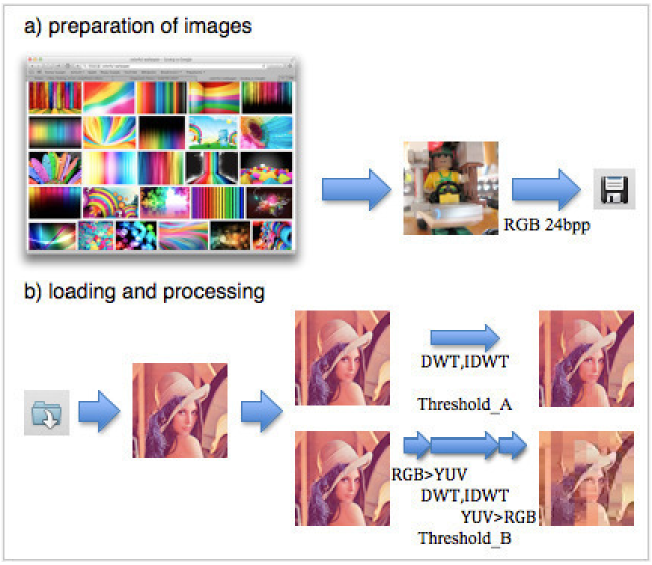


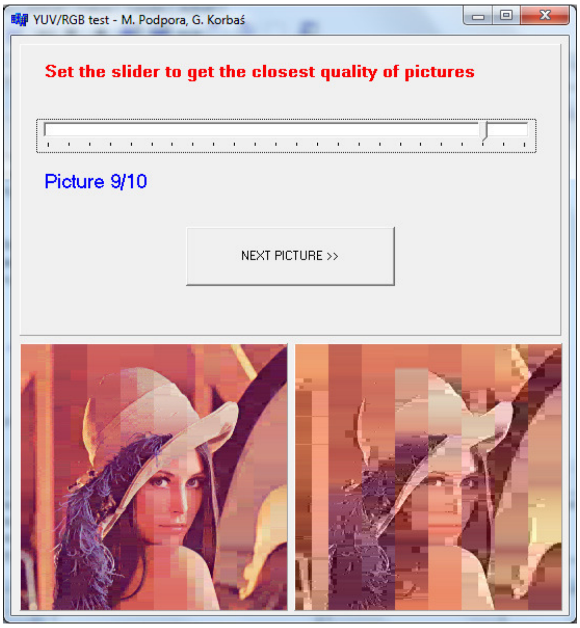
Fig. 5. Image preparation procedure for the application's test images

Since the quality/distortion threshold parameter, as the authors suggest, should be modeled on human perception rather than simply as a variance of difference between input and output image, some perceptual distortion measures should be developed. Audio compression perceptual models are relatively advanced (mp3, ogg), the perception aspect is also present in some of the compression algorithms of image data (usually available to users as a quality threshold value), but it the aspect of color space, currently there seem to be no research.

IV. Research Methodology

In order to investigate the subjective difference in images, an application has been implemented. The graphical user interface of the application is presented in fig.6. The application form includes two panels and a slider. Each of these two panels show an image, based on the same source but converted in a different way. The source images have good quality, but the images in GUI have significantly lower quality (so that the quality drop would be clearly visible to a human). The source images were transformed using different DWT threshold value for each panel. One of the GUI images was transformed directly from the RGB color space, and the second one was transformed to the YUV color space first. During the tests, the threshold value of the DWT transformation for the RGB-based image was set to a (random) fixed value, while for the YUV-based image the user was able to modify the DWT threshold using a slider. The tests included also an inverted scenario: a fixed threshold for YUV and a slider for RGB.

Ten tests were conducted in every test scheme (five various images were loaded and presented in two following procedures: at first the threshold for RGB was fixed and then the threshold for YUV was fixed). Pictures that were chosen for the research, comprised a set of interesting features, inter alia: variable complexity, clear edges as well as some gradients, good color saturation, etc. The source images were 256x256 pixels, 24 bits per pixel, RGB, uncompressed, BMP images.


YUV/RGB test - M. Podpora, G. Korbaś

Set the slider to get the closest quality of pictures

Picture 9/10

Fig. 6. Graphical user interface of the application

During the trials, users were asked to set the slider in such a way that the quality of the two images would seem similar. The slider offered 24 positions, translated into 24 threshold values of the wavelet transform, affecting the quality of one of the images. The default slider position was either 1 or 24, while the user was supposed to set the slider to 7–17 (the DWT threshold of the second image was randomized from the range: 7–17). If the user did not modify the default slidebar position (and left it on 1 or 24) it clearly indicated that the test results should be rejected. By moving the slider to an intermediate position, the user could subjectively ascertain if the quality of the two images was at a comparable level.

The study involved 170 people, aged from 10 to 40 years (most aged 17-24 years). This has provided 1700 test results. 7% of the tests were rejected because of the extreme positions of sliders (indubitable quality difference). However, further analysis was carried out for both situations: not only for tests marked as correct, but for all tests (without rejecting any) and the results were very similar - the users that did not bother to use the slider, did not use it in both test configurations: fixed-RGB and fixed-YUV, influencing both result data sets in the same way.

The authors decided that the basic parameter analyzed in this study is the number of zeros present in the matrix describing the image after wavelet transform. If the two images are subjectively the same quality, the greater the number of zeroes in the DWT matrices of one of the images means that less information is needed to describe the image. A greater number of zeroes also enables a possibility of higher level of image compression, which translates into the potential application of the results. In order to compare the usefulness of a color space, following factor has been defined:

$$P = \frac{Z_{YUV} - Z_{RGB}}{Z_{RGB}} \cdot 100 \qquad (2)$$

where Z$_{RGB}$ is the number of zeroes in the DWT matrices when using the RGB color space, and Z$_{YUV}$ is number of zeroes when using the YUV space. The P coefficient indicates percentage – the number of zeroes that have been found using YUV compared to the number of zeroes that have been found using RGB. A positive coefficient value means that (after the user had set the slider position to set similar quality of both images) more zeroes occurred in the YUV, while a negative value of the coefficient indicates the superiority of RGB color space.

It is worth noting that some of the slider positions allowed to obtain positive and some – negative values of the P coefficient. An exemplary dependence of P upon the slider's position (for one of the tests with fixed DWT threshold for the RGB and modifiable one for YUV) is shown in fig.7. The distribution of the slider's positions set by users for this particular test is shown in fig.8. (Positions 20 and 22 were excluded from evaluation.)

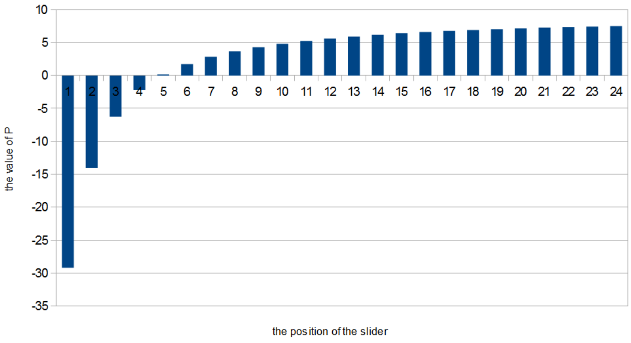


Fig. 7. Dependance of the P coefficient upon specific slider positions (i.e. specific threshold values) for one specific exemplary test

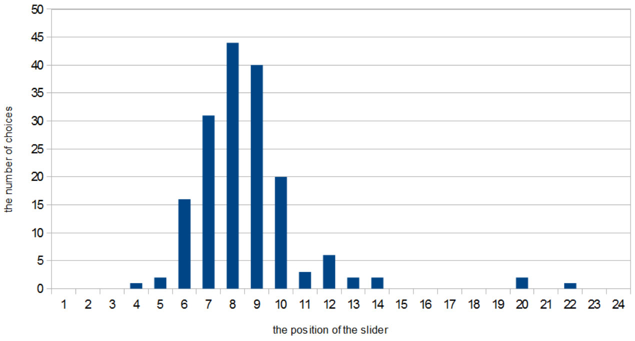


Fig. 8. Exemplary statistics of slider positions for one specific test

V. Results

The results were analyzed separately for each test. In each case, the average value of the coefficient P was positive - ranging from 0.59 to 4.40. After merging the results of all tests, the average of the P coefficient has been obtained. Statistical results for all images and tests are shown in fig.9.

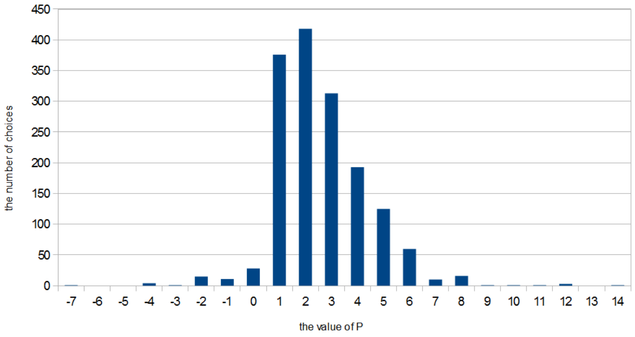


Fig. 9. General statistics of the P coefficient (all tests)

The results indicate that the use of the YUV space in some cases may be more effective than using the RGB color space. These conclusions are of a general nature, but they are also true for each image individually. The use of the YUV space results in a larger number of zeroes in the DWT matrices.

VI. Advantages for HMI Systems, Future Work

The study showed that people qualify the images as qualitatively similar even though the images are described using a different form of information. Comparison of the RGB and YUV color spaces in conjunction with the wavelet transform shows that the use of the YUV space enables efficient reduction of the amount of information necessary to represent the image of subjectively similar quality. Therefore, systems designed to map the human factor in the field of image processing should use YUV color space. A smaller number of data needed to make a decision may result, among others, in faster performance and/or reduction of the size of transmitted data in the system. This perceptual difference in quality can be used in another way - by modifying the subjective image quality (if image size reduction is not required [7]) by adjusting the luminance channel compression threshold value to improve perceptual quality while preserving comparable file size.

It is noteworthy that the conversion between RGB and YUV was performed using the specific conversion coefficient values defined by ITU-R BT.601.

The analysis of other coefficients, their values, thresholds and their impact on the effectiveness of the use of YUV space should be subjected to further research.

The positive results related to the analysis of the human factor, of course do not preclude the benefits of the YUV space in traditional decision-making systems. There are many algorithms related to the recognition of shapes of objects, searching for specific parameters, motion detection, etc. It is possible that the use of the YUV color space can bring many benefits also in the classical cases. The authors consider further research in this field is to allow a broader view and more accurate analysis of application areas of YUV instead of RGB color space.

REFERENCES

[1] Artamonov O., Xbit Laboratories, 'Contemporary LCD Monitor Parameters - Objective and Subjective Analysis', 2007, available on-line (accessed 2014-03-01) http://www.xbitlabs.com/articles/monitors/display/lcd-parameters.html

[2] Berowski P., 'Transformata Falkowa', Instytut Elektrotechniki, Warszawa, available on-line (accessed 2014-04-01) http://bambus.iel.waw.pl/pliki/ogolne/studia_doktoranckie/wyklady/wyklad_falki.pdf

[3] Color and Vision Research Labs, 'CIE Standards', 2006, available on-line (accessed 2014-03-01) http://cvrl.ioo.ucl.ac.uk/cmfs.htm

[4] Glynn E. F., 'CIE Chromaticity Diagrams - CIE Color Matching Functions', 2009, available on-line (accessed 2014-03-01) http://www.efg2.com/Lab/Graphics/Colors/Chromaticity.htm

[5] Hamilton E., 'JPEG File Interchange Format, Version 1.02', available on-line (accessed 2014-03-08) http://www.jpeg.org/public/jfif.pdf

[6] Open Source Computer Vision Library, 'OpenCV online documentation', available on-line (accessed 2014-03-01) http://opencv.org

[7] Podpora M. A., 'YUV vs. RGB – A comparison of lossy compressions for human-oriented man-machine interfaces', *III SWD conference proceedings*, Glucholazy 2009, Oficyna Wydawnicza Politechniki Opolskiej, ISSN 1429-1533, vol. 62, no. 329/2009, Opole, 2009

[8] Szeliski R., 'Computer Vision: Algorithms and Applications', Springer, 2011

[9] Wald G., Brown P. K., 'Human rhodopsin', Science, 1958

[10] Wandell B. A., 'Foundations of Vision', Sinauer Associates Inc., ISBN 978-0878938537, Sunderland, 1995, available on-line (accessed 2014-03-01) https://www.stanford.edu/group/vista/cgi-bin/FOV/

International Workshop on Artificial Intelligence in Medical Applications

THE workshop on Artificial Intelligence in Medical Applications – AIMA'2014 - provides an interdisciplinary forum for researchers and developers to present and discuss latest advances in research work as well as prototyped or fielded systems of applications of Artificial Intelligence in the wide and heterogenious field of medicine, health care and surgery. The workshop covers the whole range of theoretical and practical aspects, technologies and systems based on Artificial Intelligence in the medical domain and aims to bring together specialists for exchanging ideas and promote fruitful discussions.

Topics

The topics of interest include, but are not limited to:

- Artificial Intelligence Techniques in Health Sciences
- Knowledge Management of Medical Data
- Data Mining and Knowledge Discovery in Medicine
- Health Care Information Systems
- Clinical Information Systems
- Agent Oriented Techniques in Medicine
- Medical Image Processing and Techniques
- Medical Expert Systems
- Diagnoses and Therapy Support Systems
- Biomedical Applications
- Applications of AI in Health Care and Surgery Systems
- Machine Learning-based Medical Systems
- Medical Data- and Knowledge Bases
- Neural Networks in Medicine
- Ontology and Medical Information
- Social Aspects of AI in Medicine
- Medical Signal and Image Processing and Techniques
- Ambient Intelligence and Pervasive Computing in Medicine and Health Care

Event Chairs

Pancerz, Krzysztof, University of Information Technology and Management in Rzeszów, Poland

Piątek, Łukasz, University of Information Technology and Management in Rzeszów, Poland

Program Committee

Andrushevich, Aliaksei, Lucerne University of Applied Sciences, Switzerland

Bazan, Jan, University of Rzeszów, Poland

Cardoso, Jaime, University of Porto, Portugal

Drahansky, Martin, Brno University of Technology, Czech Republic

Grzymala-Busse, Jerzy, University of Kansas, United States

Hassanien, Aboul Ella, Cairo University, Egypt

Hiroyasu, Tomoyuki, Doshisha University, Japan

Iantovics, Barna, Petru Maior University, Romania

Kountchev, Roumen, Technical Univerity of Sofia, Bulgaria

Krawczyk, Bartosz, Wroclaw University of Technology, Poland

Kumar, Sajeesh, University of Tennessee, Health Science Center, United States

Marchenko, Dmitro, Volodymyr Dahl East Ukrainian National University, Ukraine

Min, Fan, Zhangzhou Normal University, China

Mohyuddin, Mohyuddin, King Abdullah International Medical Research Center, Saudi Arabia

Olszewska, Joanna Isabelle, University of Gloucestershire, United Kingdom

Sawada, Hideyuki, Kagawa University, Japan

Shulgin, Sergiy Kostyantynovych, Volodymyr Dahl East Ukrainian National University, Ukraine

Slezak, Dominik, University of Warsaw & Infobright Inc., Poland

Strzelecki, Michal, Lodz University of Technology, Poland

Wei, Wei, School of Computer science and engineering, Xi'an University of Technology, China

Wysocki, Marian, Rzeszow University of Technology, Poland

Yanushkevich, Svetlana, University of Calgary, Canada

Zaitseva, Elena, University of Zilina, Slovakia

DOI: 10.15439/2014F487
ACSIS, Vol. 3

Fuzzy Decision Tree Based Classification of Psychometric Data

Vitaly Levashenko*, Elena Zaitseva*, Krzysztof Pancerz[†] [‡] and Jerzy Gomuła[§] [¶]
*University of Zilina, Slovakia
Email: Vitaly.Levashenko@fri.uniza.sk, elena.zaitseva@fri.uniza.sk
[†]University of Management and Administration in Zamość, Poland
Email: kpancerz@wszia.edu.pl
[‡]University of Information Technology and Management in Rzeszów, Poland
[§]The Andropause Institute, Medan Foundation, Warsaw, Poland
Email: jerzy.gomula@wp.pl
[¶]Cardinal Stefan Wyszyński University in Warsaw, Poland

***Abstract*—For five years, the Copernicus system - a tool for computer-aided diagnosis of mental disorders based on data coming from psychometric tests - has been developed. This tool uses a variety of classification ways for differential interprofile diagnosis. In the current version of the tool, psychometric data come from the Minnesota Multiphasic Personality Inventory (MMPI) test. In this paper, we describe another machine learning approach, based on fuzzy decision trees, for classification of psychometric data. The algorithm for generation of fuzzy decision trees used by us is based on cumulative information estimations of initial data. Due to the promising results of classification of MMPI data, the presented approach will be implemented in the Copernicus system.**

I. Introduction

SO FAR, there have not been developed universal machine learning and data mining methods which could be applied for each kind of data, delivering expected results. Each kind of data requires an individual approach to them, and what follows, designing suitable, specialized methods for them. Classification models can be obtained in various ways (cf. [1], [2], [3]). Each way leads to obtaining a set of rules characterized by different coefficients describing their classification ability/quality. One of the main tasks of building decision support systems for computer-aided diagnosis of patients is to search for efficient methods of classification of new cases unseen earlier. Our research concerns psychometric data coming from the Minnesota Multiphasic Personality Inventory (MMPI) test [4]. MMPI is used to count the personality-psychometric dimensions which help in diagnosis of mental diseases. A brief information about this test is included in Section II.

In years 1998-1999, a team of researchers consisting of W. Duch, T. Kucharski, J. Gomuła, R. Adamczak created two independent rule systems devised for the nosological diagnosis of persons that may be screened with the MMPI-WISKAD test [5]. In the literature, we can also find descriptions of some other computer tools for classification of MMPI profiles, e.g., based on the Fortran program [6], "Panda" [7]. However, these tools are now mature. For five years, our research has been focused on creating a new computer tool called the Copernicus system. We started creation of the system in 2009. The main goal of this tool is to support clinical psychologists in differential and clinical diagnosis based on the overall analysis of profiles of patients examined by means of personality inventories. The tool has been designed for the Java platform. We can distinguish three main parts of the Copernicus system:

- Knowledge base.
- Multiway classification engine.
- Visualization engine.

The development of the Copernicus system (consecutive versions) over five years has been described in [8], [9], [10].

In the paper, we describe another machine learning approach that we plan to implement in the Copernicus system. This approach uses fuzzy decision trees proposed in [11]. That approach, based on cumulative information estimations of initial data, is brought back in Section III. Results of experiments made on over 1700 cases showed that the approach looks very promising.

II. MMPI Data

The Minnesota Multiphasic Personality Inventory (MMPI) test [4] delivers psychometric data on patients with selected mental disorders. Originally, the MMPI test was developed and published in 1943 by a psychologist S.R. McKinley and a neuropsychiatrist J.Ch. Hathaway from the University of Minnesota. Later, the inventory was adapted in above fifty countries. The MMPI-WISKAD personality inventory is the Polish adaptation of the American inventory. It has been used, among other modern tools, for carrying out nosological differential diagnosis in psychiatric wards. MMPI is also commonly used in scientific research. The test is based upon the empirical approach and originally was translated by M. Chojnowski (as WIO) [12] and elaborated by Z. Płużek (as WISKAD) in 1950 [13]. American norms were accepted there. Based upon received responses (*"Yes"*, *"Cannot Say"*, *"No"*) to selected questions we may make up a diagnosis for the subject being examined.

After examination by means of the MMPI test, each case (patient) x is described by a data vector $A(x)$ consisting of thirteen descriptive attributes: $A(x) = [A_1(x), A_2(x), ..., A_{13}(x)]$. A data vector is the so-called MMPI profile. The profile always has a fixed and invariable order of its constituents (attributes). If we have training data, then to each case x we also add one decision attribute Cl - a class (*nosological type* or *reference class*) to which a patient is classified. For the training data (which are used to learn or extract relationships between data), we have a tabular form (see example in Table I) which is formally called a decision system (decision table) $S = (Obj, Attr, Cl)$ in the Pawlak's form [14]. Obj is a set of cases (patients), $Attr$ is a set of descriptive attributes corresponding to scales, and Cl is a decision attribute determining a decision class.

The MMPI profile (data vector) can be divided into two parts. The validity part of the profile consists of three scales: L (laying) - attribute A_1, F (atypical and deviational answers) - attribute A_2, K (self-defensive mechanisms) - attribute A_3. The clinical part of the profile consists of ten scales: $1.Hp$ (Hypochondriasis) - attribute A_4, $2.D$ (Depression)- attribute A_5 , $3.Hy$ (Hysteria) - attribute A_6, $4.Ps$ (Psychopathic Deviate) - attribute A_7, $5.Mf$ (Masculinity/Femininity) - attribute A_8, $6.Pa$ (Paranoia) - attribute A_9, $7.Pt$ (Psychasthenia) - attribute A_{10}, $8.Sc$ (Schizophrenia) - attribute A_{11}, $9.Ma$ (Hypomania) - attribute A_{12}, $0.It$ (Social introversion) - attribute A_{13}. The clinical scales have numbers attributed to them so that a profile can be encoded to avoid negative connotations connected with the names of scales. Values of attributes are expressed by the so-called T-scores. The T-scores scale, which is traditionally attributed to MMPI, represents the following parameters: offset ranging from 0 to 100 T-scores, average equal to 50 T-scores, standard deviation equal to 10 T-scores.

In our experiments, we have used a data set, collected by T. Kucharski and J. Gomuła from the Psychological Outpatient Clinic, consisting of over 1700 patients (women) classified by a clinic psychologist. The data for the analysis (i.e., profiles of patients) were selected using the competent judge method (the majority of two-thirds of votes of three experts). Each case is assigned to one of nineteen nosological classes and the reference class (*norm*). Each class corresponds to one of psychiatric nosological types: neurosis (*neur*), psychopathy (*psych*), organic (*org*), schizophrenia (*schiz*), delusion syndrome (*del.s*), reactive psychosis (*re.psy*), paranoia (*paran*), sub-manic state (*man.st*), criminality (*crim*), alcoholism (*alcoh*), drug addiction (*drug*), simulation (*simu*), dissimulation (*dissimu*), and six deviational answering styles (*dev1*, *dev2*, *dev3*, *dev4*, *dev5*, *dev6*).

III. Fuzzy Decision Trees

A new greedy version of the Fuzzy ID3 algorithm based on cumulative information estimations of initial data has been proposed in [11]. It can be used to generate understandable fuzzy classification rules. Cumulative information estimations have been introduced in [15]. In this section we briefly recall basics of the Fuzzy ID3 algorithm.

Let us assume, that data describing cases are collected in the form of a decision system (decision table) $S = (Obj, Attr, Cl)$ (cf. Section II). In the recalled approach, for each real valued attribute $A_i \in Attr$ describing cases, fuzzy partitions A_{i_1}, A_{i_2}, ..., $A_{i_{k_i}}$, with ranges $[0, 1]$, are determined (see an example in Figure 1). The fuzzification of attribute values $A_i(u)$ for a given case $u \in Obj$ is performed by analysing the corresponding values of membership functions, i.e., $\mu_{A_{i1}}(u)$, $\mu_{A_{i2}}(u)$, ..., $\mu_{A_{ik}}(u)$. Analogously, the decision attribute Cl can be fuzzified, i.e., we obtain Cl_1, Cl_2, ..., Cl_m. In case of symbolic decision attribute values, a number of partitions is equal to the number of possible values of this attribute and membership functions take values either 0 or 1. Each attribute value can be seen as a likelihood estimate. In the approach, a particular case is assumed when the sum of membership function values of all partitions equals to 1. To transform numeric attribute values to triangular fuzzy data, the algorithm presented in [16] is used.

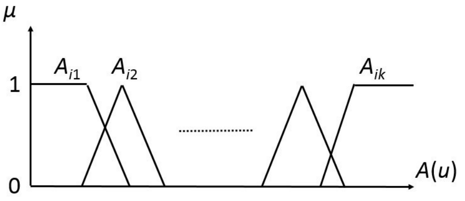


Fig. 1. An example of fuzzy partitions.

The cardinality measure $M(A_{i_1})$ for each A_{i_1} is calculated as:

$$M(A_{i_1}) = \sum_{u \in Obj} \mu_{A_{i1}}(u).$$

The cumulative joint information JI for Cl_j is:

$$JI(Cl_j) = \log_2 N - \log_2 M(Cl_j),$$

where N is a number of cases, i.e., the cardinality of Obj.

The cumulative joint information JI for Cl_j and $\mathbf{A}_q$ is:

$$JI(Cl_j, \mathbf{A}_q) = \log_2 N - \log_2 M(Cl_j \times A_{i_1} \times \cdots \times A_{q_{k_q}})$$

The cumulative conditional entropy H between Cl and A_{i_q} with a given $\mathbf{A}_q$ is:

$$H(B|\mathbf{A}_q, A_{i_q}) = \sum_{j=1}^{m} M(B_j \times \mathbf{A}_q) \times [JI(B_j, \mathbf{A}_q) - JI(\mathbf{A}_q)]$$

The cumulative conditional entropy H between Cl and A_i with a given $\mathbf{A}_q$ is:

$$H(B|\mathbf{A}_q, A_i) = \sum_{j=1}^{k_i} H(B|\mathbf{A}_q, A_{i_{k_i}}).$$

The cumulative mutual information I for A_i with a given $\mathbf{A}_q$ is:

$$I(B|\mathbf{A}_q, A_i) = H(B|\mathbf{A}_q) - H(B|\mathbf{A}_q, A_i).$$

These information estimations are used for forming new criteria of fuzzy decision tree induction (see [11]).

TABLE I
THE TRAINING DATA (FRAGMENT)

Attribute	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	Cl
Scale	L	F	K	$1.Hp$	$2.D$	$3.Hy$	$4.Ps$	$5.Mf$	$6.Pa$	$7.Pt$	$8.Sc$	$9.Ma$	$0.It$	
#1	55	65	50	52	65	57	63	56	61	61	60	51	59	$norm$
#2	50	73	53	56	73	63	53	61	53	60	69	45	61	org
#3	56	78	55	60	59	54	67	52	77	56	60	68	63	$paran$
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

TABLE II
EXEMPLARY FUZZIFICATION OF DESCRIPTIVE ATTRIBUTE VALUES (FOR CASE $\#1$)

Attribute A_i	A_{i_1}	A_{i_2}	A_{i_3}	A_{i_4}
A_1	0.077734	0.922266	0.000000	0.000000
A_2	0.298226	0.701774	0.000000	0.000000
A_3	0.000000	0.771691	0.228309	0.000000
A_4	1.000000	0.000000	0.000000	0.000000
A_5	0.453095	0.546905	0.000000	0.000000
A_6	0.858034	0.141966	0.000000	0.000000
A_7	0.238494	0.761506	0.000000	0.000000
A_8	0.119084	0.880916	0.000000	0.000000
A_9	0.725949	0.274051	0.000000	0.000000
A_{10}	0.579492	0.420508	0.000000	0.000000
A_{11}	0.873296	0.126704	0.000000	0.000000
A_{12}	0.781460	0.218540	0.000000	0.000000
A_{13}	0.000000	0.782395	0.217605	0.000000

There are two tuning parameters α and β used in the algorithm. Expanding a tree branch is stopped when either the frequency f of the branch is below α or when more than β per cent of cases left in the branch have the same decision class label.

IV. EXPERIMENTS

Experiments have been performed on a data set described in Section II. Each descriptive attribute has been fuzzified using four partitions (see an example in Table II). An exemplary fuzzy decision tree is shown in Figure 2. A decision attribute Cl has been fuzzified using 20 partitions because it has symbolic character (see an example in Table III). Cases are classified into 20 decision classes.

We have calculated fuzzy decision trees for several parameters α and β. The results (classification ability) of experiments are collected in Table IV. The best classification result has been obtained for $\alpha = 0.001$ and $\beta = 0.999$. We can compare classification ability of the presented approach with other classic machine learning approaches (see Table V).

We also refer the readers to our earlier papers where results of classification of MMPI profiles using a variety of approaches are presented. In [17], we have described the tests of several algorithms included in the following software tools:

- The Rough Set Exploration System (RSES) - a software tool featuring a library of methods and a graphical user interface supporting a variety of rough set based computations [18].
- NGTS - a system developed to generate decision rules using the algorithm called GTS (General-To-Specific) [19].

TABLE IV
RESULTS OF EXPERIMENTS FOR DIFFERENT PARAMETERS α AND β

α	β	Average classification error	Standard deviation
0.001	0.999	0.0823	0.0121
0.001	0.989	0.0823	0.0121
0.001	0.979	0.0829	0.0122
0.001	0.969	0.0838	0.0123
0.001	0.959	0.0846	0.0123
0.001	0.949	0.0854	0.0124
0.001	0.939	0.0865	0.0125
0.001	0.929	0.0878	0.0128
0.001	0.919	0.0896	0.0130
0.001	0.909	0.0915	0.0135
0.001	0.899	0.0937	0.0137
0.001	0.889	0.0959	0.0138
0.001	0.879	0.0978	0.0141
0.001	0.869	0.1001	0.0146
0.001	0.859	0.1030	0.0149
0.011	0.999	0.1646	0.1288
0.011	0.989	0.1648	0.1288
0.011	0.979	0.1650	0.1288
0.011	0.969	0.1653	0.1288
0.011	0.959	0.1656	0.1288
0.011	0.949	0.1660	0.1288
0.011	0.939	0.1667	0.1289
0.011	0.929	0.1675	0.1289
0.011	0.919	0.1684	0.1289
0.011	0.909	0.1693	0.1289
0.011	0.899	0.1703	0.1289
0.011	0.889	0.1715	0.1289
0.011	0.879	0.1727	0.1290
0.011	0.869	0.1745	0.1290
0.011	0.859	0.1766	0.1291
0.021	0.999	0.2100	0.1816
0.021	0.989	0.2102	0.1816
0.021	0.979	0.2102	0.1816
0.021	0.969	0.2103	0.1816
0.021	0.959	0.2104	0.1816
0.021	0.949	0.2110	0.1817

TABLE V
COMPARISON OF CLASSIFICATION ABILITIES OF DIFFERENT APPROACHES

Method	Average classification error	Standard deviation
FDT	**0.0823**	**0.0121**
C4.5	0.0952	0.0149
CART	0.1407	0.1287
Bayes	0.1531	0.1288
k-NN	0.0937	0.0129

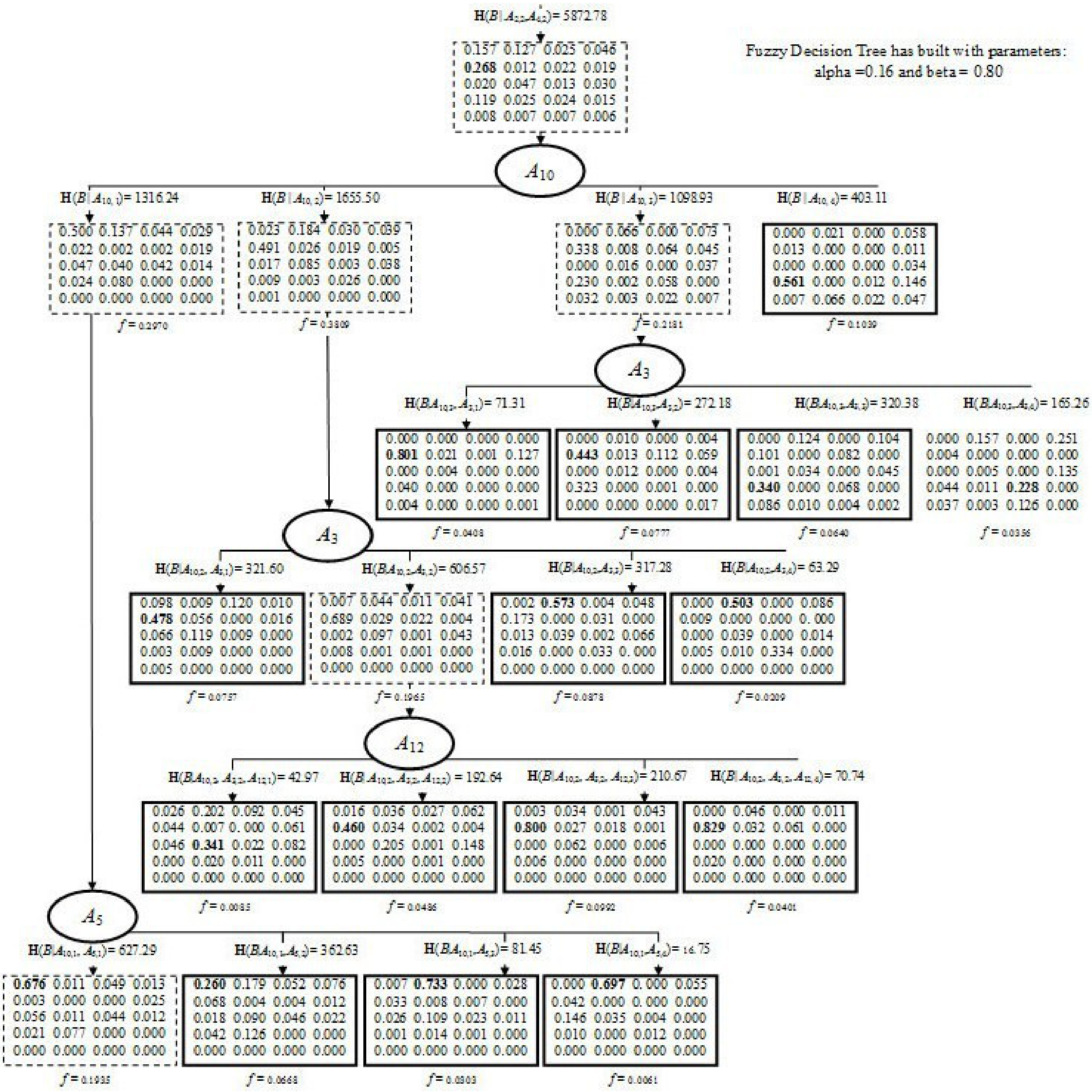


Fig. 2. An exemplary fuzzy decision tree.

- TreeSEEKER - a system containing several algorithms to generate decision trees [20].
- RuleSEEKER - a tool for generation and optimization of rule sets [21].

The main goal of experiments described in [22] was generation of efficient classification (decision) rules via decision trees on the basis of profiles of patients and selected indexes (e.g. Eichmann's indexes, Goldberg's indexes, Leary's indexes), calculated for profiles. Indexes added to profiles (scales) have been calculated using the Copernicus system. Next, for decision tree generation, the well-known C4.5 algorithm [23] (implemented in WEKA) has been used. Experiments with basic profiles (validity and clinical scales) and the extended ones (by adding different specialized indexes defined in the professional domain literature) have also been described in [24]. In [25], we have shown that it is possible to improve classification accuracy of MMPI profiles by reduction or extension of the number of attributes with relation to the original data table. In [26], a problem of hybridization and optimization of the knowledge base for the Copernicus system has been presented. Another tests of classification algorithms, based on decision trees available in the WEKA system [3], have been described in [27]. The review of the variety of approaches (not only machine learning) implemented so far in the Copernicus system is shown in [10].

V. Conclusions and Further Work

In the paper, we have presented another machine learning approach, based on fuzzy decision trees, for classification of psychometric data that is planned to be implemented in the Copernicus system - a tool for computer-aided diagnosis of mental disorders based on data coming from psychometric

TABLE III
EXEMPLARY FUZZIFICATION OF DECISION ATTRIBUTE VALUES (FOR CASE #1)

Cl_1	Cl_2	Cl_3	Cl_4	Cl_5	Cl_6	Cl_7	Cl_8	Cl_9	Cl_{10}	Cl_{11}	Cl_{12}	Cl_{13}	Cl_{14}	Cl_{15}	Cl_{16}	Cl_{17}	Cl_{18}	Cl_{19}	Cl_{20}
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

tests. The presented approach is based on cumulative information estimations of initial data. Results of experiments against a background of our previous research looks very promising. The main goal of further work is to develop the interpretation of fuzzified attribute values and obtained decision rules from the point of view of diagnosticians. Moreover, there is a need to optimize the transformation method of attribute values from numeric to fuzzy set partition variables. The current version seems to be not optimal.

ACKNOWLEDGMENT

The research has been partially supported by the grant No. SK-PL-0023-12 under Research and Development Cooperation Slovakia-Poland.

REFERENCES

[1] K. Cios, W. Pedrycz, R. Swiniarski, and L. Kurgan, *Data mining. A knowledge discovery approach*. New York: Springer, 2007. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-36795-8

[2] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[3] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[4] D. Lachar, *The MMPI: Clinical assessment and automated interpretations*. Fate Angeles: Western Psychological Services, 1974.

[5] W. Duch, T. Kucharski, J. Gomuła, and R. Adamczak, *Machine learning methods in analysis of psychometric data. Application to Multiphasic Personality Inventory MMPI-WISKAD (in polish)*, Toruń, 1999.

[6] W. E. Hatcher, "Automated classification of MMPI profiles into psychotic, neurotic or personality disorder types," *Computer Programs in Biomedicine*, vol. 8, no. 1, pp. 77–80, 1978.

[7] P. Pancheri and D. De Fidio, "Dal minnesota multiphasic personality inventory al panda: il mmpi-2 automatico," Tech. Rep.

[8] J. Gomuła, K. Pancerz, and J. Szkoła, "Analysis of MMPI profiles of patients with mental disorders - the first unveil af a new computer tool," in *Applications of Systems Science*, A. Grzech, P. Świątek, and K. Brzostowski, Eds. Warsaw, Poland: Academic Publishing House EXIT, 2010, pp. 297–306.

[9] ——, "Computer-aided diagnosis of patients with mental disorders using the copernicus system," in *Proceedings of the International Conference on Human System Interaction (HSI 2011)*, Yokohama, Japan, 2011. doi: 10.1109/HSI.2011.5937378. [Online]. Available: http://dx.doi.org/10.1109/HSI.2011.5937378

[10] D. Jachyra, K. Pancerz, and J. Gomuła, "Multiway classification of MMPI profiles," in *Proceedings of the Ninth International Conference on Digital Technologies (DT 2013)*, E. Zaitseva and V. Levashenko, Eds., Zilina, Slovakia, 2013. doi: 10.1109/DT.2013.6566294 pp. 119–127. [Online]. Available: http://dx.doi.org/10.1109/DT.2013.6566294

[11] V. Levashenko, E. Zaitseva, and S. Puuronen, "Fuzzy classifier based on fuzzy decision tree," in *Proceedings of the International Conference on Computer as a Tool (EUROCON 2007)*. IEEE, 2007. doi: 10.1109/EURCON.2007.4400614 pp. 823–827. [Online]. Available: http://dx.doi.org/10.1109/EURCON.2007.4400614

[12] M. Choynowski, *Multiphasic Personality Inventory (in polish)*, Psychometry Laboratory, Polish Academy of Sciences, Warsaw, 1964.

[13] Z. Płużek, *Value of the WISKAD-MMPI test for nosological differential diagnosis (in polish)*, The Catholic University of Lublin, 1971.

[14] Z. Pawlak, *Rough Sets. Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publishers, 1991. [Online]. Available: http://dx.doi.org/10.1007/978-94-011-3534-4

[15] V. Levashenko and E. Zaitseva, "Usage of new information estimations for induction of fuzzy decision trees," in *Proceedings of the Third International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2002)*, ser. Lecture Notes in Computer Science, H. Yin, N. Allinson, R. Freeman, J. Keane, and S. Hubbard, Eds. Springer Berlin Heidelberg, 2002, vol. 2412, pp. 493–499. [Online]. Available: http://dx.doi.org/10.1007/3-540-45675-9_74

[16] H.-M. Lee, C.-M. Chen, J.-M. Chen, and Y.-L. Jou, "An efficient fuzzy classifier with feature selection based on fuzzy entropy," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 3, pp. 426–432, 2001. doi: 10.1109/3477.931536. [Online]. Available: http://dx.doi.org/10.1109/3477.931536

[17] J. Gomuła, W. Paja, K. Pancerz, and Szkoła, "A preliminary attempt to rules generation for mental disorders," in *Proceedings of the International Conference on Human System Interaction (HSI 2010)*, Rzeszów, Poland, 2010. doi: 10.1109/HSI.2010.5514483. [Online]. Available: http://dx.doi.org/10.1109/HSI.2010.5514483

[18] J. G. Bazan and M. S. Szczuka, "The Rough Set Exploration System," in *Transactions on Rough Sets III*, ser. LNAI, J. Peters and A. Skowron, Eds. Berlin Heidelberg: Springer-Verlag, 2005, vol. 3400, pp. 37–56. [Online]. Available: http://dx.doi.org/10.1007/11427834_2

[19] Z. Hippe, "Machine learning: a promising strategy for business information processing?" in *Business Information Systems*, W. Abramowicz, Ed. Poznan: Academy of Economics Editorial Office, 1997, pp. 603–622.

[20] M. Knap, "Research on new algorithms for decision trees generation (in polish)," Ph.D. dissertation, AGH University of Science and Technology, Krakow, 2009.

[21] W. Paja and Z. Hippe, "Feasibility studies of quality of knowledge mined from multiple secondary sources. I: Implementation of generic operations," in *Intelligent Information Processing and Web Mining*, ser. Advances in Intelligent and Soft Computing, M. Klopotek, S. Wierzchon, and K. Trojanowski, Eds. Berlin Heidelberg: Springer-Verlag, 2005, vol. 31, pp. 461–465. [Online]. Available: http://dx.doi.org/10.1007/3-540-32392-9_53

[22] J. Gomuła, K. Pancerz, and J. Szkoła, "Copernicus - an expert system supporting differential diagnosis of patients examined using the MMPI test: an index-rule approach," in *Proceedings of the International Conference on Health Informatics (HEALTHINF 2011)*, V. Traver, A. Fred, J. Filipe, and H. Gamboa, Eds., Rome, Italy, 2011. doi: 10.5220/0003172503230328 pp. 323–328. [Online]. Available: http://dx.doi.org/10.5220/0003172503230328

[23] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1992.

[24] J. Gomuła, K. Pancerz, and J. Szkoła, "Rule-based classification of MMPI data of patients with mental disorders: Experiments with basic and extended profiles," *International Journal of Computational Intelligence Systems*, vol. 4, no. 5, 2011.

[25] ——, "Classification of MMPI profiles of patients with mental disorders - experiments with attribute reduction and extension," in *Rough Set and Knowledge Technology*, ser. LNAI, J. Yu *et al.*, Eds. Berlin Heidelberg: Springer-Verlag, 2010, vol. 6401, pp. 411–418. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-16248-0_58

[26] J. Gomuła, W. Paja, K. Pancerz, T. Mroczek, and M. Wrzesień, "Experiments with hybridization and optimization of the rules knowledge base for classification of MMPI profiles," in *Advances on Data Mining: Applications and Theoretical Aspects*, ser. LNAI, P. Perner, Ed. Berlin Heidelberg: Springer-Verlag, 2011, vol. 6870, pp. 121–133. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-23184-1_10

[27] D. Jachyra, K. Pancerz, and J. Gomuła, "Classification of mmpi profiles using decision trees," in *Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P 2011)*, M. Szczuka, L. Czaja, A. Skowron, and M. Kacprzak, Eds., Pułtusk, Poland, 2011, pp. 397–407.

1st Complex Events and Information Modelling

COMPLEX Events and Information Modelling (CEIM). Event modelling is a method of intelligent analyzing streams of information (data, percepts) about things that happen (events), and deriving conclusions from them. The goal of CEIM is to identify meaningful events and respond to them appropriately and quickly. We define the complexity of the event as both the complexity of the modelled physical phenomenon (fire, weather, chemical reaction, biological process) as well as the heterogeneity of the data (digital images, percepts, sensory data, natural language, semi-structured and structured data). In addition, the emphasis should be placed on the intelligent aspect of these models. This means that systems should semi- autonomously perceive their environment and take action.

The workshop on Complex Events and Information Modelling provides an interdisciplinary forum for researchers and developers. The workshop is mostly intended for applications of Artificial Intelligence in Fire Safety domain, but we also encourage researchers from other fields: health care, smart buildings, ubiquitous computing, process mining and others. The workshop covers the whole range of theoretical and practical aspects, technologies and systems and aims at bringing together specialists for exchanging ideas and promote interdisciplinary discussions.

Topics

- Artificial Intelligence techniques in Fire Safety,
- Data Assimilation and Smart Buildings,
- Evacuation models,
- Cognition and Decision Making models during Emergency,
- Sensory data storage representation and processing,
- Ubiquitous computing,
- Uncertainty modelling,
- Automated reasoning,
- Risk Management,
- Knowledge Discovery, Data and Process Mining,
- Decision Support Systems,
- Knowledge Modelling.

Event Chairs

Krasuski, Adam, The Main School of Fire Sevice, Poland
Rein, Guillermo, Imperial College London

Program Committee

Bargiela, Andrzej, University of Nottingham, United Kingdom

Bazan, Jan, University of Rzeszów, Poland

Butler, Bret W., Missoula Fire Sciences Laboratory

Chaudhury, Santanu, Indian Institute of Technology Dehli, India

Gałaj, Jerzy, The Main School of Fire Service, Poland

Gelenbe, Erol, Imperial College London, United Kingdom

Hamins, Anthony, National Institute of Standard and Technology, United States

Hostikka, Simo, VTT Technical Research Centre of Finland, Finland

Jahn, Wolfram, Raindance Science International, Chile

Jankowski, Andrzej, Warsaw University of Technology, Poland

Jin, Peng, Leshan Normal U. China

Liu, Jiming, Hong Kong Baptist University, Hong Kong S.A.R., China

Meina, Michał, Nicolaus Copernicus University

Mendonca, David, Rensselaer Polytechnic Institute, United States

Merci, Bart, Ghent University, Belgium

Mirończuk, Marcin, Institute of Computer Science Polish Academy of Sciences

Muzy, Alexandre, National Center for Scientific Research

Nguyen, Hung Son, University of Warsaw, Poland

Ohsawa, Yukio, The University of Tokyo, Japan

Pilemalm, Sofie, Linköping University, Sweden

Robinson, Karen, NICTA, Australia

Ronchi, Enrico, Lund University, Sweden

Roszkowska, Ewa, University of Bialystok

Rykaczewski, Krzysztof, Nicolaus Copernicus University

Salamonowicz, Zdzisław, The Main School of Fire Service

Sikora, Beata, Silesian University of Technology

Simeoni, Albert, University of Edinburgh

Tarapata, Zbigniew, Military University of Technology, Poland

Tavares, Rodrigo Machado, RMT Fire & Crowd Safety, Brazil

Velasquez Silva, Juan D., Web Intelligence Research Chile Centre

Welch, Stephen, University of Edinburgh, United Kingdom

Goczyła, Krzysztof, Gdansk University of Technology, Poland

Haralambous, Yannis, Institut Telecom - Telecom Bretagne, France

Homenda, Wladyslaw, Warsaw University of Technology, Poland

Jin, Qun, Waseda University, Japan

Kaczmarek, Janusz, Lódz University, Poland

Kakkonen, Tuomo, University of Eastern Finland, Finland

Krawczyk, Bartosz, Wroclaw University of Technology, Poland

Kulicki, Piotr, John Paul II Catholic University of Lublin, Poland

Lai, Cristian, CRS4, Italy

Leonelli, Sabina, University of Exeter, United Kingdom

Ludwig, Simone, North Dakota State University, United States

Martinek, Jacek, Poznan University of Technology, Poland

Mirenkov, Nikolay, University of Aizu, Japan

Mozgovoy, Maxim, University of Aizu, Japan

Nalepa, Grzegorz J., AGH University of Science and Technology, Poland

Palma, Raúl, Poznan Supercomputing and Networking Center, Poland

Piasecki, Maciej, Wroclaw University of Technology, Poland

Pyshkin, Evgeny, St. Petersburg State Polytechnical University, Russia

Reformat, Marek, University of Alberta, Canada

Shtykh, Roman, CyberAgent Inc., Japan

Slezak, Dominik, University of Warsaw & Infobright Inc., Poland

Soldatova, Larisa, Brunel University, United Kingdom

Suárez-Figueroa, Mari Carmen, Ontology Engineering Group, Scool of Computer Science at Universidad Politécnica de Madrid, Spain

Tadeusiewicz, Ryszard, AGH University of Science and Technology, Poland

Vacura, Miroslav, University of Economics, Czech Republic

Vazhenin, Alexander, University of Aizu, Japan

Wang, Haofen, Shanghai Jiao Tong University, China

Wu, Shih-Hung, Chaoyang University of Technology, Taiwan

Zadrozny, Slawomir, Systems Research Institute, Poland

Ławrynowicz, Agnieszka, Poznan University of Technology, Poland

Position papers of the 2014 Federated Conference on
Computer Science and Information Systems pp. 45–52

DOI: 10.15439/2014F475
ACSIS, Vol. 3

Evaluation of a Heat Release Rate based on Massively Generated Simulations and Machine Learning Approach

Mateusz Fliszkiewicz*, Adam Krasuski* and Karol Krenski*
*Section of Computer Science,
The Main School of Fire Service
Słowackiego 52/54,
01-629 Warsaw, Poland
fliszkiewicz@inf.sgsp.edu.pl, krasuski@inf.sgsp.edu.pl, krenski@inf.sgsp.edu.pl

***Abstract*—We present an approach for evaluation of a heat release rate of compartment fires. The approach is based on the idea of matching the actual condition of the fire to the pregenerated CFD simulations. We use an IR image of imprint of the temperature on the ceiling as a similarity relationship between actual fire and the set of the simulations. We extract the invariants, features and similarity measures of the fires using machine learning approach.**

***Index Terms*—Inverse Fire Modelling, Artificial Intelligence, Classification, Image Processing, Fire Services**

I. Introduction

MODELING of fire dynamics in buildings attracts many researchers from the fire safety science domain [1], [2]. The achievements of the research are successfully implemented in the processes of building design such as installation of fire detection and suppression, segmentation of the spaces and ventilation. All these activities are destined to prevent fire spreading in the building. However, there are rare cases when fire overcome these controls and is threatening the occupants and the building. In such cases the management of the emergency scene is delegated to the Incident Commanders (IC). Old buildings were designed with currently obsolete fire engineering approaches. In these buildings fires occur more frequently and generate much more losses. Currently, the forecasting of fire dynamics is bound to the designing of the buildings and not really to the management of the emergency scene by the commanders. This is the result of the complexity of the calculation, which forces the long computation times and the uncertainty of the input parameters for the fire modelling. The geometry of the room and the heat release rate (HRR) are the key input parameters for the successful simulation in the fuel controlled regime. The new achievements in the computer aided design (Building Information Management BIM[1]) allow to hope that in the near future the layouts of the insides of the buildings will be more available, as opposed to the availability of the HRR, unfortunately. The HRR is strongly dependent on the type, amount and distribution of combustible material inside the room, which is generally difficult to predict or evaluate. Therefore, a set of researches has been conducted which were focused on the evaluation of HRR based on the occurring fire [3], [4], [5]. Each of the approaches in the domain were grouped under one umbrella called inverse fire modelling. However, the approaches are strongly dependent on high sensor density in the compartment, which makes them infeasible during the F&R actions, taking into account the poor sensors infrastructure in buildings today.

In the previous works on evaluating of the HRR, the authors focused on creating the physical model of the phenomenon. In these models all the parameters represent the physical features. Moreover, their goal was to obtain the strict values of the HRR. These approach is not feasible in the case of F&R action, where the IC is interested in the approximation of HRR only. During the F&R action the evaluation of the HRR with accuracy of $\pm 100\,\mathrm{kW}$ is good enough to estimate the type of the fire (furniture, TV, curtains) and its future course of action. Therefore in the case of supporting the IC during the action some conditions can be relaxed.

In this paper we address the problem of inverse fire modelling. We analyze the possibilities of utilizing the inverse fire modeling approach in order to facilitate the management of the emergency scene by the IC. We present a new approach to inverse fire modeling which can be practically used on the emergency scene. The rest of the paper is structured as follows. Section II contains description of our approach and experiments conducted to validate our model. Section III introduces the results obtained in our experiments. In the Section IV we discuss the result obtained and the perspectives of future work for model improvements.

II. Method

Our method is based on the simple intuition that during the compartment fire, the unique – for a given type of fire – imprint of temperature on the ceiling is created. The imprint depends on a set of physical parameters such as HRR, the type of the materials the ceiling is produced from, the geometry of the compartment and others. We also assumed that if we relax

[1] http://en.wikipedia.org/wiki/Building_information_modeling

the conditions for the approximation of HRR purposes then the category of similar compartment will be created allowing for the reduction of invariants considered. It opens the way for running a set of simulations for each of the categories and store them in the database. Then, by comparing the imprint of the actual fire against the completed simulations in the database we can evaluate the HRR and the fire dynamics.

We also assumed that for the approximation purposes during the F&R action the physical interpretation of the parameters of our model are not important. It allows us for the usage of the approaches from artificial intelligence domain. According to these approaches the feature extraction and their impact on the decision class (HRR value in our case) is obtained automatically during the training process. There is only one demand: large enough amount of training data. This condition is however, easy to meet in our case, because we are able to massively generate the simulations.

In our analysis we focused mainly on the measurement of the temperature distribution on the structural elements of the compartment using IR camera. This method allows to obtain large number of individual measurement points. Number of points and quality of the obtained measurements depends on the accuracy of measuring device. The long-wave IR cameras have become increasingly common in the F&R actions [6], [7]. In many countries almost all fire departments have at least one IR camera. It helps rescuers to navigate through smoke filled compartments and also speeds up the localization of the fire source. Except from the camera, the rescuers could be equipped with the module for analyzing the temperature distribution on structural elements. This analysis could dynamically estimate many fire parameters such as the HRR, smoke layer temperature, fire area or critical conditions e.g. flash over (combined with information from BIM).

In most cases the thermal response depends on the building materials from which the compartment was made. Building obstructions are mainly heated by radiative and convective heat transfer from fire and conductivity. For large spaces, the uniform heating of walls and ceiling may prove to be a very long process. However, the ceiling above the fire source is always heated identically and comparatively fast. This approach can be universal for all compartments sizes. However, there are many factors that can affect the final imprint of temperatures. Ceiling temperature distribution depends on many variables i.e. HRR, combusted material, fire area, base of the fire, height of the compartment, type of building materials and many others but final imprint is related to the amount of heat released from fire which means that there is a physical dependence of these two parameters.

A. Preliminary experiments

We conducted the preliminary experiment in order to recognize, whether there are possibilities of using the machine learning algorithms to resolve the problem. We generated, for this purpose, a representative numer of simulations of various fire scenarios in a single compartment. We ran all simulations on the Fire Dynamics Simulator (FDS) [8] software version 6.0.1. FDS is a computational fluid dynamics model of fire-driven fluid flow, with an emphasis on smoke and the heat transport from fires. Launching the FDS simulation requires a large number of input parameters. The parameters affect the simulation results. In order to simplify the conditions of preliminary experiment we divided the input parameters into invariants and variants. The setup of invariant parameters is as follows:

- the properties of the building materials,
- height of the room – 2.6 m,
- base of the fire – 0.0 m,
- room dimensions – $5.0 \times 3.0\,\mathrm{m}$,
- burning material – ethyl alcohol,
- FDS model settings except radiation model,
- cell size – $0.2\,\mathrm{m}$,
- fire always is located at least $1.0\,\mathrm{m}$ away from the building walls,
- ventilation hole – entrance door with $1.0\,\mathrm{m}$ width and $2.0\,\mathrm{m}$ height,
- simulation time – 600 s.

The properties of the building materials are crucial in this analysis. Depending on physical properties the different rate of heat loss by the hot gases to compartment boundaries is transfered. We considered the room made of fire resistant calcium silicate boards (both the walls and the ceiling). The assumed physical properties are presented in the Table I.

TABLE I: The physical properties of calcium silicate boards

Parameter	Value
Density	$480\,\mathrm{kg/m^3}$
Thermal conductivity	$0.09\,\mathrm{W/(m \times K)}$
Specific heat at 293 K	$1.074\,\mathrm{kJ/(kg \times K)}$
Specific heat at 473 K	$1.000\,\mathrm{kJ/(kg \times K)}$
Emissivity	0.9

The parameter *specific heat* was introduced in order to define the boundary conditions. This parameter is a linear function of the temperature and is defined by two points: $y1 = 1.074$ at $293\,\mathrm{K}$ and $y2 = 1.0$ at $473\,\mathrm{K}$ (physical parameters of the boards). We also assumed that the height from the fuel source to the ceiling is constant (fire basis) for all the experiments. This parameter as well as the height of the compartment may strongly influence the temperature distribution on the ceiling. The higher is the fire basis the higher temperatures may be reached by ceiling jet. It mainly depends on the plume mass flow above the flames [1]. Moreover, the position of the fire in the room affects these temperatures. If the burning fuel is located close to the walls the cool air is entrained into the plume only from one or two directions. This causes higher temperatures and higher flames from the same fire [9].

We used ethyl alcohol as a burning material. It is commonly used in the real experiments because of its well known physical and chemical properties. It allows to precisely estimate the

HRR using the following function [10]:

$$H = 345 + 1139A - 1108A^2 + 320A^3$$

where H is HRR and A is the area of the fire.

As variant parameters we assumed the total HRR, the area of the fire and its location in the compartment. We considered steady-state fires with constant HRR which was changing in the range of 50 to 1,000 kW every 50 kW for all fire areas and locations. The total number of combinations were 600.

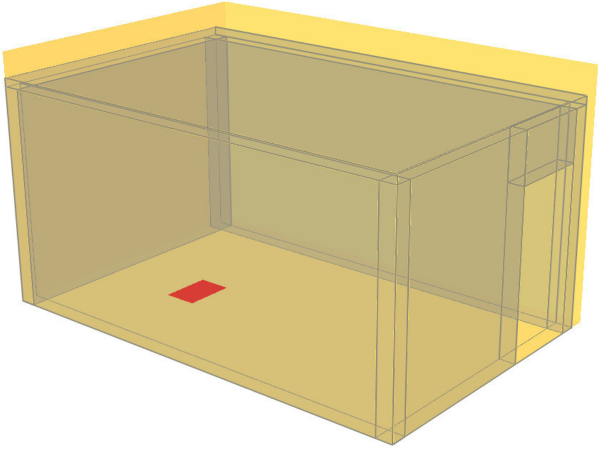


Fig. 1: The compartment view used in the CFD simulations.

Next step of our analysis was the extraction of the data from the simulations. The results e.g., the wall temperature was collected for each cell on all boundaries in the compartment. We recorded the values every 5 seconds. The binary output from the FDS was converted to comma separated values file (CSV) with the fds2ascii tool. All cells from the ceiling were inserted to monetDB[2] database. Each record contains the information about temperature, location (x, y coordinate) of the cell (z axis fixed) and other data needed to distinguish the fire scenario i.e. HRR, fire area and position. From the total ceiling area only $1\,\mathrm{m}^2$ was considered for the analysis – this area is chosen by spotting the hottest cell and centering the one square metre around it.

The data from the database was used as a training set for classifiers. We constructed our information system by defining the object as an avaraged values of attributes within defined (30 s) time window of the given fire test. The object in our information system was represented by following attributes: time from ignition, maximum temperature, average temperature form selected area and standard deviation. As a decision class the HRR was used. We used in our experiment the total number of objects (samples) equal 36,600.

Due to the large amount, the data were first discretized. The original values of HRR were in range from 50 to $1{,}000\,\mathrm{kW}$ with $50\,\mathrm{kW}$ growing step. We conducted two tests with HRR discretized into 20 and 10 equal-width sets. The time of simulation was in the range from 0 to 600 second. We discretized this parameter into 20 equal-width sets in both cases.

Then we used the orange-canvas[3] software to run the classifications. We used the 5-folds cross-validation technique for estimating the performance of predictive models. In 5-fold cross-validation, the simulations data were randomly partitioned into five equal size subsamples. Out of the five subsamples, a single subsample was retained as the validation data for testing the classifiers, and the remaining four subsamples were used as training data. The cross-validation process was then repeated five times, with each of the five subsamples used exactly once as the validation data. We averaged the five results from the folds to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

We used in our experiments two classification algorithms, i.e., Classification Trees [11], [12] and Support Vectors Machine (SVM) [13]. We decided to use these two algoritms because they are based on two different foundations. Classification Trees represents the algoritms with descriptive method of classification. The model created as a result of training process is interpretable and understandable by the fire safety domain experts – the trees with decision rules. The selected algorithm allowed us for better controling the experiments and keeping the physical interpretation of the obtained results. SVM represents respectively the algoritms with procedural method of classification. The model defined within trainig process is not interpretable by domain experts. However, we chose this algorithm for the comparision with Classification Trees and because of good performance in similar problems [14].

The goal of the classifiers was to predict the value of HRR from testing subsample. We used the Balanced Classification Accuracy (BCA)[4] measure for measuring the performance of the classifiers. At the current state of our researches, the conducted experiments were aimed at checking the possibility of using proposed method. Our intention was not to tune up classifier settings for better performace rather we used default options. We used the following settings of classifiers: SVM type: C-SVM (cost = 1, complexity bound = 0.50), SVM Kernel: RBF[5], SVM Numerical tolerance = 0.0010. Classification Tree attribute selection criterion: information gain, Classification Tree Pre-Running: min. instances in leaves = 2, pruning with m-estimate m = 2.

B. *Real experiments*

In order to check whether the method could be applied in the real situation, when the distribution of the temperature on the cell is obtained by the physical equipment, we conducted a set of real experiments. Three surveys of three different HRRs were made within the compartment with dimensions $5.25\,\mathrm{m}$ length, $2.54\,\mathrm{m}$ width and $2.55\,\mathrm{m}$ height. The whole building was made of prefabricated concrete slabs, however

[2] http://www.monetdb.com/Home

[3] http://orange.biolab.si/

[4] http://en.wikipedia.org/wiki/Accuracy_and_precision

[5] http://en.wikipedia.org/wiki/Radial_basis_function_kernel

the analyzed compartment was protected by non-combustible low density calcium silicate boards. Fire resistant boards were mounted on steel construction. The compartment had three exits with dimensions $2.1 \times 1.0\,\mathrm{m}$ each. Two of them were closed and the third one, the exit from the building was opened.

The fire source was placed in the compartment at least $1\,\mathrm{m}$ from the walls. $90\,\%$ methylated spirits was used as a burning material. Alcohol was burned in steel fire trays based on ISO international paper sizes i.e. A3 and A2, located directly on the floor. The total HRR was estimated according to Australian Standard [10] i.e. $60\,\mathrm{kW}$ $140\,\mathrm{kW}$ and $200\,\mathrm{kW}$ for A3, A2 and A3 + A2 fire trays, respectively. Tests were planned to reach steady-state conditions with constant HRR. During the tests the following parameters were recorded:

- imprint of the temperature on the ceiling captured by IR camera,
- temperature distribution in the compartment measured by thermocouple trees,
- video image.

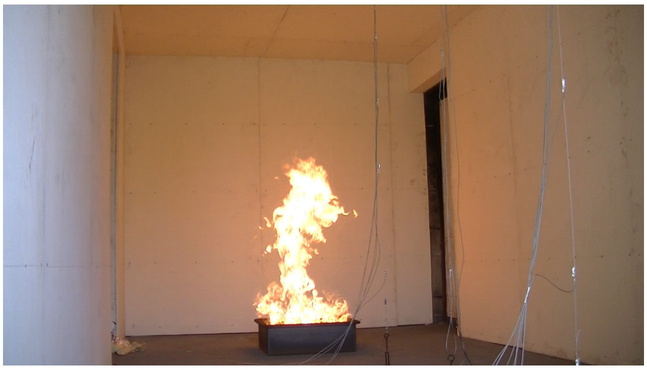


Fig. 2: Experimental setup with A2 fire tray.

Temperature distribution on the ceiling was captured by long-wavelength IR camera – wavelength $8 - 14\,\mu\mathrm{m}$. Resolution of the camera was 640×480 pixels. The distance from the measured area to the camera was approximately $2.7\,\mathrm{m}$. The pictures were captured every 1 second. We also measured the temperature distribution in compartment using three thermocouple trees which held 6 (Type-K) thermocouples each. These trees were spread in compartment with distance $1\,\mathrm{m}$, $2\,\mathrm{m}$, $3\,\mathrm{m}$ from center of the fire source. The values from the thermocouples were logged at frequency of $1\,\mathrm{Hz}$. One video camera was used to monitor the fire growth.

After the fire tests we processed 60 pictures from IR camera. We selected pictures form all tests with a $30\,\mathrm{s}$ step. Next we extracted the attributes needed for the predictions methods i.e. maximum temperature, average temperature and standard deviation from $1\,\mathrm{m}^2$. Finally we used already trained classifiers from preliminary experiments to match most suitable decision class.

C. *Towards the Generalization of the Model*

The goal of the experiments described above was a general validation of our approach. It allows for proving that the concepts are reasonable and for evaluating the results. However, the method, even if it obtained quality is satisfying, is not quite practical so far. To make the approach usable, the generation of massive number of simulations for every particular compartment is required. Moreover, taking into account that for a given compartment different conditions of the ventilation could appear (windows and doors could be opened or closed) the generation of the simulation scenarios requires some sampling method such as Monte Carlo. Therefore our next experiments were concentrated on the generic features of the compartments. The extracted features allow to create groups of similar compartments – from the fire scenario point of view. Such grouping, thus allows for relaxing the conditions of generation of the simulations for every particular compartment.

Similarly to preliminary experiments we set a couple of invariant parameters. Most of them were the same except from the following:

- building materials properties,
- height of the room – $2.8\,\mathrm{m}$,
- simulation time – $1{,}200\,\mathrm{s}$.

This time the test was performed in the room made of gypsum boards. It is a more common material for typical offices or flats. Moreover, we changed the height of the compartment. The previous one was matched to the test room where the real experiments were made.

TABLE II: Variant parameters assumed in simulations.

Parameter	Number of combination
Room dimensions	14
Number of the fire location	140
Area of the fire	10
Total heat release rate	10
Total	14,000

The assumed variant parameters were presented in Table II. Dimensions of the room were changing from $3\times3\,\mathrm{m}$ to $7\times7\,\mathrm{m}$ which is corresponding to the real conditions in a dwelling or an office. The increment step was set to $1\,\mathrm{m}$ in both directions except from the repeated dimensions. Also the location of the fire was moving by $1\,\mathrm{m}$, however, the minimal distance from obstructions was set to $1\,\mathrm{m}$. We assumed ten combinations of fire area and total heat release rate i.e. from $0.4 \times 0.4\,\mathrm{m}^2$ to $1.0\times1.0\,\mathrm{m}^2$ and from $100\,\mathrm{kW}$ to $1{,}000\,\mathrm{kW}$, respectively. The total numer of combinations was 14,000.

Due to the large number of generated input files the coarse grid size of $20\,\mathrm{cm}$ was chosen. According to previous works [4] and results from preliminary experiments we decided to study the case with the mentioned size of the grid.

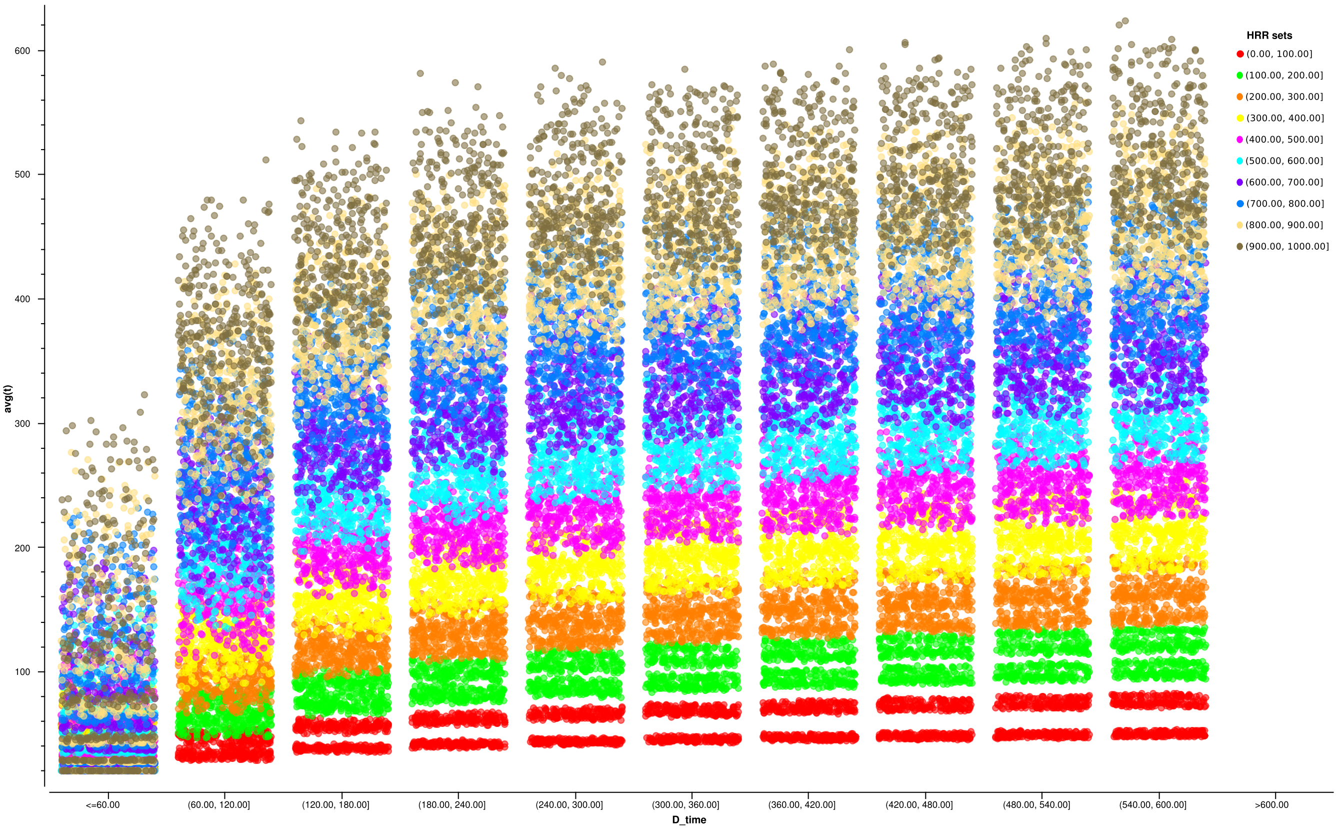


Fig. 3: Temperature distribution on the ceiling in the simulation. Legend: X-axis – discretized time, Y-axis – average temperature.

We chose a similar methodology as in the preliminary experiments for preparing and processing the data for classification methods. We decided upon 10 discrete equal-width sets of the HRR.

III. Results

The results were collected in each step of the experiments. Firstly we tested the classification accuracy for preliminary experiments. Then we compared the results from the classifiers against the real experiments. Finally we measured the classification accuracy for generalized model. In all steps we considered BCA for Classification Tree and Support Vector Machine (SVM) classifiers. We also provide the confusion matrix[6] as the measure of classifiers' performance.

A. *Preliminary experiments*

The preliminary experiments aimed at verifying whether the machine learning approach is appropriate for characterizing the HRR from the temperature imprint on the ceiling. We prepared 600 simulations for a single room with varying fire area, fire localization and Heat Release Rate Per Unit Area (HRRPUA). For all the simulations we registered the time, the average and maximum temperatures and the standard deviation of the temperatures in the measured (1 square metre) area. In total there were 36,600 records registered.

The Figure 3 illustrates the dependence of the HRR on the average temperature and time. In this case the HRR were discretized into 10 sets. Each set is presented in a distinct color. The figure shows that the sets of the HRRs are quite separable, especially for the lower HRR. For the higher HRR the classification is less certain. However, the additional attributes – maximum temperature and the standard deviation of the temperatures improve the classification of the HRR.

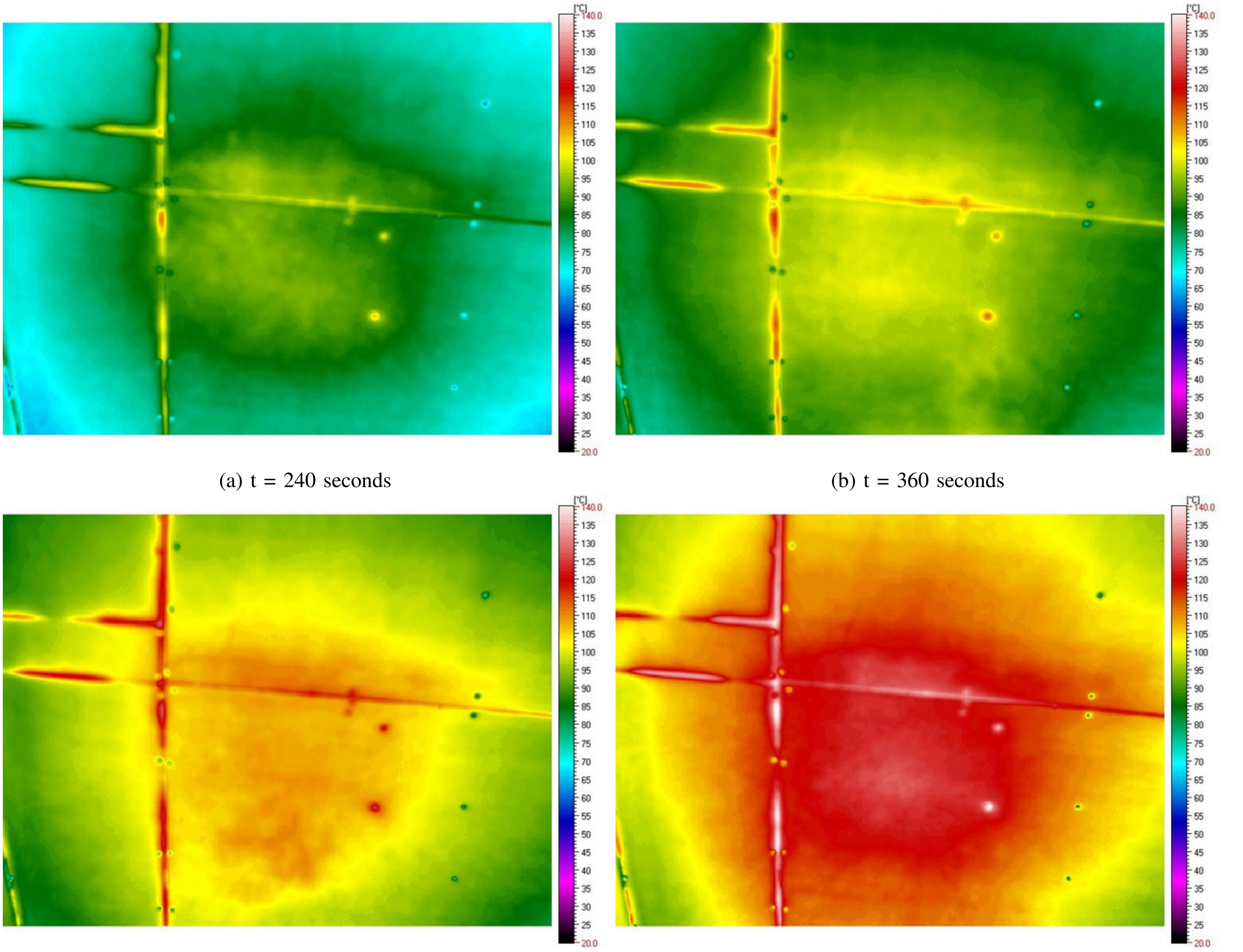
In the Table III we collected the results of the classification accuracy. It shows the results of the test learners for all the classifiers used. We distinguish two groups of the results dependent on the HRR discretization. First group assumes discretization into 20 sets (every 50 kW) and the second into 10 sets (every 100 kW).

TABLE III: The test learners obtained by the classifiers in the preliminary experiments.

Test learners	BCA	BCA	AUC	AUC
(HRR step)	50 kW	100 kW	50 kW	100 kW
Classification Tree	0.6292	0.7713	0.9745	0.9701
SVM	0.3711	0.6727	0.9619	0.9781

In the Table IV and V we present the confusion matrix for the Classification Tree. The results are for best and worst HRR

[6] http://en.wikipedia.org/wiki/Confusion_matrix



(a) t = 240 seconds

(b) t = 360 seconds

(c) t = 480 seconds

(d) t = 600 seconds

Fig. 4: Temperature distribution on the ceiling.

classification classes for 20 and 10 equal-width sets (50 and 100 kW HRR step). It can be observed that the accuracy of the classifier is high because of the low number of false negatives and because these false negatives were assigned to the nearby sets.

B. Real experiments

The purpose of the study was to validate the classification methods against the results of the full scale experiments. During the real experiments we captured the imprint of the temperature on the compartment ceiling using IR camera. The pictures were captured every second. However, only sixty pictures were used to extract the average and maximum temperatures and the standard deviations of the temperatures. Figure 4 presents the imprints of the ceiling temperature for burning A2 fire tray.

Afterwards we used the prediction module in orange-canvas to verify the classifiers accuracy. In Table VI we present the performance of the classifiers for HRR discretized into 20 and 10 equal-width sets.

TABLE VI: The performance obtained by the classifiers in the real experiments.

Classifier	BCA	BCA
	50 kW	100 kW
Classification Tree	0.5833	0.9167
SVM	0.6667	0.900

C. Generalized model

In Table VII we present the performance obtained by the classifiers. The total number of all instances was 293,976.

TABLE IV: Confusion matrix for Classification Tree (50 kW step).

	(0.00, 50.00]	(100.00, 150.00]	(200.00, 250.00]	âĂę	(700.00, 750.00]	(800.00, 850.00]	(900.00, 950.00]	
(0.00, 50.00]	1,781	7	0	âĂę	0	0	0	1,830
(50.00, 100.00]	13	15	2	âĂę	0	0	0	1,830
(100.00, 150.00]	5	1705	7	âĂę	0	1	0	1,830
(150.00, 200.00]	4	43	57	âĂę	1	0	0	1,830
(200.00, 250.00]	1	11	1578	âĂę	3	0	0	1,830
(250.00, 300.00]	0	4	91	âĂę	0	0	3	1,830
(300.00, 350.00]	0	7	22	âĂę	4	0	1	1,830
(350.00, 400.00]	0	2	16	âĂę	7	1	0	1,830
(400.00, 450.00]	0	8	6	âĂę	5	3	4	1,830
(450.00, 500.00]	0	3	10	âĂę	8	3	9	1,830
(500.00, 550.00]	0	0	6	âĂę	13	2	0	1,830
(550.00, 600.00]	0	0	3	âĂę	21	8	5	1,830
(600.00, 650.00]	0	1	2	âĂę	69	10	5	1,830
(650.00, 700.00]	0	4	0	âĂę	381	24	6	1,830
(700.00, 750.00]	0	1	7	âĂę	735	72	18	1,830
(750.00, 800.00]	0	0	13	âĂę	378	370	21	1,830
(800.00, 850.00]	0	0	2	âĂę	89	757	101	1,830
(850.00, 900.00]	0	0	0	âĂę	23	379	375	1,830
(900.00, 950.00]	0	0	0	âĂę	23	123	661	1,830
(950.00, 1,000.00]	0	0	0	âĂę	9	44	453	1,830
	1,804	1,811	1,822	âĂę	1,769	1,797	1,662	36,600

TABLE V: Confusion matrix for Classification Tree (100 kW step).

	(0.00, 100.00]	(100.00, 200.00]	(200.00, 300.00]	âĂę	(700.00, 800.00]	(800.00, 900.00]	(900.00, 1,000.00]	
(0.00, 100.00]	3,552	41	25	âĂę	23	10	0	3,660
(100.00, 200.00]	49	3,430	91	âĂę	22	16	0	3,660
(200.00, 300.00]	14	101	3,248	âĂę	35	13	4	3,660
(300.00, 400.00]	5	38	243	âĂę	41	21	2	3,660
(400.00, 500.00]	4	27	48	âĂę	35	29	18	3,660
(500.00, 600.00]	2	21	32	âĂę	80	35	17	3,660
(600.00, 700.00]	0	21	18	âĂę	594	55	34	3,660
(700.00, 800.00]	0	17	28	âĂę	2,285	607	60	3,660
(800.00, 900.00]	0	20	18	âĂę	624	2,185	696	3,660
(900.00, 1000.00]	0	12	25	âĂę	113	634	2,800	3,660
	3,626	3,728	3,776	âĂę	3,852	3,605	3,631	36,600

TABLE VII: The performance obtained by the classifiers in the generalized method.

Classifier	BCA
Classification Tree	0.6569
SVM	0.5436

IV. Discussion

The presented results prove that there is a potential in the described approach. We reached a high value of BCA (0.77) in the preliminary experiments. No less optimistic are the results of the real experiments with the IR camera. These experiments resulted in BCA equal 0.92. The separability of the (especially lower) HRR in the analyzed data were observed. In order to improve the performance in higher HRR, the observed area of the ceiling should be extended by lower focal length of the objective of IR camera. There is a room for other more sophisticated statistics and operations on the features from the pictures, e.g. calculating the parameters of spatial distribution of temperature instead of the mean etc.

The main goal of the research was to asses whether the proposed approach may be used on the fire ground. The results from the real experiments showed that this method may be used to characterizing HRR. The BCA ratio for full scale experiments could be higher, however there were problems with selecting correct values of the attributes from IR camera pictures. In many cases products of combustion i.e. water vapour and aerosols radiation veil created imprint on the ceiling. In order to overcome these shortcomings we consider processing of the images to get the lowest temperatures from the analyzed set of frames. For this purpose we will record the sequence of frames and then use the fluctuating phenomenon of flame and gases to obtain the points of interest. Obviously we may face further problems with more dense smoke where soot yield is higher and the combustion reaction is incomplete. However, we also consider to use mid-wave IR camera to find regions in infrared spectrum where these particles emit less electromagnetic wave.

The proposed method has a number of general assumptions. Most of them determined the final results of generated simulations and the type of the temperature imprint. In our opinion most decisive parameters are the fire base, fire localization and physical properties of the building materials. In all tests we located the fire on the floor and at least 1 m from the walls. Both of these parameters may strongly affect the amount of the incoming air to the convection column which influences the temperatures reached in combustion process. Moreover various building materials show different speed of conduction process. Even though the density, specific heat, thermal conductivity or emissivity may affect the results, we hope to distinguish the specific groups of materials in common buildings.

Further problems may be encountered if the fire varies in its area and its total HRR. In our analyses we considered the fire with fixed area and constant HRRPUA. Both factors have crucial influence on temperature imprint on the ceiling, however to check usefulness of machine learning method these assumptions were made.

Another problem may arise when few fire sources occur in considered compartment. Our method can deal with only one source of fire. The main goal is to localize the warmest area on the ceiling and then extract the needed attributes. This assumption may be inappropriate if there are multiple sources of fire.

At this stage of the research we are also uncertain about the number of exemplars of compartment which should be used in order to evaluate most of the fire scenarios. If we were to consider each small difference in geometry, ventilation conditions and other parameters then we would be forced to generate and store an enormous amount of simulations which is impractical. However, we choose the sensible experiments scenarios and thus limit their number.

One of the parameters which was used as a descriptor of the current condition of the fire (except from the temperature) was time from the beginning of the fire. This requires that the IC provides the accurate (within 1 minute) time of the ignition. Less experienced ICs may be not able to evaluate the time so accurately, which makes the whole approach fail. This will not be an issue for the fires detected by fire detecting systems with the time log.

In our further works we will focus on finding principal factors determining the results of the simulations. For this purpose we will consider the application of Principal Component Analysis (PCA)[7], Non-negative matrix factorization (NMF)[8] methods or rough set approach[9]. All these approaches will be used in order to determine the most important features which affect the temperature imprint on the ceiling. This allows us for the more comprehensive addressing the problem of reduction of the number of generated simulations for a specific compartment. After that we will try to generate a new set of the simulations including research results and create new classifiers for this set. Even later we will prepare the full-scale experiments with various ceiling materials and various HRR.

Acknowledgments

The research was supported by the Polish National Centre for Research and Development (NCBiR) - Grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: “Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings”.

References

[1] D. Drysdale, *An introduction to fire dynamics*. John Wiley & Sons, 2011.

[2] B. Karlsson and J. Quintiere, *Enclosure fire dynamics*. CRC press, 2002.

[3] W. Jahn, G. Rein, and J. L. Torero, “Forecasting fire growth using an inverse zone modelling approach,” *Fire Safety Journal*, vol. 46, no. 3, pp. 81–88, 2011.

[4] W. Jahn, J. L. Torero, and G. Rein, “Forecasting fire dynamics using inverse computational fluid dynamics and tangent linearisation,” *Advances in Engineering Software*, vol. 47, no. 1, pp. 114–126, 2012.

[5] K. J. Overholt and O. A. Ezekoye, “Characterizing Heat Release Rates Using an Inverse Fire Modeling Technique,” *Fire Technology*, vol. 48, no. 4, pp. 893–909, 2012.

[6] J. Martinez-de Dios, B. C. Arrue, A. Ollero, L. Merino, and F. Gómez-Rodríguez, “Computer vision techniques for forest fire perception,” *Image and vision computing*, vol. 26, no. 4, pp. 550–562, 2008.

[7] J. San-Miguel-Ayanz and N. Ravail, “Active fire detection for fire emergency management: Potential and limitations for the operational use of remote sensing,” *Natural Hazards*, vol. 35, no. 3, pp. 361–376, 2005.

[8] R. M. J. F. C. W. K. O. K. McGrattan, S. Hostikka, *Fire Dynamics Simulator, User's Guide*, 2013.

[9] B. Y. Lattimer, “Heat Fluxes from Fires to Surfaces,” in *The SFPE Handbook of Fire Protection Engineering, 3^{rd} Edition*, 2002, pp. 2–269 – 2–296.

[10] *Australian Standard AS 4391-1999, Smoke managment systems - Hot smoke test*, 1999.

[11] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, “Top 10 algorithms in data mining,” *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[12] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[13] M. A. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, “Support vector machines,” *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[14] S. Wang, W. Zhu, and Z.-P. Liang, “Shape deformation: Svm regression and application to medical image segmentation,” in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 209–216.

[7] http://en.wikipedia.org/wiki/Principal_component_analysis

[8] http://en.wikipedia.org/wiki/Non-negative_matrix_factorization

[9] http://en.wikipedia.org/wiki/Rough_set

7th Workshop on Computational Optimization

Many real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

We invite original contributions related to both theoretical and practical aspects of optimization methods.

Topics

The list of topics includes, but is not limited to:

- unconstrained and constrained optimization
- combinatorial optimization
- global optimization
- multiobjective optimization
- optimization in dynamic and/or noisy environments
- large scale optimization
- parallel and distributed approaches in optimization
- random search algorithms, simulated annealing, tabu search and other derivative free optimization methods
- nature inspired optimization methods (evolutionary algorithms, ant colony optimization, particle swarm optimization, immune artificial systems etc)
- hybrid optimization algorithms involving natural computing techniques and other global and local optimization methods
- optimization methods for learning processes and data mining
- computational optimization methods in statistics, econometrics, finance, physics, medicine, biology, engineering etc

Event Chairs

Fidanova, Stefka, Academy of Sciences, Bulgaria
Mucherino, Antonio, IRISA, France
Zaharie, Daniela, West University of Timisoara, Romania

Program Committee

Bartl, David, University of Ostrava, Czech Republic
Breaban, Mihaela
Cremonesi, Paolo
Gualandi, Stefano
Hoai Ann, Le Thi
Hosobe, Hiroshi, Hosei University, Japan
Iiduka, Hideaki, Kyushu Institute of Technology, Japan
Lavor, Carlile, IMECC-UNICAMP, Brazil
Marinov, Pencho, Bulgarian Academy of Science, Bulgaria
Michini, Carla
Miettinen, Kaisa, University of Jyvaskyla, Finland
Mihalas, Stelian, West University of Timisoara
Muscalagiu, Ionel, Politehnica University Timisoara
Nannicini , Giacomo
Parsopoulos, Konstantinos, University of Patras
Pop, Petrica
Roeva, Olympia, Institute of Biophysics and Biomedical Engineering, Bulgaria
Siarry, Patrick, Universite Paris XII Val de Marne, France
Slezak, Dominik, University of Warsaw & Infobright Inc., Poland
Stefanov, Stefan, South-West University ""Neofit Rilski, Bulgaria
Stuetzle, Thomas, Université Libre de Bruxelles (ULB), Belgium
Suganthan, Ponnuthurai Nagaratnam, Nanyang Technological University, Singapore
Tamir, Tami, The Interdisciplinary Center (IDC), Israel
Tvrdik, Josef, University of Ostrava, Czech Republic
Vrahatis, Michael, University of Patras, Greece
Wolfler Calvo, Roberto
Zilinskas, Antanas, Vilnius University

DOI: 10.15439/2014F381
ACSIS, Vol. 3

Probabilistic traveling salesman problem: a survey

Abir Henchiri
Laboratory RIADI, University of Manouba,
National School of Computer Sciences (ENSI),
Tunisia.
Email: abirhenchiri89@gmail.com

Monia Bellalouna, Walid Khasnaji
Laboratory CRISTAL-GRIFT, University of Manouba,
National School of Computer Sciences (ENSI),
Tunisia.
Email: monia.bellalouna@gmail.com,mwkhaznaji@yahoo.fr

***Abstract*—The Probabilistic Traveling Salesman Problem (PTSP) is a variation of the classic Traveling Salesman Problem (TSP) in which only a subset of potential nodes needs to be visited on any given instance of the problem. The number of nodes to be visited each time is a random variable. The objective is to find an a priori tour which minimizes the expected length,with the strategy of visiting the present nodes in a particular instance in the same order as they appear in the a priori tour. In this paper, we survey a number of results obtained for PTSP and we present the different approaches used for solving it.**

I. Introduction

OVER the past fifty years, the study of Combinatorial Optimization Problems (COP) has continued to grow in importance and has become one of the most active branches of discrete mathematics. This deterministic model is not adequacy with reality, where often the number of data of the studied problem is a random variable in [[0; n]].

In the late of 80's, studies have developed on a class of combinatorial optimization problems, characterized by the fact that probabilistic elements are explicitly associated with data i,e, given an instance of the problem, only a subinstance of it will eventually be solved. This concept was called Probabilistic Combinatorial Optimization Problems (PCOPs) and was initially introduced by Jaillet [1], [2].

There are several motivations for studying the effect of including probabilistic elements in combinatorial optimization problems. The two most important motivations are, firstly, the desire to define and analyze models which are more appropriate with reality where randomness is a major source of concern. For example, for many delivery companies, only a subset of their customers requires a delivery each day. Ideally we would like to re-optimize, i .e., find an optimal TSP tour for every day. However, we may not have the resources to do this, or even if we have them it may be very time consuming to do that. It is therefore necessary to adopt a model that takes into account random phenomena. Secondly, the possibility to analyze the stability of optimal solutions to deterministic problems when the instances are disturbed by the absence of certain data.

The first problem studied in the probabilistic combinatorial optimization problems was the probabilistic traveling salesman problem [1]. Later, this approach was extended to other problems such as the probabilistic vehicle routing problem [3], the probabilistic spanning tree problem [4]. Studies on this probabilistic approach continued in many others domains such as the probabilistic maximum independent set problem [5], [6], the probabilistic longest path problem [7], the probabilistic minimum vertex covering problem [8], the probabilistic minimum coloring problem [9], the probabilistic graph-coloring in bipartite problem [10] and the probabilistic steiner tree problem [11]. The probabilistic approach has been extended on the combinatorial problems not defined on graph such probabilistic bin packing problem [12], [13] and the probabilistic scheduling problem [14].

In this paper, we interest to the well-known problem in optimization under uncertainty: the probabilistic traveling salesman problem (PTSP). The organization of the paper is as follows: section 2 presents the PCOP and its formulation and section 3 is devoted to the presentation of the research aspects of PCOP. Section 4 gives the definition and the formulation of the PTSP. A review of the main results in the literature is presented in section 5. Then, solving methods for the resolution of PTSP are presented in section 6. Finally the last section gives the concluding remarks.

II. Probabilistic combinatorial optimization problems

The probabilistic combinatorial optimization problems, noted PCOP are generalized versions of COP. Formally, a PCOP is defined as follows: Let $L_n = x_1, x_2, ..., x_n$ a finite set of data and S a finite set of feasible solutions. Consider a cost function $f : S \to R$. In this model, we define on L_n a probability law: each element $x_i \in L_n$ has a probability of p_i. The problem consists in resolving only on a subset of L_n and then the size of the problem is a random variable. The objective is to minimize the expected objective function $E[f]$ through all parts I of $P(L_n)$.

The most natural approach that comes in mind is to consider each potential instance as a new problem defined through the present data and to optimally solve the instance considered. This approach is called $reoptimization\ strategy$. This approach is optimal, however, it can be very much time and space consuming, in particular when the combinatorial optimization problem considered is NP-hard [15], [16], [17].

It is therefore necessary to adopt another resolution strategy, which is less costly in terms of computations. This new approach is called an *a priori optimization* and has been introduced in [1], [18]. It consists of determining a solution of the initial instance, where all data are present, called an *a* *priori solution*, and applying a strategy called *a modification strategy* to adapt as quickly as possible the *a priori solution* to the subinstance that must effectively be solved.

III. The research aspects of PCOP

Since the introduction of probabilities in the formulation of COP [1], works on this subject have been considerable. Thus, we will present in this section the different research aspects addressed to these new problems.

A. Study of complexity of the different strategies

This research aspect concerns advanced study of the border between easy PCOP and PCOP hard. Most PCOP have been proven as difficult problems: like the probabilistic generalization of the shortest path which is NP-hard while the deterministic version is in *P* and very easily solvable [18]. An interesting line of research is to try to find for a given COP case in which the probabilistic version remains easy, for example Bellalouna and al [19] found particular cases for the PTSP to be solved in an easy way.

B. Asymptotic behaviour

The asymptotic analysis of PCOPs (when the size of the problem tends to infinity) is an important and successful area of research. The results obtained in this field are very useful for several reasons, they allow to obtain approximations for problems of very large size, analyze the performance of some heuristics and finally to explore the boundary between good and bad algorithms in the probabilistic sense. An Asymptotic study was proposed for the first time by Jaillet [20] for the probabilistic traveling salesman problem. The 2-dimensional probabilistic bin packing problem was asymptotically studied by Bellalouna and al. [13].

C. Stability Analysis

Several research works studied the stability of COPs. For example, the studies of Hromkowic [21], Forlizzi and al. [22] etc. The objective is to study the interdependence between a solution of a given COP and the parameters that define the problem. For the probabilistic version, we call a PCOP stable if the real random variables associated to the re-optimization strategy and to the *a priori* strategy follow the same law.

Several studies have been devoted to the study of stable problem. We quote here Bellalouna [23], who has interested to the study of stability of the probabilistic traveling salesman problem , Boria and al [24] studied the stability of probabilistic min spanning tree in complete graph, Bouyahia and al [14] analyzed the stability of the probabilistic scheduling problems.

D. Solving methods

Several algorithms were implemented for the resolution of PCOP and have shown very satisfactory performance. Exact methods have been used to solve the POCP such as exact branch and bound algorithm developed by Rosenow [25] for the probabilistic traveling salesman problem. Besides that, approximate methods have received wide interests in researchers'effort to solve large scale PCOP. Among the applied approximate methods we cite the work of Bertsimas [3] who proposed and analyzed heuristics for probabilistic vehicle routing problem. In the works of Bellalouna and al ([23], [26], [12]) algorithms based on classic heuristics were proposed for the probabilistic bin packing problem. Metaheuristics were also used to solve POCP for example for probabilistic traveling salesman problem simulated annealing and tabu search were implemented [26]. A Tabu Search was implemented by Gendreau and al. [27] for the vehicle routing problem with stochastic demands and customers. Experimental studies allow to choose the best parameters for these solving methods in the probabilistic framework.

IV. Definition of the PTSP

The Probabilistic Traveling Salesman Problem (PTSP) is a variation of the classic Traveling Salesman Problem (TSP) and is introduced for the first time by Jaillet [1] in which only a subset of the nodes may be present in any given instance of the problem. The goal is to find an *a priori* tour of minimal expected length, with the strategy of visiting the present nodes in a particular instance in the same order as they appear in the a priori tour ([20], [28], [29]). The TSP can be treated as a special case of the PTSP. The main difference between PTSP and TSP is that in PTSP the probability of each node being visited is between 0 and 1 while in TSP the probability of each node being visited is 1. In a given instance, the nodes present should be visited based on the sequence of the *a priori* tour while the others nodes will simply be skipped.

A formulation of the problem is the following [1]: We are given an a priori PTSP tour t through n points of a given graph G. Each point i is present with a probability p_i independently of the others. Let $d(i,j)$ the distance between points i,j and we assume, without loss of generality, that the *a priori* tour is $T=(1,2,...,n,1)$, then our problem is to find an *a priori* tour through all n potential nodes, which minimizes the expected length of a specific *a priori* PTSP tour T, denoted $E[L_T]$:

$$E[L_T]=\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}p_ip_j \prod_{k=i+1}^{j-1}(1-p_k)$$
$$+\sum_{j=1}^{n}\sum_{i=1}^{j-1} d_{ji}p_ip_j \prod_{k=j+1}^{n}\prod_{k=1}^{i-1}(1-p_k) \qquad (1)$$

When all points i have the same probability of presence($p_i = p\,\forall i$) ,we note $q=1-p$. In this case the expected length $E[L_T]$ is:

$$E[L_T]=p^2\sum_{r=0}^{n-2} q^r L_{T^r} \qquad (2)$$

Where $L_{T^r}=\sum_{i=1}^{n} d(i,T^r(i))$. T^r consists in jumping r points from the initial tour T, therefore T^r is formed by $pgcd(n,r+1)$ sub-tours. $T^r(i)$ the point after i along the permutation T^r, thus L_{T^r} is the length of the permutation T^r. We note that T^0 is the tour T and L_{T^0} is the length of the tour T.

V. Review of previous results

In [1], Jaillet derived several theoretical properties of optimal tours including the fact that such a tour may intersect itself. In [20], he bounded the relationship between optimal PTSP and TSP solutions. The analysis presented in [20] implies that under specific conditions the optimal PTSP and TSP solutions coincide. Later, others works examined some further properties of the PTSP and improved some bounds proved by Jaillet.
In this section we briefly review here the previous results.

A. Properties of PTSP

The TSP is a special case of the PTSP in which all the nodes are present ;so it is interesting to understand the relationship between the TSP and PTSP. In Jaillet's dissertation he found very special cases where TSP is stable.

Property 1. The optimum TSP tour is guaranteed to solve the PTSP optimally for problems with only 5 or fewer points (and with only 3 or fewer points when the matrix of distances between points is not symmetric).

Property 2. When the n points lie at the corner points of a convex n-gon then TSP is stable.

In [20] he exhibited examples where TSP is a very bad solution for PTSP. Let us denote by T_{PTSP} the optimal PTSP tour and by T_{TSP} the optimal TSP tour.

Property 3. Given $D = (d_{ij})$ the distance matrix through the n points. If D satisfies the triangle inequality, if the number of deterministically present points is m, then for $m \geq 1$ or (for m = 0 and n prime) :

$$E(L_{T_{PTSP}}) \geq p(1-(1-p)^{n-1})L_{T^0_{PTSP}} \tag{3}$$

$$\frac{E(L_{T_{TSP}})-E(L_{T_{PTSP}})}{E(L_{T_{PTSP}})} \leq \frac{1-p}{p}$$

$$\frac{L_{T^0_{PTSP}}-L_{T^0_{TSP}}}{L_{T_{TSP}}} \leq \frac{1-p}{p^2}$$

These bounds are the best possible. We note that if p is close to 1, the error $\frac{1-p}{p}$ is close to zero and therefore the TSP tour is a very good approximate solution for the PTSP. If p is very small the error $\frac{1-p}{p}$ tends to infinity and there is no information about the behavior of the optimal tour under TSP as solution for PTSP.

The PTSP seems much more complex than the TSP, The following results underscore the point that the probabilistic aspects of the PTSP induce some characteristics which are distinctly different from those of the TSP:

Property 4. For Euclidean plane, the optimal PTSP tour may intersect itself .

Property 5. The dynamic programming approach proposed for TSP fails to solve the PTSP.

Property 6. Given $h = (l, ...n)$ a path through n vertices, if h is decomposed into two paths $h_1 = (1, ..k)$ and $h_2 = (k, ..., n)$ and if we note $h = h_1 \oplus h_2$ then :

$$E(L_{h_1 \oplus h_2}) \leq E(L_{h_1}) + E(L_{h_2}) \tag{4}$$

Functional associated to PTSP is not additive and we cannot decompose the PTSP into sub problems. The optimality principle is not verified.

Bellalouna [23], [26] found special cases where the problem is polynomial and showed that under some conditions on the distance matrix denoted C, TSP is stable.

She gave conditions on the constant matrices of the form $c_{ij} = a_i + b_j$ basing on a result of Berenguer [30], showing that constant matrices are the only matrices where all the permutations of vertices have the same length.

Property 7. Constant matrices are the only ones that have the same expectation for every a priori tour T:

$$E(L_{T_{PTSP}}) = p(1-(1-p)^{n-1})L_{T^0_{TSP}} \tag{5}$$

In this case PTSP is polynomial.

Let us call a matrix C small if there exist two vertices, a and b, such that $c_{ij} = min\{a_i, b_i\}, i, j = 1.....n$. C is called small with distinct values where a_i and b_j are all distinct. In this case, let d_i be the $i-th$ smallest value between the $2n$ values a_k and b_j ; $D = \{d_1, ...d_n\}, \overline{D} = \{d_{n+1}, ...d_{2n}\}$ and $d = \sum_{i=1}^{n} d_i$. The vertices can be partitioned into four sets: $D_2 = \{i : a_i, b_i \subseteq D\}, D_o = \{i : a_i, b_i \subseteq \overline{D}\}, D_a = \{i : a_i \in D, b_i \in \overline{D}\}, D_b = \{i : b_i \in D, a_i \in \overline{D}\}$.

We remind Gabovitch's theorem [47] where he shows that for small matrices the TSP is an easy problem. In particular, he shows that the length of the optimal tour is equal to d if and only if D satisfies one of the following conditions:
(i) $D_2 \neq \emptyset$,
(ii) $D = \{a_l.....a_n\}$,
(iii) $D = \{b_1.....b_n\}$.

And if the length of the optimal tour is different from d then the length of the optimal tour equals $d' = d - d_n + d_{n+l}$ if and only if $D' = D \cup d_{n+l}\{d_n\}$ satisfies one of the following conditions:
(i') $D'_2 \neq \emptyset$, where D'_2 is defined analogously to D_2,
(ii') $D' = \{a_1.....a_n\}$,
(iii') $D' = \{b_1, ...b_n\}$.

Based on the results of Gabovitch [31], Bellalouna [23] gave conditions under which the TSP is stable.

Property 8. Let C be a small matrice.Then $T'_{PTSP} = T'_{TSP}$ if and only if $[(d_n = b_l) \vee [(d_n = a_l) \wedge (d_{n-1} = b\ )] \wedge [(d_{n+l} = a_n) \vee [(d_{n+l} = b_n) \wedge (d_{n+2} = a_n)]]$. In this case:

$$E(L_{T_{PTSP}}) = p(1-(1-p)^{n-1})d' - p^2(d_{n+1} - d_n) \tag{6}$$

Property 9. Let C a small with distinct values and consider the following conditions:
(cl) $[d_n \neq b_1] \wedge [(d_n \neq a_1) \vee (d_{n-1} \neq (b_1))] \wedge [(d_{n+1} = a_n) \vee (d_{n+1} = b_n) \wedge (d_{n+2} = a_n)]$
(c2) $[(d_n = b_1) \vee (d_n = a_1 \wedge d_{n-1} = b_1)] \wedge [(d_{n+1} \neq a_n) \wedge (d_{n+1} \neq b_n) \vee (d_{n+2} \neq a_n)]$
Then if (cl) is verified, we get:

$$E(L_{T_{PTSP}}) = p(1-(1-p)^{n-1})(d+a_n) \\ -p^2 \sum_{r=0}^{n-2} q^r max(a_{n-r}, b_1) \quad (7)$$

On the other hand, if (c2) is verified, we have:

$$E(L_{T_{PTSP}}) = p(1-(1-p)^{n-1})(d-b_1) \\ +p^2 \sum_{r=0}^{n-2} q^r min(a_n, b_{1+r}) \quad (8)$$

Bellalouna [23] was also based on the result of Lawler [31]to prove the stability of TSP. Lawler [31] showed that for triangular inequality matrix the TSP is easy then transport problem.

Property 11. Let C a non negative matrix and verifies the triangular inequality, if 1 et n are presents and the shortest path between 1 and n is (1 , n) then TSP is stable.

Property 12. Let C a non negative matrix and for i < j we have $c_{i,j} \leq C_{k,j}$ $\forall k$ $i+1 \leq k \leq j-1$ then TSP is stable.

In Bellalouna'thesis [23], Christofides heuristic for TSP [31] was studied and it is proved that its approximation ratio is bounded by a constant even for the case of PTSP.

Property 13. If the TSP matrix is positive and verifies the triangular inequality and if X is a random variable representing the number of present vertices and verifying $Pr(W \leq n-k-1) = 0$ and $Pr(W = n-k) > 0$ then Christofides heuristic is an heuristic in the worst case for the PTSP.

$$\frac{E(L_{C_{TSP}})}{E(L_{T_{PTSP}})} \leq \frac{3}{2}[1 + \frac{k^2(k+1)}{n-2}] \quad (9)$$

where $L_{C_{TSP}}$ is the tour provided by Christofides' algorithm.

Basing on research of Bellalouna relating to the small PTSP [23], [26] a study of particular cases which can be solved in an easy way was addressed in [19].

Let $T[i]$ the i^{th} city of the tour $T, i \in \{1, ..., n\}$, $T[i-1]$ is the predecessor of $T[i]$ and $T[i+1]$ is the successor of $T[i]$.

Property 14. Let C be a small matrix with distinct values, suppose that $D_2 = \{1\}, D_0 = \{n\}$ and $D_b = \emptyset$. Without loss of general information,let $a_1 < a_2 < ... < a_{n-1} < a_n$. Then: TSP is stable if and only if $a_{n-1} < b_1$ and $a_n < min\{b_i\}_{2 \leq i \leq n-1}$

Property 15. Let C be a small matrix with distinct values, suppose that $D_2 = \{1\}, D_0 = \{n\}$ and $D_b = \{\}$. Without loss of general information,let $a_1 < a_2 < ... < a_{n-1} < a_n$ and $b_1 < ... < b_{n-1}$ Then:

$$T_{PTSP} = T^*_{TSP}$$

B. Bounds

Bertsimas and Howell [16] improved the best upper and lower bounds for PTSP in three cases: 1) $p_i = p$, 2) $p_i \neq p_j$ and 3) $p_1 = 1, p_i \neq p_j$.

1) The case $p_i = p$. They proved that the bounds proved by Jaillet for n prime (3) holds even if n is not prime.

Result 2. If T_{PTSP} is the optimal PTSP tour,then for n not prime ($n = 2k+1$):

$$E[L_{T_{PTSP}}] \geq pL_{T_{PTSP}}(1 + (1-p)^{2k-1} - (1-p)^k(2-p)) \quad (10)$$

2) The case $p_i \neq p_j$. In the case of unequal probabilities, they obtained lower bounds for the expected length of the PTSP by using a mathematical programming rather than a combinatorial approach. They used an idea suggested by Berman and al. in [33].

Result 3. If T_{PTSP} is the optimal PTSP tour, then

$$E[L_{T_{PTSP}}] \geq z^* \quad (11)$$

where z^* is the optimal solution to the transportation problem.

$$z^* = min \sum_{i,j} x_{i,j} d(i,j) \ ,$$
$$\text{s.t} \sum_i x_{i,j} = p_j(1 - \prod_{k \neq i}(1-p_k)) \ ,$$
$$\sum_j x_{i,j} = p_i(1 - \prod_{k \neq i}(1-p_k)) \ ,$$
$$x_{i,j} \geq 0$$

For the upper bound Bertsimas and Howell [16] used the triangle inequality:

Result 4. Under the triangle inequality,

$$E[L_{T_{PTSP}}] \leq L_{TSP} \quad (12)$$

3) The case $p_1 = 1, p_i \neq p_j$. In this case they improved the bound (12) if the triangle inequality holds:

Result 5. Under the triangle inequality, any tour T satisfies

$$E[L_T] \leq \sum_{i=2}^{n} pi[d(i,1) + d(1,i)] \quad (13)$$

Bellalouna [23] gave also bounds for PTSP based on the work of Jaillet [1], who showed that the inequality (3) in the case where the number of points is prime. Bellalouna gave Similar results for any n.

Result 6 Let T_{PTSP} the optimal PTSP tour, if $n = 2k+1$ then:

$$E[L_{T_{PTSP}}] \geq p^2 L_{T^0_{PTSP}} \frac{1-(1-p)^{n-1}}{1-(1-p)^k} \qquad (14)$$

Result 7. Let T_{PTSP} the optimal PTSP tour and $n = 2k > 6$ then:

$$E[L_{T_{PTSP}}] \geq p^2 L_{T^0_{PTSP}} \frac{1-(1-p)^{n+h_{r_0}}}{1-(1-p)^n} \qquad (15)$$

and $gcd(n, r_0+1) = 1 \Rightarrow \exists(h_n, h_{r_0})/nh_n + (r_0+1)h_{r_0} = 1$.

C. Asymptotic results

In [2], [17] Jaillet presented an interesting analysis of the PTSP in the plane in order to find convergence results for PTSP similar to those demonstrated by Beardwood and al [34] for TSP.R

We note $x = \{x_1, x_2, ...\}$ a sequence of points of R^2 and $x^n = \{x_i, ..., x_n\}$. If the position of the points is random then the sequence is represented by $X = \{X_1, X_2, ...\}$.

Result 1. Let X be a sequence of points uniformly and independently distributed within the unit square $[0,1]^2$ and each point has probability p of being present, independently of the others then there is a constant $c(p)$ as:

$$\lim_{n\to\infty} \frac{E(L_{T_{PTSP}}(x^n, p))}{\sqrt{n}} = c(p) \qquad (16)$$

where

$$\beta\sqrt{p} \leq c(p) \leq min(\beta, 0.9204\sqrt{p})$$

β is the TSP constant in the theorem of Beardwood and al [34], who showed that $0.625 \leq \beta \leq 0.9204$.

VI. Solving methods

There are several algorithms for solving PTSPs. Some papers use exact algorithms to solve PTSPs to optimality, we cite Berman and Simchi-Levi [33] who suggested a lower bound and explained how to combine this bound with a branch-and-bound algorithm to find an optimal a priori tour, Laporte and al. [35] who proposed an exact branch and cut algorithm based on an integer two-stage stochastic programming formulation to solve to optimality instances involving up to 50 vertices. Most approaches in the PTSP literature focus on heuristics that efficiently find good but not necessarily optimal solutions, these include Clarke and Wright, nearest neighbor [36], spacefilling curve, 2-OPT and 1-Shift techniques ([16], [37], [38], [39]). Recently, metaheuristics have been proposed to solve the PTSP such as simulated annealing algorithm [40], scatter search algorithm ([41], [42]), Ant Colony Optimization ([41], [43], [44], [45]) Greedy Randomized Adaptive Search Procedure (GRASP) ([46], [47]), A hybrid Honey Bees Mating Optimization (HBMO) [48], iterative local search algorithms ([49], [50]), memetic algorithms [50].

VII. Conclusion

Probabilistic Combinatorial Optimization problems are very suitable and real-life problems where probabilities are associated with the data. In this paper we have interested on the PTSP the first problem studied in PCOP and we have surveyed the main results obtained on it include their combinatorial properties, bounds and asymptotic results. We have also presented the different solving methods for PTSP. A wide variety of exact and approximate algorithms have been proposed for solving it. Exact algorithms can only solve relatively small problems. As for the approximate methods, a number of heuristics and metaheuristics have proved very satisfactory for large problems. Nowadays, approximate algorithms are the main interests of many researchers who still trying to find the best algorithm which give a very good approximate solution in a proper running time.

References

[1] P. Jaillet, "Probabilistic Traveling Salesman Problems," PhD thesis. MIT, Cambridge,MA, USA, 1985.

[2] P. Jaillet, "Analysis of the probabilistic traveling salesman problem in the plane," Ricerca Operativa, vol. 36, 1986.

[3] D. Bertsimas, "A Vehicule Routing Problem with Stochastic Demand," Operational Research, 1992, pp. 574–585.

[4] D. Bertsimas, "The Probabilistic Minimum Spanning Tree Problem," Networks, vol. 20, 1990, pp. 245–275.

[5] C. Murat and V. Paschos, "The probabilistic maximum independent set problem," Proceedings ASMDA, 1995.

[6] C. Murat and V. Paschos, "The probabilistic maximum independent set problem," Theoretical Computer Science, vol. 270, 2002.

[7] C. Murat and V. Paschos, "The probabilistic longest path problem," Networks, vol. 33, 1999.

[8] C. Murat and V. Paschos, "The probabilistic minimum vertex covering problem," International Transactions on Operational Research, vol. 9, 2002.

[9] C. Murat and V. Paschos, "The probabilistic minimum coloring problem," Proceedings WGÂŠ03, vol. 2880, pp. 346–357, 2003.

[10] N. Bourgeois, F. Della Croce, B. Escoffier, C. Murat and V. Paschos, "Probabilistic graph-coloring in bipartite and split graphs," Journal of Combinatorial Optimization, 2009.

[11] V. Paschos, O. Telelis, V. Zissimopoulos, "Probabilistic models for the Steiner Tree problem," Networks, vol. 56, pp. 39–49, 2010.

[12] M. Bellalouna, S. Souissi and B. Ycart, "Average-Case Analysis for the Probabilistic Bin Packing Problem," Trends in Mathematics, BirkhÃď'user Verlag Basel, Switzerland, pp. 149–159, 2004.

[13] M. Bellalouna and L. Horchani, "The Two-Dimensional Probabilistic Bin-Packing Problem : An average case analysis," International Journal of mathematics and computer in simulation, pp. 42–46, 2008.

[14] Z. Bouyahia, M. Bellalouna, P. Jaillet and K. Ghedira, "A Priori Parallel Machines Scheduling," Computers Industrial Engineering, vol. 58, pp. 488–500, 2010.

[15] D.J. Bertsimas, P. Jaillet, A. Odoni, "A priori optimization,“ Operations Research, vol. 38(6),pp. 1019–1033, 1990.

[16] D.J. Bertsimas, Louis H . Howell "Further results on the probabilistic traveling salesman problem," European Journal of Operational Research, vol. 65, pp. 68–95, 1993.

[17] P. Jaillet, "Analysis of combinatorial optimization problems in the euclidean spaces," Mathematics of Operation Research, vol. 18(1), 1993.

[18] D.J. Bertsimas,"Probabilistic combinatorial optimization problems," PhD thesis, MIT, Cambridge, MA, 1988.

[19] M. Bellalouna,V.Th.Paschos and W.Khaznaji, "Well solved cases of probabilistic traveling salesman problem," 42emes Journees de Statistique, 2010.

[20] P. Jaillet, "On some probabilistic combinatorial optimization problems defined on graphs," Flow Control of Congested Network, vol. 38, pp. 255–267, 1987.

[21] J. Hromkovic, "Stability of approximation algorithms for hard optimization problems," PSOFSEM'99: Theory and Practice of Informatics, vol. 1725, pp. 29–47, 1999.

[22] L. Forlizzi, J. Hromkovic, G.Proietti, and S. Seibert, "On the stability of approximation for hamiltonian path problem," Proc. Theory and Practice of Comp. Sci., vol.3381, pp. 147–156, 2005.

[23] M. Bellalouna, "Problèmes d'optimisation combinatoires probabilistes," PhD thesis, CERMA, Ecole Nationale des Ponts et Chausstes, Paris, 1993.

[24] N. Boria, C. Murat, V. Paschos, "The small traveling salesman problem," Journal of Mathematical Modelling and Algorithms, vol. 11, pp. 45–76, 2012.

[25] S. Rosenow, "Comparison of an Exact Branch-and-Bound and an Approximative Evolutionary Algorithm for the Probabilistic Traveling Salesman Problem," Operations Research Proceedings, pp. 168–174, 1998.

[26] M. Bellalouna, C. Murat, and V.Th. Paschos, "Probabilistic combinatorial optimization problems on graphs : A new domain in operational research," European Journal of Operational Research, vol. 87(3), pp. 693–706, 1995.

[27] M. Gendreau,G. Laporte, R. Séguin, "A Tabu Search Heuristic for the Vehicle Routing Problem with Stochastic Demands and Customers," Operations Research, vol. 44, pp. 469–477, 1996.

[28] P. Jaillet, "A priori solution of a Travelling Salesman Problem in which a random subset of the customers are visited," Operations research, vol. 36, pp. 929–936, 1988.

[29] P. Jaillet, "Rates of convergence of quasi additive smooth Euclidean funcionals and application to combinatorial optimization problems," Mathematics of Operations Resaerch, vol. 17, pp. 965–980, 1992.

[30] X. Berenguer, "A characterisation of linear admissible transformation for the m-traveling salesmen problem," European J.Oper.Res., vol. 3, pp. 232–249, 1979.

[31] Gabovitch, E.Y., "The small traveling salesman problem," Trudy Vychisl. Tsentra Tratu., Gos. Univ, vol. 19, pp. 27–51, 1970.

[32] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B.Shmoys, "The Traveling Salesman Problem," John Wiley and Sons, New york, 1985.

[33] O. Berman and D. Simchi-Levi, "Finding the optimal a priori tour and location of a traveling salesman with nonhomogeneous customers," Transportation Science, vol. 22(2),pp. 148–154, 1988.

[34] J. Beardwood, J. Halton, and J. Hammersley, "The shortest path through many points," Proc. Camb. Phil Soc, vol. 55, pp. 299–327, 1959.

[35] G. Laporte, F. Louveaux, and H. Mercure, "A Priori Optimization of the Probabilistic Traveling Salesman Problem," Operations Research, vol. 42(3),pp. 543–549, 1994.

[36] F. A. Rossi and I. Gavioli, "Aspects of heuristic methods in the probabilistic traveling salesman problem," Advanced School on Statistics in Combinatorial Optimization, pp. 214–227, 1987.

[37] D.J. Bertsimas, P. Chervi, M. Peterson,"Computational approaches to stochastic vehicle routing problems," Transportation Science, vol. 29(4),pp. 342–352, 1995.

[38] L. Bianchi, J. Knowles, N. Bowler, "Local search for the probabilistic traveling salesman problem : Correction to the 2-p-opt and 1- shift algorithms" European Journal of Operational Research, vol. 162(1), pp. 206–219, 2005.

[39] L. Bianchi, A. M. Campbell, "Extension of the 2-p-opt and 1-shift algorithms to the heterogeneous probabilistic traveling salesman problem," European Journal of Operational Research vol. 176, pp. 131–144, 2007.

[40] N. E. Bowler, T. M. A. Fink, and R. C. Ball, "Characterization of the probabilistic traveling salesman problem," Physical Review, E 68, 036703 ,2003.

[41] Y.-H. Liu, "A hybrid scatter search for the probabilistic traveling salesman problem," Comput. Oper. Res, vol. 34, pp. 2949–2963, 2007.

[42] Y.-H. Liu, "Diversified local search strategy under scatter search framework for the probabilistic traveling salesman problem," Eur. J. Oper. Res., vol. 191,pp. 332–346, 2008.

[43] L. Bianchi, L. M. Gambardella, and M. Dorigo, "An ant colony optimization approach to the probabilistic traveling salesman problem," 7th International Conference on Parallel Problem Solving from NatureI, vol. 2439, pp 883–892,2002.

[44] L. Bianchi, L. M. Gambardella, and M. Dorigo, "Solving the homogeneous probabilistic traveling salesman problem by the ACO metaheuristic," 3rd International Workshop on Ant Algorithms, vol. 2463, pp 176–187,2002.

[45] L. Bianchi, "Ant colony optimization and local search for the probabilistic traveling salesman problem : a case study in stochastic combinatorial optimization," Ph.D. Thesis, Univ. Libre de Bruxelles, 2006.

[46] Y. Marinakis, M. Marinaki, "A hybrid multi-swarm particle optimization algorithm for the probabilistic traveling salesman problem," Comput. Oper. Res., vol. 37, pp. 432–442, 2010.

[47] Y. Marinakis, A. Migdalas, P.M. Pardalos, "Expanding neighborhood search - GRASP for the probabilistic traveling salesman problem," Optim. Lett, vol. 2, pp. 351–361, 2008.

[48] Y. Marinakis, M. Marinaki, "A hybrid honey bees mating optimization algorithm for the probabilistic traveling salesman problem," BIBLIOGRAPHIE 24 Proc.IEEE Congr. Evol. Comput., CEC '09, pp. 1762–1769, 2010.

[49] P. Balaprakash, M. Birattari, T. StÃijtzle, M. Dorigo, "Estimation-based local search for stochastic combinatorial optimization using delta evaluations : a case study on the probabilistic traveling salesman problem," INFORMS J. Comput, vol. 20, pp. 644–658, 2008.

[50] P. Balaprakash, M. Birattari, T. Stützle, M. Dorigo, "Adaptive sample size and importance sampling in estimation-based local search for the probabilistic traveling salesman problem," European J. Oper.Res., vol. 199, pp.98–110, 2009.

DOI: 10.15439/2014F483
ACSIS, Vol. 3

Transdimensional sequential Monte Carlo for hidden Markov models using variational Bayes - SMCVB

Clare A. McGrory
Centre for Applications in Natural Resource Mathematics
University of Queensland
Brisbane
Queensland, Australia
Email: c.mcgrory@uq.edu.au

Daniel C. Ahfock
Centre for Applications in Natural Resource Mathematics
University of Queensland
Brisbane
Queensland, Australia
Email: daniel.ahfock@uqconnect.edu.au

***Abstract*—In this paper we outline a transdimensional sequential Monte Carlo algorithm - SMCVB - for fitting hidden Markov models. Sequential Monte Carlo (SMC) involves generating a weighted sample of particles from a sequence of probability distributions with the aim of converging to the target Bayesian posterior distribution. SMCVB makes use of variational Bayes (VB) in combination with SMC principles to create an algorithm which targets the posterior distribution more efficiently thereby saving on time and computational storage requirements. Another key feature of our methodology is that the variational-Bayes-generated proposals can vary in dimension. We have found in our simulation studies that we are able to obtain sensible estimates of the model dimensionality in this one-step procedure. This introduces very valuable additional flexibility in the modelling approach and opens up the potential for use of the algorithm in on-line settings where efficient and reliable estimation of dimensionality and parameters is required.**

I. Introduction

SEQUENTIAL Monte Carlo (SMC) approaches for Bayesian inference were first introduced to meet the requirement for efficient and tractable methods for analysing large amounts of data that arose sequentially over time (see [1] for an overview). In SMC, the procedure begins by initially proposing a population of samples, which are referred to as particles, from an initial target posterior distribution, reweighting these particles through importance sampling and then resampling from them to approximate the next target posterior density in the sequence. Subsequently there has also been a significant amount of research into the application of SMC to static problems, i.e. the data are treated as if they had arisen sequentially even though the whole dataset is available at the start of the analysis. For example, [2] and [3], provide examples of SMC in a static setting. Another example of a static SMC algorithm is given in [4]. This is a data-tempering SMC approach and this will form the basis for our new proposed algorithm.

Within the context of finite mixture estimation, [5] proposed a new transdimensional SMC algorithm based on the idea of using the variational Bayes (VB) approach [6], [7], [8] within an SMC framework. The resulting hybrid algorithm is called SMCVB. The SMC algorithm is initialised with particles drawn from a VB approximation to the posterior distribution rather than from a prior distribution; the aim of this is to make the algorithm more efficient. The underlying SMC algorithm takes the form of the data-tempering algorithm described in [4] as noted above. A significant advantage of the SMCVB algorithm is that it is not restricted to fixed-dimensional space. This is a highly useful feature for practical application since estimating a suitable dimension for the model is usually an important part of the analysis. In particular, in applications where new batches of data arise over time, the dimension size that achieves the most appropriate fit might change throughout the analysis as new information becomes available. Our approach has the advantage over existing schemes that it is able to adapt to such changes in an automated fashion. This feature means that there is a lot of potential for application and extension of the hybrid approach to modern applications where datasets are ever increasingly large and there is a demand for fast or even online analysis capabilities.

In this paper we describe how the algorithm proposed in [5] can be extended to the context of hidden Markov modelling, and we show that this leads to a novel scheme which is time-efficient and provides reliable results.

The article is organised as follows. In Section 2 we outline the model. In section 3, we describe the VB approach for hidden Markov modelling with Gaussian noise. In Section 4 we present the transdimensional VB-based SMC (SMCVB) algorithm. In Section 5 we show some results from the analysis of simulated data, and Section 6 concludes the paper.

II. Variational Bayesian Inference for Hidden Markov Models with Gaussian Noise

Following the approach that is described in [9], we assume a Gaussian HMM where the system can be in any one of K states at any time-point i, but the actual state sequence is hidden. Our observations correspond to a noisy realisation of the actual state sequence. We assume a discrete first-order Markovian dependence structure, therefore the current state depends only on the state occupied at the last time-point. We will follow the notation set out in [8] for specifying the HMM and we will apply the algorithm described in that article for estimation of the model. Given that the system is in state j_1 at time-point i, the transition matrix π represents the probability of moving to state j_2 at time-point $i+1$.

The transition matrix is defined as $\pi = \{\pi_{j_1j_2}\}$ where $\pi_{j_1j_2} = p(z_{i+1} = j_2|z_i = j_1)$ and z_i is the latent variable representing the state at time i; all transition probabilities are non-negative and columns of the transition matrix must sum to 1. No structure is imposed on the transition matrix π, it will be estimated as part of the analysis. The observed data is denoted by $\{y_i; i = 1, \dots, n)\}$, and the emission probabilities, i.e., the conditional probabilities of state membership at each time-point, are denoted by $p(y_i|z_i = j) = p_j(y_i|\phi_j)$. Since we are assuming Gaussian noise in the observations, the $\phi = \{\phi_j\}$ correspond to the parameters of the univariate Gaussian noise distribution corresponding to the relevant states $j = 1, \cdots, K$. Then, the model parameters are given by $\theta = (\pi, \phi)$ and we have

$$\begin{aligned} p(y, z, \theta) &= \prod_{i=1}^{n}\prod_{j=1}^{K}(p_j(y_i|\phi_j))^{z_{ij}} \\ &\times \prod_{i=1}^{n-1}\prod_{j_1=1}^{K}\prod_{j_2=1}^{K}(\pi_{j_1j_2})^{z_{ij_1}z_{i+1j_2}} \\ &\times \prod_{j=1}^{K}p_j(\phi_j)\prod_{j_1=1}^{K}p(\pi_{j_1}), \end{aligned}$$

where z_{ij} is a latent indicator variable such that $z_{ij} = 1$, if $z_i = j$, and $z_{ij} = 0$, if $z_i \neq j$. The terms $p_j(\phi_j)$ and $p(\pi_{j_1})$ correspond to the prior distributions over the parameters of the univariate Gaussian noise distribution, and the transition probabilities, for the relevant state j_1.

We use the same prior specifications for this model as the ones used in [8]. For more detail, we refer the reader to that paper. We follow the notation used in [8] in order to facilitate comparison with the more detailed descriptions of the corresponding derivations provided therein. The standard conjugate prior distributions are used for the model parameters. For each state j_1, there is an independent Dirichlet prior distribution for the transition probabilities $\{\pi_{j_1j_2} : j_2 = 1, \dots, K\}$, with hyperparameters $\{\alpha_{j_1j_2}{}^{(0)}\}$. The noise model for the observations is univariate Gaussian with unknown means and precisions such that for each state j, we have a Gaussian prior distribution with mean μ_j and precision τ_j. Each of the means μ_j themselves have independent univariate Gaussian conjugate prior distributions, conditional on the precisions, with means and precisions given by $m_j^{(0)}$ and $\beta_j{}^{(0)}\tau_j$, respectively. The precisions τ_j have independent Gamma prior distributions with shape and scale parameters given by $\frac{1}{2}\eta_j{}^{(0)}$ and $\frac{1}{2}\delta_j{}^{(0)}$, respectively.

III. Variational Bayesian Inference for Hidden Markov Models with Gaussian Noise

In the variational Bayesian inferential approach, we do not sample from the posterior distribution, as we would in a Markov chain Monte Carlo (MCMC) based approach, instead we find a close approximation to it; this approximation to the posterior is referred to as the variational posterior distribution. The fact that the VB estimate of the posterior does not require iterative sampling makes it a very useful approach in terms of time efficiency, which is of course an important consideration when working with large datasets. We will briefly outline the key concepts of the variational approach in this section. As we have stated, our aim is to find the VB approximation to our desired posterior distribution, i.e. $p(\theta|y)$. This posterior can be obtained as the marginal distribution of $p(\theta, z|y)$; this distribution is typically a complex expression and it has to be approximated in this method. In the VB approach, we approximate $p(\theta, z|y)$ by another distribution which we call the variational approximating distribution. The variational approximating distribution is denoted by $q(\theta, z)$, and the idea is to take this distribution as being the minimiser of the the Kullback-Leibler(KL) divergence between $q(\theta, z)$ and $p(\theta, z|y)$. To make the minimisation of the KL divergence between these quantities tractable, the standard assumption made is that $q(\theta, z)$ can be factorised as $q(\theta, z) = q_\theta(\theta)q_z(z)$. Derivation of the variational function leads to a set of coupled equations for $q_\theta(\theta)$ and $q_z(z)$ for updating the estimates of the parameters and latent variables in the mode. Note that another way to view the motivation for this approach, is that the variational approximation to the posterior provides a tight lower bound on the observed-data log-likelihood. The variational Bayesian algorithm proceeds by iteratively updating these coupled expressions for the model parameters and the latent variables until they converge, at least locally, in the sense that subsequent updates no longer improve estimates. The values in the converged algorithm are then the estimates of the model parameters in the variational posterior distribution.

The Forms of the Variational Posterior Estimated Distributions over Model Parameters and Latent Variables

After applying the standard VB approximation to the Bayesian posterior of the HMM we find the following forms for the variational posterior distributions over the model parameters [8].

$$\begin{aligned} q_{j_1}(\pi_{j_1}) &= \mathrm{Dir}(\pi_{j_1}|\{\alpha_{j_1j_2}\}), \\ q(\mu_j|\tau_j) &= \mathrm{N}(\mu_j|m_j, (\beta_j\tau_j)^{-1}), \\ q(\tau_j) &= \mathrm{Ga}\left(\tau_j|\frac{1}{2}\eta_j, \frac{1}{2}\delta_j\right). \end{aligned}$$

The parameters of these distributions can be computed by iteratively solving the set of coupled equations outlined in Algorithm 1. The well-known forward-backward algorithm [10] has to be used to make computation of the required marginal probabilities possible [11]. The forward-backward algorithm gives us estimates of the forward and backward variables, $\mathrm{fvar}_i(j)$ and $\mathrm{bvar}_i(j)$, respectively, for each i and j. In the forward-backward algorithm the $a^*_{j_1j_2}$ are estimates of the probabilities of transition from states j_1 to state j_2, and the b^*_{ij}'s are estimates of the emission probabilities given that the system is in state j at time point i. These are then used in the update equation for $q_{ij} = q_z(z_i = j) = p(z_i = j_1|y_1, \dots, y_n)$

Algorithm 1 VB Algorithm for Fitting a Hidden Markov Model with Gaussian Noise

```
Set initial values for parameters
$\alpha_{j_1j_2}{}^{(0)}, m_j^{(0)}, \beta_j^{(0)}, \eta_j^{(0)}, \delta_j^{(0)}, K, q_z(z_i = j_1, z_{i+1} = j_2)$
and $q_{ij}$, for $j, j_1, j_2 \in 1, \cdots, K$, $i \in 1, \cdots, n$
while not converged do
    Update the VB posterior parameter estimates ($\Psi$ denotes the digamma function)
    $\alpha_{j_1j_2} = \alpha_{j_1j_2}{}^{(0)} + \sum_{i=1}^{n-1} q_z(z_i = j_1, z_{i+1} = j_2)$
    $\beta_j = \beta_j{}^{(0)} + \sum_{i=1}^{n} q_{ij}$
    $\eta_j = \eta^{(0)} + \sum_{i=1}^{n} q_{ij}$
    $\delta_j = \delta^{(0)} + \sum_{i=1}^{n} q_{ij}y_i{}^2 + \beta_j{}^{(0)}m_j{}^{(0)^2} - \beta_j m_j^2$
    $m_j = \frac{\beta_j{}^{(0)}m_j{}^{(0)} + \sum_{i=1}^{n} q_{ij}y_i}{\beta_j}$
    $a^*_{j_1j_2} = \exp\left(\Psi(\alpha_{j_1j_2}) - \Psi\Big(\sum_{j=1}^{K}\alpha_{j_1j}\Big)\right) = p(z_{i+1} = j_2 | z_i = j_1)$
    $b^*_{ij} = \exp\left(\frac{1}{2}\Psi\Big(\frac{1}{2}\eta_j\Big) - \frac{1}{2}\log\Big(\frac{\delta_j}{2}\Big) - \frac{1}{2\beta_j} - \frac{1}{2}\Big(\frac{\eta_j}{\delta_j}\Big)(y_i - m_j)^2\right) = p(y_{i+1} | z_{i+1} = j_2)$
    $q_z(z_i = j) = \frac{\mathrm{fvar}_i(j_1)\mathrm{bvar}_i(j_1)}{\sum_{j_2}\mathrm{fvar}_i(j_2)\mathrm{bvar}_i(j_2)}$
    $q_{ij} = q_z(z_i = j_1, z_{i+1} = j_2) = \frac{\mathrm{fvar}_i(j_1)a^*_{j_1j_2}b^*_{i+1j_2}\mathrm{bvar}_{i+1}(j_2)}{\sum_{j_1}\sum_{j_2}\mathrm{fvar}_i(j_1)a^*_{j_1j_2}b^*_{i+1j_2}\mathrm{bvar}_{i+1}(j_2)}$
    If any state has a weighting approaching zero then eliminate this state and reduce model dimension to $K = K-1$
end while
```

and $q_z(z_i = j_1, z_{i+1} = j_2)$. The $\sum_{i=1}^{n} q_{ij}$ is the VB estimate of the number of observations expected to belong to state j, and the $q_z(z_i = j_1, z_{i+1} = j_2)$ are the VB estimates of the transition probabilities.

We would like to draw the readers attention to the automatic state elimination feature that is an intrinsic part of the VB approach when estimating mixture models and HMMs. As a result of this property, given the initially chosen value for model dimension, K, the final estimated solution in the VB posterior will have dimension less than or equal to K. Provided that K is chosen sufficiently large, we expect that a suitable dimension for the model is estimated as part of the VB procedure. Note that model selection criteria could also be computed to provide an alternative way to select the most appropriate dimension size.

IV. Transdimensional Variational Bayes Sequential Monte Carlo Algorithm (SMCVB)

The SMC framework which underpins the algorithm is a modification of the SMC algorithm described in [4]. The algorithm described in [4] is initialised with a small batch of data and then proceeds to incorporate data in sequential batches of increasing size which is what is meant by data tempering. What distinguishes our SMCVB approach from other SMC algorithms is that we use a VB posterior mean estimate of the model parameters in order to generate proposal particles rather than generating them from the prior. This is an intuitively logical and sensible hybrid modification of the data-tempering SMC algorithm. The complete-data target posterior distribution that we ultimately wish to estimate is

$$\pi(\theta) = \pi(\theta | y_1, \cdots, y_n),$$

and the target posterior at each subsequent iteration t ($t = 1, \cdots, T$) is

$$\pi_t(\theta) = \pi_t(\theta | y_1, \cdots, y_{n_t}),$$

where $n_1 \leq n_2 \leq \cdots \leq n_T = n$ is an increasing set of sample sizes. This separation of the whole dataset into smaller sub-batches leads to the formation of a sequence of target posteriors which on average smoothly converge to the final complete data target posterior. Our proposed SMCVB algorithm for HMMs is outlined in Algorithm 2.

Due to the VB algorithm intrinsic state elimination property, VB solutions obtained for each batch, which are in turn used to generate new proposed sets of particles, can vary in dimension. This allows us to explore a range of models with various numbers of states at various points in the analysis. We suggest that the distribution of particles over the various dimension sizes might be used as a guide for deciding on the most appropriate number of states to include in the final model. In our analyses of simulated datasets we found that this strategy led to reliable estimates of the most appropriate number of states to include in the fitted model.

V. Some Results from Application of the SMCVB Algorithm for HMMs to Simulated Data

For illustration we present here results obtained from using our hybrid algorithm to analyse synthetic data generated from a three-state hidden Markov model. The parameter settings used to simulate the data are outlined in table I, note that this corresponds to a considerably noisy set of data. We generated 1000 datapoints from this model and the data were read in in batches of 100 points. We used vague priors in the analysis. The transition matrix was given by

Algorithm 2 SMCVB Algorithm

```
**Initialise**: estimate the VB partial posterior $\pi_{t_0}(\theta) = \pi_{VB}(\theta|y_1, \cdots, y_{n_0})$ using Algorithm 1
**Particle set**: generate a set of $R$ particles $(\theta_r^{(0)}, W_r^{(0)})_{r=1,\cdots,R}$ with associated weights $\{W_r^{(0)}\}$ which target the initial posterior $\pi_{t_0}(\theta)$.
**Draw**: draw $R$ particles from these estimated posteriors, which results in vectors of the form $\{\theta_R^{(0)} = (\mu_r^{(0)}, \tau_r^{(0)}, \rho_r^{(0)})\}$, with weights given by
    $W_r^{(0)} \propto \frac{p(y_1, \cdots, y_{n_0}|\theta_r^{(0)})p(\theta_r^{(0)})}{\pi_{VB}(\theta_r^{(0)}|y_1, \cdots, y_{n_0})}$
Normalise the weights to obtain $W_r^{(0)}$.
**while** $n_t < n$ **do**
    Reweight: update the weights at iteration $t$ using the $n_t$th batch of data to give
        $W_r^{(t)} \propto W_r^{(t-1)} \times p(y_{n_{t-1}+1}, \cdots, y_{n_t}|\theta_r^{(t-1)})$,
    where $r = 1, \cdots, R$.
    **if** Effective sample size < $\frac{n}{2}$ **then**
        Resample $R$ values from the current set of particles using multinomial sampling. i.e. we resample the $\{(\theta_r^{(t-1)}, W_r^{(t-1)})\}_{r=1,\cdots,R}$ to get $\{(\theta_r^{\prime(t)}, 1/R)_{r=1,\cdots,R}\}$.
    **end if**
    **Move**: move to a new set of particles, these become the $\{\theta_r^{(t)}\}$ to be carried forward. Propose these from distributions from the VB posterior mean of the parameters based on the current batch of data and use a standard Metropolis–Hastings update to choose the new particles.
**end while**
```

$$\pi = \begin{pmatrix} 0.15 & 0.80 & 0.05 \\ 0.50 & 0.10 & 0.40 \\ 0.30 & 0.40 & 0.30 \end{pmatrix}$$

The estimated posterior parameter values for the noise model are displayed in table II and table III. We compare the fits we obtained with our hybrid SMCVB algorithm to those obtained from a standard MCMC analysis and a standard VB analysis. Plots of the posterior distributions corresponding to the fitted parameters are shown in Figures 1 and 2.

Using our approach, we estimated the most suitable number of states by calculating the proportions of particles corresponding to the different model sizes. The majority of particles in the final set corresponded to a three-state model, which we know accurately reflects the true underlying model in this case. Further work is required to better explore the reliability and justification for the general use of the distribution of the number of states in the final particles for estimating the most appropriate number of states for the fitted model. However, our explorations of some different simulated datasets suggest good potential for this strategy. Model selection criteria could also be checked to assess the most appropriate dimension.

TABLE I
PARAMETERS OF THE GAUSSIAN NOISE DISTRIBUTIONS FOR EACH STATE IN THE MODEL USED TO SIMULATE THE EXAMPLE DATA.

State	Mean	Standard Deviation
1	1.00	0.50
2	2.00	0.15
3	2.50	0.30

TABLE II
POSTERIOR MEANS OF THE MEAN PARAMETERS OF THE GAUSSIAN NOISE DISTRIBUTIONS FOR EACH STATE ESTIMATED USING THE DIFFERENT APPROACHES

SMCVB	VB	MCMC
1.01	1.01	1.01
2.00	2.00	2.00
2.56	2.56	2.55

TABLE III
POSTERIOR MEANS OF THE STANDARD DEVIATION PARAMETERS OF THE GAUSSIAN NOISE DISTRIBUTIONS FOR EACH OF THE STATES ESTIMATED USING THE DIFFERENT APPROACHES

SMCVB	VB	MCMC
0.53	0.53	0.53
0.15	0.15	0.15
0.26	0.26	0.27

The results shown demonstrate that this approach leads to reliable estimates of model parameters. This new hybrid SMCVB scheme produces posterior estimates which are even closer to MCMC estimates for the same model than the VB approximation is. Note then that another way to view this scheme is as a way to further improve on the VB approximation to the Bayesian posterior distribution.

SMCVB is efficient in terms of computing time and due to the nature of the SMC structure, it is ideally suited to applications where new batches of data continually become available. That feature, combined with the improved time efficiency that is achieved through the use of the targeted VB guided proposals, has created an algorithm which has much potential to be of practical use in modern applications where large volumes of sequentially occurring data have to be processed and the traditional MCMC-based approaches may not be feasible due to computational limitations.

VI. CONCLUSION

We have extended the recently proposed transdimensional SMC algorithm, SMCVB, to the setting of estimating parameters and dimension of hidden Markov models. The algorithm allows us to explore the dimension of the posterior distribution and achieves increased computational efficiency over other SMC approaches by using a VB algorithm to generate the independent proposals at each iteration of the procedure.

Current work involves applying this algorithm to analysing large time series data with the aim of performing climate regime shift detection.

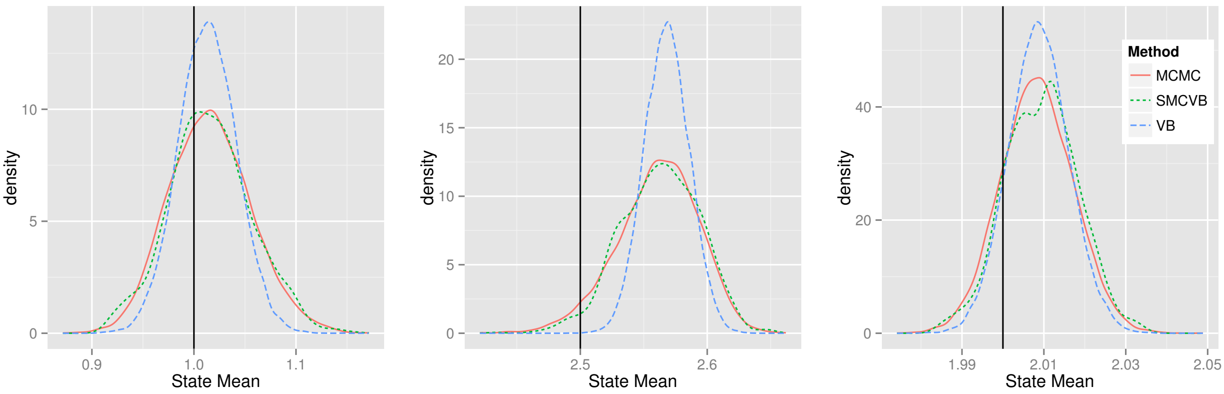


Fig. 1. Comparison of the posterior distributions for the mean parameter of the Gaussian noise distribution for the three states fitted using the different methods. The solid black line marks the true mean for the corresonding state in the model the data were simulated from.

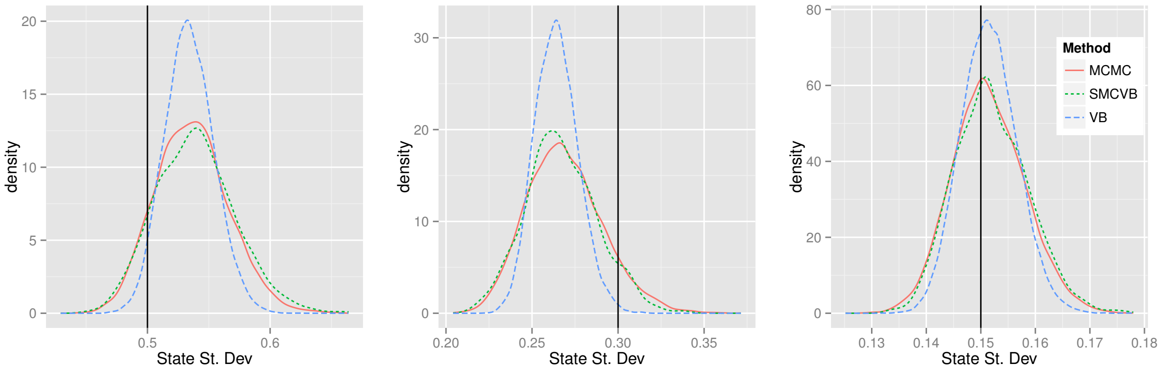


Fig. 2. Comparison of the posterior distributions for the standard deviation parameter of the Gaussian noise distribution for the three states fitted using the different methods. The solid black line marks the true standard deviation for the corresponding state in the model the data were simulated from.

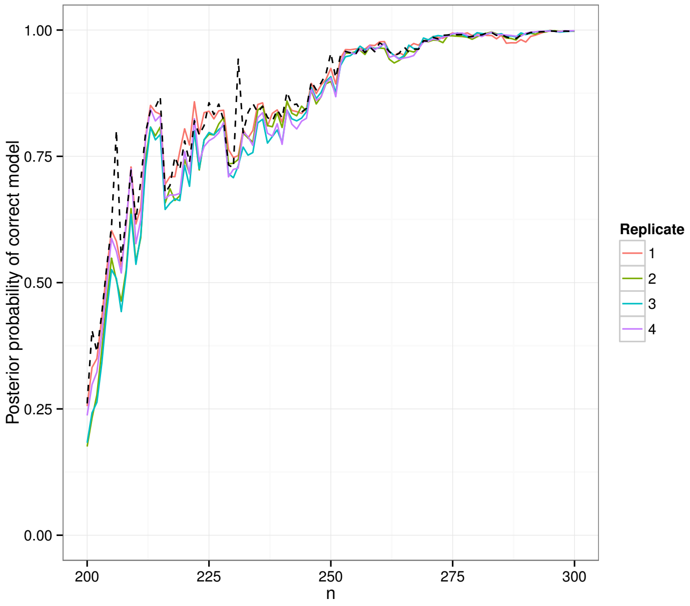


Fig. 3. Posterior probability associated with the correct number of hidden states for the model as estimated based on the proportion of particles having that dimension in the final set of particles. The plot shows results from four replicate runs of the SMCVB algorithm and for different numbers of particles.

References

[1] A. Doucet, J.F.G. De Freitas, and N.J. Gordon, *Sequential Monte Carlo Methods in Practice.* NewYork, NY: Springer, 2001.

[2] P. Del Moral, A. Doucet, and A. Jasra, “Sequential Monte Carlo samplers,” *J. Roy. Statist. Ser. B,* vol. 68, 2006, pp. 411–436.

[3] A. Jasra, D.A. Stephens, and C.C. Holmes, “On population based simulation for static inference,” *Statist. Comput.* vol. 17, 2007, pp. 263–279.

[4] N. Chopin, “A sequential particle filter method for static models,” *Biometrika* vol. 89, 2002, pp. 539–551.

[5] C. A. McGrory, A. N. Pettitt, D.M. Titterington, C.L. Alston, and M. Kelly, “Transdimensional Sequential Monte Carlo using Variational Bayes - SMCVB,” *Unpublished,* 2014.

[6] C. A. McGrory, and D. M. Titterington, “Variational approximations in Bayesian model selection for finite mixture distributions,” *Comput. Stat. Data An.* vol. 51, 2007, pp. 5352–5367.

[7] H. Attias, “Inferring parameters and structure of latent variable models by variational Bayes,” *in Proceedings of the 15th International Conference on Uncertainty in Artificial Intelligence,* Morgan Kaufmann, San Francisco, 1999, pp. 21–30.

[8] M. Wand, J. Ormerod, S. Padoan, and R. Fruhwirth, “Mean Field Variational Bayes for Elaborate Distributions,” *Bayesian Analysis* vol. 6, 2011, pp. 847–900.

[9] C. A. McGrory, and D. M. Titterington, “Variational Bayesian analysis for hidden Markov models,” *Aust. and New Zealand J. Statist.* vol. 51, 2009, pp. 227–244.

[10] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” *Ann. of Math. Statist.* vol. 41, 1970, pp. 164–171.

[11] T. Rydén, T. Teräsvirta, T., and S. Åsbrink, "Stylized facts of daily return series and the hidden Markov model," *J. Appl. Econometr.* vol. 13, 1998, pp. 217–244.

Computer Science & Systems

CSS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- CANA'14 - 7th Workshop on Computer Aspects of Numerical Algorithms
- MMAP'14 - 7th International Symposium on Multimedia Applications and Processing
- SCoDiS-LaSCoG'14 - 3nd Workshop on Scalable Computing in Distributed Systems and 8th Workshop on Large Scale Computations on Grids

7th Computer Aspects of Numerical Algorithms

NUMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

Topics

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on GPUs
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

Event Chairs

Bylina, Beata, Maria Curie-Sklodowska University, Poland

Bylina, Jarosław, Maria Curie-Sklodowska University, Poland

Stpiczyński, Przemysław, Maria Curie-Sklodowska University, Poland

Program Committee

Amodio, Pierluigi, Università di Bari, Italy

Anastassi, Zacharias, Qatar University, Qatar

Banaś, Krzysztof, AGH University of Science and Technology, Poland

Brugnano, Luigi, Universita' di Firenze, Italy

Czachorski, Tadeusz, IITiS

Filippone, Salvatore, University Rome Tor Vergata, Italy

Fourneau, Jean-Michel

Gansterer, Wilfried, University of Vienna, Austria

Georgiev, Krassimir, IICT - BAS, Bulgaria

Gimenez, Domingo, University of Murcia, Spain

Gravvanis, George, Democritus University of Thrace, Greece

Kozielski, Stanislaw

Kucaba-Pietal, Anna, Politechnika Rzeszowska, Poland

Lirkov, Ivan, Institute of Information and Communication Technologies , Bulgarian Academy of Sciences, Bulgaria

Maksimov, Vyacheslav, Institute of Mathematics and Mechanics, Russia

Marowka, Ami, Bar-Ilan University, Israel

Meini, Beatrice, Universita di Pisa, Italy

Minev, Peter, University of Alberta, Canada

Mycka, Jerzy, UMCS

Pekergin, Nihal

Petcu, Dana, West University of Timisoara, Romania

Pultarova, Ivana, Czech Technical University in Prague, Czech Republic

Satco, Bianca-Renata, Stefan cel Mare University of Suceava, Romania

Sedukhin, Stanislav, The University of Aizu, Japan

Sergeichuk, Vladimir, Institute of Mathematics of NAS of Ukraine, Ukraine

Srinivasan, Natesan, Indian Institute of Technology, India

Szajowski, Krzysztof, Institute of Mathematics and Computer Science, Poland

Telek, Miklos

Trivedi, Kishor S., Duke University, United States

Tudruj, Marek, Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland

Tůma, Miroslav, Academy of Sciences of the Czech Republic, Czech Republic

Ustimenko, Vasyl, Marie Curie-Sklodowska University, Poland

Vazhenin, Alexander, University of Aizu, Japan

DOI: 10.15439/2014F83
ACSIS, Vol. 3

Attractor Selection Mechanism Simulink Model

Dmitrijs Finaskins
Riga Technical University
st. Azenes 16, LV-1048, Riga, Latvia
Email: dmitrijs.finaskins@rtu.lv

Gunars Lauks
Riga Technical University
st. Azenes 16, LV-1048, Riga, Latvia
Email: gunars.lauks@rtu.lv

***Abstract*—Telecommunication network topologies are changing nowadays. The core telecommunication networks are based on fiber optics infrastructure. WDM technology is used to transmit multiple flows of traffic over a single fiber. Virtual network topology (VNT) is used to route IP traffic over WDM networks. VNT control system must be adaptive to traffic changes and dynamically reconfigure VNT. Langevin equation is used in statistical physics to describe the time evolution of a subset of the degrees of freedom. These degrees of freedom typically are collective (macroscopic) variables changing only slowly in comparison to the other (microscopic) variables of the system. The fast variables are responsible for the stochastic nature of the Langevin equation. Langevin equation could be used to model attractor selection mechanism, which is used by biological systems to adopt to unknown changes in environment. In this paper we describe attractor selection mechanism Simulink model and analyze simulation results.**

I. Introduction

CORE telecommunication networks nowadays are based on Wavelength Division Multiplexing (WDM) technology. In such division multiplexing technology transmitting/receiving channels are divided by wave length. It allows transmitting multiple traffic channels over a single fiber. Traffic by fiber optics can be transmitted over long distances without any additional equipment. Because of that WDM networks are commonly used to carry Internet traffic at backbone level. The major protocol of Internet is IP protocol. One of possibilities to carry IP traffic over WDM network is to construct virtual network topology, which includes transmitting/receiving channels (lightpaths) and IP routers. There are many VNT control methods, which configure/reconfigure VNT according to traffic demand matrices. Traffic demand matrices show how traffic flows are distributed via lightpaths.

One of the most exciting Internet opportunities for end-users is to share their pictures, videos and so on with other users; to communicate with each other using online services such as Skype. These cause constant and rapid changes in traffic flows between IP routers in VNT. There is a need to reconfigure VNT over a period of time in order to provide high-level service with minimal delays. VNT needs to be adaptable to changes in traffic demand.

There are two modes of constructing traffic demand matrices—offline and online or dynamical mode. In offline approach traffic demand matrices are constructed using previous known information about changes in traffic demand in [1], [2], [6]. The major weakness of this approach—offline methods will not work correctly if traffic flow changes will be different from ones expected. Online approach allows reconfiguring VNT dynamically. In this case periodical measurement results are used. To evaluate VNT status, information about average or maximum link utilization, packet delays can be used. Based on this information new lightpaths are added to source-destination pair of nodes if, for example, link utilization between these nodes is more than threshold and deleted if lightpath is underutilized.

The majority of online VNT control methods are used when traffic demand is changing periodically and gradually [5]. This approach will not work if changes of traffic demand are not predictable. There is a need to develop such VNT control method, which is adaptable to unknown changes in network environment.

One of the approaches is to use attractor selection, which represents mechanism of adaptation to unknown changes of biological systems [4]. The main idea of attractor selection—the system is driven by two components—deterministic and stochastic. Attractors are a part of the equilibrium points in the solution space. Conditions of such system are controlled by very simple feedback. When conditions of a system are suitable (close to one of the attractors), it is driven almost only by deterministic behaviour, stochastic influence is very limited. When conditions of the systems are poor, deterministic behaviour influence is close to zero and in this case system is driven by stochastic behaviour. It randomly fluctuates searching for a new attractor. When this attractor is found, deterministic behaviour again dominates over stochastic.

There are some publications, which show how to use attractor selection mechanism to manage network resources (bandwidth), but in most cases authors do not write which tool was used to create a model. In this paper we propose a Simulink model, which describes approach proposed by [4]. The model is relatively large and it could be reduced by assigning multiple tasks to one block, but in this case the model will not be so easy to understand.

II. Attractor selection mechanism

Every pair of nodes, between which connection can be established, is represented by control unit u_{ij}, where i and j are indexes of nodes. Every control unit has its control value $x_{u_{ij}}$. Indexes s and d refers to source and destination nodes. How control values change over time can be expressed by

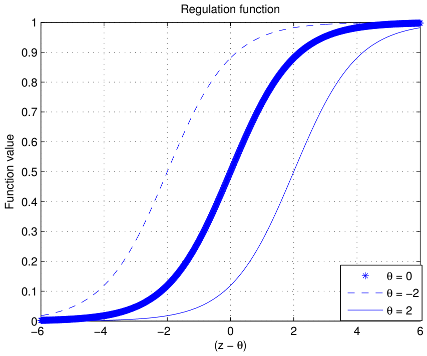


Fig. 1. Regulation function

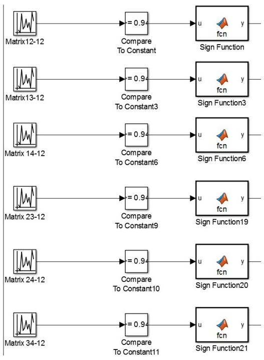


Fig. 2. Regulation matrix generation

differential equation, also known as Langevin equation:

$$\frac{dx_{u_{ij}}}{dt} = v_g \cdot f\{\sum_{u_{sd}} W(u_{ij}, u_{sd}) \cdot x_{u_{ij}} - \theta_{u_{ij}}\} - v_g \cdot x_{u_{ij}} + \eta, \quad (1)$$

where v_g indicates conditions of VNT, $W(u_{ij}, u_{sd})$—regulatory matrix of VNT, $\theta_{u_{ij}}$ is coefficient, which depends on minimum and maximum loads on links in VNT and η—Gaussian noise. We use Gaussian noise with 0 mean value and 0.1 variance. The first term in Equation (1) represents regulation function f(z):

$$f(z) = \frac{1}{1 + e^{-z}}; z = \sum_{u_{sd}} W(u_{ij}, u_{sd}) \cdot x_{u_{ij}} - \theta_{u_{ij}} \quad (2)$$

Regulatory matrix $W(u_{ij}, u_{sd})$ is the most important parameter of Equation (1), since it shows the relationships between node pairs. Every element of this matrix can be -1, 0, or 1; it represents the influence of node pair u_{ij} on node pair $u_s d$. According to model, proposed in [3], -1 corresponds to inhibition of the node u_{ij} by u_{sd}, 0 corresponds to no relation and 1 to activation. These values are multiplied by corresponding control values of u_{ij}. In such way node pairs affect each other—if u_{sd} is activated by u_{ij}, increasing in $x_{u_{ij}}$ will increase $x_{u_s d}$ and, as it was mentioned above, increase the amount of lightpaths between nodes s and d. In [3] has been shown that if node pairs can activate or inhibit each other with probability 0.03 and no relation with probability 0.94, such network will be extremely adaptable to changes in traffic demand.

So for every node pair we construct regulation function, which is defined by Equation (2). Adjusting $\theta_{u_{ij}}$, we can move regulation function in the negative or positive direction, as it is shown in Figure 1.

The amount of assigned lightpaths is a function of $x_{u_{ij}}$ normalized by all control unit values of all node pairs, which use transmitters and receivers from node i to node j. The number of lightpaths $G_{u_{ij}}$ between nodes i and j is defined as follows [4]. The total number of lightpaths between nodes i and j is K_r:

$$G_{u_{ij}} = min\{K_r \cdot \frac{x_{u_{ij}}}{\sum_s x_{u_{sj}}}\} \quad (3)$$

III. Model limitations, assumptions, out of scope elements

In our Simulink model, described in this paper, we simulate 4-node network. We chose 4 nodes because creating wider network does not add additional value to the model, but only makes it more difficult. But it is not a model limitation, using our model a network with unlimited number of nodes could be simulated.

In our model we assume that connections between all node pairs have equal parameters (bandwidth, delay, etc.). This was done to simplify the model, but, in general, it is not a mandatory requirement and links with not equal parameters could be also used. As [4] we used only link utilization for network activity control. But more complex metrics could be used as well, e.g. link utilization and total delay.

Loads on links are randomly generated at certain time moments. Initial loads on links are randomly generated as well. We use uniform distribution function for this purpose. We chose uniform distribution to make generation results unpredictable and therefore to simulate such network in with traffic flow can change rapidly and unpredictably. We also assume that between two loads on links generation moments total traffic does not change.

IV. Attractor selection mechanism Simulink model

A. Regulatory matrix generation

As it was mentioned above, each node pair in the network affects all other nodes, including itself. We simulate 4-node network, it means that we have 6 node pairs—$x_{12}, x_{13}, x_{14}, x_{23}, x_{24}, x_{34}$. Regulatory matrix's $W(u_{ij}, u_{sd})$ dimensions will be 6x6. In Figure 2 we show how the first row of this matrix is generated. We use "Uniform random number" block, "Compare to constant" block and "User defined function" block for generating 1 element of matrix with the given parameters (0.94 probability of no relation, 0.03 of inhibition and 0.03 for activation). First of all we generate a uniformly distributed random number $\epsilon[0; 1]$. Next, we compare this generated number with constant—0.94. In 94 cases "Compare to constant" block will output 0, because in 94 cases out of

Algorithm 1 Sign function block

```
function y = fcn(u)
if u==0;
    u=0;
else m=sign(0.5 −rand);
             if m>=0
                  u=1;
              else u=−1;
             end
end
y = u;
```

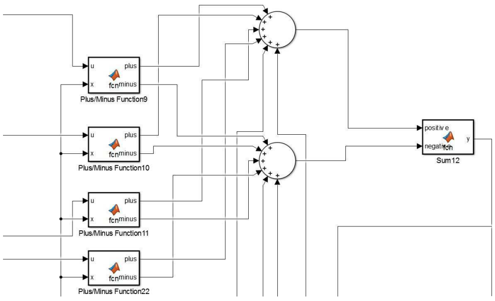


Fig. 3. Calculation of total inhibition/activation value

100 generated number will be less or equal to 0.94. In 6 cases out of 100, this block will output 1. This signal (0 or 1) is an input variable for "Sign function". In this block it is decided -1, 0, or 1 will be put in the regulatory matrix. The code of this block is listed in Algorithm 1. If "Sign function" block receives 0 as an input signal, it does nothing, but if 1, it is assumed that -1 or 1 decision should be made. Both inhibition and activation probabilities are equal, so result of rand function is subtracted from 0.5. By this we ensure that in 50 cases out of 100 we have -1 and in other 50 cases 1. In Figure 2 it could be seen how the signal is processed further. We need to calculate the total inhibition and activation value—this is done by multiplying -1 or 1 with $x_{u_{ij}}$. We divide this flow into two parts—positive and negative, corresponding to activation and inhibition. The total value of $W(u_{ij}, u_{sd})$ for $x_{u_{ij}}$ is calculated in block "Sum 12". The code for this block could be found in Algorithm 2. This value is used for regulatory function calculation.

B. Load on links generation

As it was mentioned above, we use uniform distribution function to generate load on links. We randomly change load on links at time moments t $\epsilon\epsilon$ [0,2,5,8]. At all other time moments we use load on links calculated by attractor selection mechanism. Please refer to Algorithm 3 to see how it happens. We use variable t to control the current simulation time. ry_{12} variable is used for random load on link generation, y_{12} corresponds to load on link calculated by algorithm and yo_{12} is load on link, which is processed to the next blocks—"Min" and "Max" blocks. We multiply yo_{12} by 100 to reflect the

Algorithm 2 Calculation of the total value of $W(u_{ij}, u_{sd})$ for $x_{u_{ij}}$

```
function y = fcn(positive,negative)
m=0;
n=0;
if positive == 0
    m=0;
else m = (1.08*6)/positive

end

if negative == 0
    n=0;
else n = (1.08*6)/negative
end
y = m + n;
```

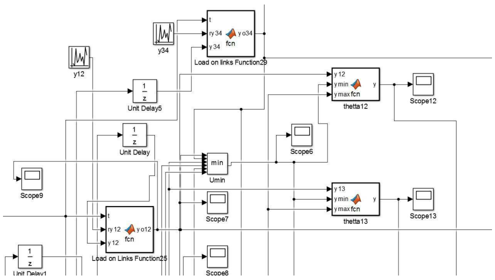


Fig. 4. Load on links generation

Algorithm 3 Load on links management

```
function yo12 = fcn(t,ry12,y12)
if t==0
    yo12=100*ry12;

elseif t==2
    yo12=100*ry12;

elseif t==5
    yo12=100*ry12;

elseif t==8
    yo12=100*ry12;

else
    yo12=y12;

end
end
```

Algorithm 4 $\theta_{u_{ij}}$ value calculation

```
function y = fcn(y12,ymin,ymax)
tc=5;
if ymin==ymax
     m=0;
      y=m;
else
m= -((y12-ymin)/(ymax-ymin))*2*tc-tc
y = m;
end
```

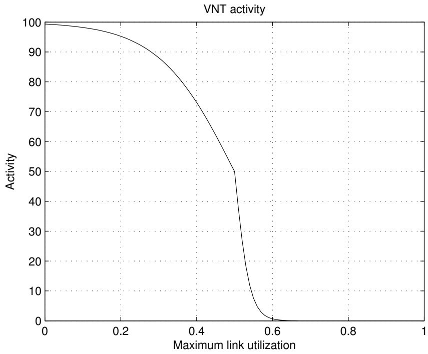


Fig. 5. VNT Activity

fact that 100 Gbps links are used, but this coefficient is not mandatory and could be excluded.

We need to know minimal and maximal load on link in our network to calculate regulation function. "Min" and "Max" blocks allow to do this very easily. These values are transferred to block "Theta", which calculates $\theta_{u_{ij}}$ value. Please refer to Algorithm 4 for code details. Condition $y_{min} = y_{max}$ is mandatory for catching exceptions and is highly recommended to be used.

C. Network activity and regulation function

We have discussed already how initial control values x_{sd} and load on links y_{sd} are generated, to integrate Equation (1) we need to calculate activity function. Activity function is defined [4] as follows:

$$v_g = \frac{100}{1 + exp\delta \cdot (u_{max} - \xi)}, u_{max} \geq \xi \tag{4}$$

$$v_g = \frac{100}{1 + exp\frac{\delta}{5} \cdot (u_{max} - \xi)}, u_{max} < \xi \tag{5}$$

Figure 5 shows how activity function looks like. From Equations (4) and (5) it can be seen that ξ is a threshold for VNT activity v_g. If maximum link utilization is more than ξ, activity of VNT dramatically degraded and stochastic behaviour dominates over deterministic and the system is searching for a new attractor. Algorithm 5 shows Simulink code for activity calculation and Figure 6 shows network activity v_g for 4-node network. For regulation function calculation we use Equation (2). All needed parameters/variables have been already calculated in the previous steps.

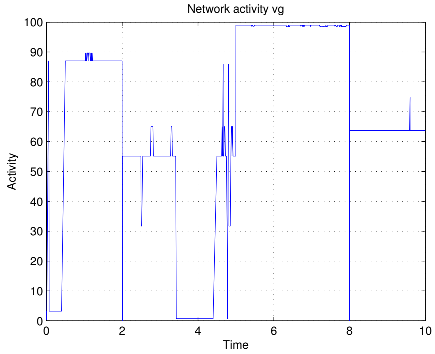


Fig. 6. Activity function

Algorithm 5 v_g calculation

```
function y = fcn(umax)
igrek = 100;
beta = 50;
ksi = 0.5;
if umax>=ksi
vg=(igrek/(1+exp(beta*(umax-ksi))));
else
vg=(igrek/((1+exp((beta/5)*(umax-ksi)))));
end
y = vg;
```

For Gaussian noise generation random number generation block is used. As it was mentioned above, mean value of this signal is 0 and variance 0.1. It was shown in [3], using Gaussian noise with these parameters makes attractor selection mechanism extremely adaptable for unknown and unpredictable changes in environment.

D. Numerical integration of system of differential equations

Equation (1) has 3 terms—regulation function multiplied by activity function, control value $x_{u_{ij}}$ multiplied by activity function and Gaussian noise. As it can be seen in Figure 7, we use sum block with 3 inputs to sum these three terms. Regulation and activity functions calculations are represented

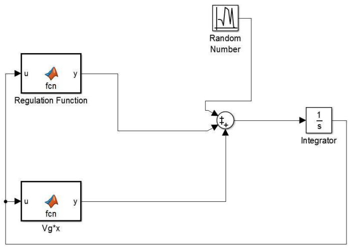


Fig. 7. Integration concept

as "User defined function" blocks. Both blocks use during the previous timeslot integrated $x_{u_{ij}}$ value. We have to solve such differential equation for each node pair in the network. As it was mentioned above, each two nodes in the network form a node pair, regardless where these nodes are located and with no respect to the fact if they are directly connected. For example, in our simulated network we have 4 nodes. It means that we have 6 nodes pairs. So at each time moment we need to solve a system of 6 differential equations. In general, if we have N nodes in the network, we will have $\frac{N \cdot (N-1)}{2}$ differential equations.

E. Lightpaths assignment and load on links recalculation

As soon as we have integrated Equation (1), that is, have $x_{u_{ij}}$ values, we need to assign/reassign lightpaths to node pairs according to Equation (3). We use the following model for load on links recalculation—at time moment $t = 0$ we randomly generate load on links. When we have $x_{u_{ij}}$, we assign additional lightpaths for node pairs (At time moment $t = 0$ each node pair have only 1 lightpath assigned). When additional lightpaths are assigned, we recalculate load on links by dividing current load on link with assigned number of lightpaths. For example, if at time moment $t = 0$ we have $y_{12} = 80$ Gbps and there are 3 additional lightpaths assigned to this node pair, then the new load on this link would be $\frac{80}{4} = 20$ Gbps.

After one integration period (sample period) passed and we got new $x_{u_{ij}}$ values, we need to recalculate load on links. First of all, we calculate the total load on link for y_{ij} by multiplying current load on link distributed by all lightpaths with the number of lightpaths. For example, if we take y_{12} mentioned in the previous paragraph, we multiply 20 Gbps with number of lightpaths (4) and get initial value 80 Gbps. Then we recalculate number of lightpaths which will be assigned to this node pair according to Equation (3) and divide load on link with this new amount of lightpaths. For example, at step 2, 5 lightpaths were assigned to y_{12}. It means that new load on link for this node pair will be $\frac{80}{5} = 16$ Gbps.

To make this algorithm work, we need to somehow save the previous value of lightpaths assigned to node pair y_{ij} at time moment T_{t-1} as well as current number of lightpaths for time moment T_t. This could be easily done in Simulink by using one of two blocks—"Memory block" or "Unit delay block". Both these blocks delay an input signal for one integration period (sample time), we used "Unit delay block" in our simulation. You can see a fragment of our Simulink model which recalculate load on links in Figure 8 and Algorithm 6 shows source code. If we refer to Figure 8, t represents simulation time, y_{ij} current load on links, G_{ij} new amount of lightpaths assigned to y_{ij}, dG_{ij}—number of lightpaths for y_{ij} from the previous integration step and ey_{ij} is a new load on link. As it can be seen from Algorithm 6, at time moments $t \epsilon [0, 2, 5, 8]$ load on links are randomly generated and in that cases we do not use previous number of lightpaths to restore the original load on link.

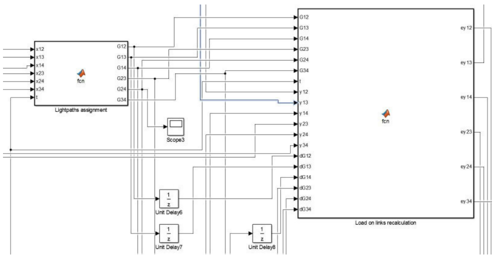


Fig. 8. Load on links recalculation

Algorithm 6 Load on links recalculation

```
if t==0
    ey12=y12/G12;
ey13=y13/G13;
ey14=y14/G14;
ey23=y23/G23;
ey24=y24/G24;
ey34=y34/G34;
elseif t==2
     ey12=y12/G12;
ey13=y13/G13;
ey14=y14/G14;
ey23=y23/G23;
ey24=y24/G24;
ey34=y34/G34;
elseif t==5
     ey12=y12/G12;
ey13=y13/G13;
ey14=y14/G14;
ey23=y23/G23;
ey24=y24/G24;
ey34=y34/G34;
elseif t==8
     ey12=y12/G12;
ey13=y13/G13;
ey14=y14/G14;
ey23=y23/G23;
ey24=y24/G24;
ey34=y34/G34;
else
ey12=(y12*dG12)/G12;
ey13=(y13*dG13)/G13;
ey14=(y14*dG14)/G14;
ey23=(y23*dG23)/G23;
ey24=(y24*dG24)/G24;
ey34=(y34*dG34)/G34;
end
end
```

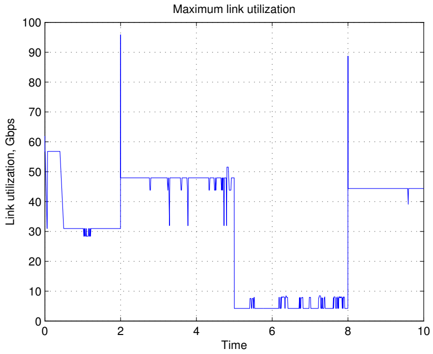


Fig. 9. Maximum link utilization in the network

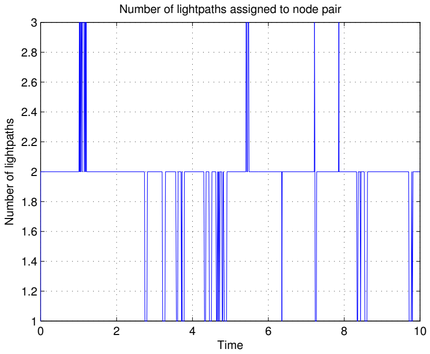


Fig. 10. Number of lightpaths assigned to node pair y_{ij}

V. Simulation results

In this section we present results of simulating 4-node network using proposed attractor selection mechanism Simulink model. Please refer to Figure 9 to see how maximum link utilization in the network changes over time. The maximum link capacity is 100 Gbps, at time moments $t\epsilon[0, 2, 5, 8]$ traffic was randomly generated. As it can be seen from Figure 9, attractor selection mechanism works well and when load on links changes, network adopts to this changes and maximum link utilization decreases.

The one could see an interesting effect at time moment $t\epsilon[2; 5]$ and $t\epsilon[5; 8]$. Although traffic amount does not change over time during these time intervals (according to our simulation traffic changes are done only at time moments $t\epsilon[0, 2, 5, 8]$, at all other time moments traffic is constant), we can see fluctuations in maximum link utilization. It is a side effect of using Gaussian noise in Equation (1). This effect was described in [4]. The system conditions are good, but the maximum link utilization increases unexpectedly (see Figure 9), although there are no changes in traffic demand at this time moment. It happens because stochastic behaviour always influences system conditions, even if VNT conditions are good. It causes VNT activity degradation. At this moment noise drives the system and a new attractor is being searched. After some time a new attractor has been found and system activity recovers. Because of noise random nature it cannot be predicted when it will happen.

But on the other hand the same effect helps to reduce maximum link utilization at time moments $t\epsilon[0, 2, 5, 8]$. At these time moments, due to changes in traffic demand and as a result changes in maximum link utilization, VNT activity degrades. In these moments stochastic behaviour dominates over deterministic as the first two terms in Equation (1) are zero or close to zero. Control unit values randomly fluctuate searching for a new attractor. As soon as a new attractor is found, VNT activity recovers and deterministic behaviour again drives VNT.

Figure 10 shows how amount of lightpaths assigned to node pair y_{ij} changes over time. As it can be seen from this figure, there are some fluctuations in number of lightpaths assigned to y_{ij}. There are two main reasons for that. The first one, as it was mentioned above, is stochastic fluctuations which affect network activity and network activity affects all other network parameters. The second reason is that we do not use any mechanisms for such fluctuation filtering. If filtering is used, we could prevent number of lightpaths from flapping.

Overall, simulation results show that our proposed attractor selection mechanism Simulink model works well. This proves that Simulink provides great opportunity for researches and scientists to use this program to get practical results for their theories/algorithms.

VI. Discussion and further work

Uniformly distributed function for traffic demand matrix generation has been used. This distribution has been selected as it allows to study how control method will adapt to unknown changes if traffic demand will vary from 0 to 1 (with respect to link capacity) with equal probability. On the other hand, it is vital to know how attractor selection mechanism deals with traffic in case when traffic demand matrix is generated using normal distribution (Gaussian distribution). Some other distribution functions such as logarithmic and Poison are also needed to be studied—if attractor selection Simulink model is able to successfully deal with such distributed traffic.

Another improvement is to use more complicated network for simulation. In this case, a model should be modified, as it does not support transit lightpaths from some nodes to others, because in real networks each node is not directly connected any other. Because of that we need to modify block which assigns lightpaths to support some lightpaths reservation for transit purposes.

We also want to make metrics for network activity measurements more complex. Now the decision about network conditions is made based only on maximum link utilization. Some others metrics could be added, for example, total delay between x_s and x_d.

References

[1] G. Agrawal and D. Medhi, "Lightpath topology configuration for wavelength-routed IP/MPLS networks for time-dependent traffic," Proceedings of GLOBECOM, 2006, pp. 1-5.

[2] B. Chen, G. N. Rouskas, and R. Dutta, "On hierarchical traffic grooming in WDM networks," IEEE/ACM Transactions on Networking," vol. 16, 2008, pp. 1226-1238.

[3] C. Furasawa and K. Kaneko, "A generic mechanism for adaptive growth rate regulation," PLoS Computational Biology, vol. 4, 2008, p. e3.

[4] Y. Koizumi, T. Miyamuta, S. Arakawa, E. Oki, K. Shiomoto, M. Murata, "Application of attractor selection to adaptive virtual network topology control," Proceedings of the 3rd International Conference on Bio-Inspired Models of Network, Information and Computing Systems, 2008, Article No. 9.

[5] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E.D. Kolaczyk and N. Taft, "Structural analysis of network traffic flows," Proceedings of ACM Sigmetrics, 2004, pp. 61-72.

[6] F.Ricciato, S. Salsano, A. Belmonte and M. Listanti, "Off-line configuration of a MPLS over WDM network under time-varying offered traffic," Proceedings of IEEE INFOCOM, vol. 1, 2002, pp. 57-65.

DOI: 10.15439/2014F281
ACSIS, Vol. 3

Parallel implementation of linear repetitive processes identification using subspace algorithms

Dominik Kujawa
Institute of Control and Computation Engineering University of Zielona Góra
ul. Podgórna 50, 65-246 Zielona Góra, Poland
Email: dkujawa@weit.uz.zgora.pl, dominikinf@gmail.com

***Abstract*—This paper presents a new parallel approach to identification of linear repetitive processes based on subspace algorithms. Parallel realizations of these algorithms are tested on various graphic cards that use NVIDIA CUDA technology. The paper describes implementation of subspace identification algorithms and their parallel speedup, efficiency, throughput, and delay. The parallel approach to the identification of linear repetitive processes based on subspace methods, presented in this paper, is very useful not only for time invariant LRPs but also for processes with dynamics that changes rapidly from pass to pass. A simulation example is included to illustrate the effectiveness of the proposed approach.**

***Index Terms*—state space models, subspace methods, identification algorithms, parameter estimation.state space models, subspace methods, identification algorithms, parameter estimation.s**

I. Introduction

FOR SEVERAL years, the growth of maximum clock frequency of digital implementations has stopped, particularly in general purpose processors. This phenomenon is presented in [1]. It seems that further progress in computer system performance is still possible despite reaching the limit of switching frequency of digital integrated circuits. The further increase in computing power can be achieved by multiplying the number of computational units in multi-core general purpose Digital Signal Processors (DSP), Grapihics Processing Units (GPU), and reconfiguration Field Programmable Gate Array (FPGA) chips [2].

Subspace identification methods are an attractive alternative to the well-known prediction error methods due to simple and general parameterizations in the MIMO case. They do not need any canonical parameterizations as well. Moreover, no nonlinear optimization is performed and reliable state-space models for complex multi-input multi-output dynamical systems are derived directly from the input-output data. Also, computational complexity of subspace identification methods is modest in comparison with the well-known prediction error methods [3], [4], [5], [10]. Subspace identification algorithms consist of two steps. In the first step, based on the input-output data, the repetitive process order is determined and the repetitive process sequence of states (N4SID algorithm) or the extended observability matrix of system (MOESP algorithm) are computed [7]. In the second step, the unknown system matrices are determined.

The repetitive control theory has been an area of intense research since the beginning of 1990's and considerable results have been achieved both in the analysis and synthesis problems, (see: [6]) for the actual state of art. Contrary to the LRP control theory, the identification of LRPs has attracted very limited attention. The aim of this paper is to propose a new approach to the identification of the LRPs based on subspace algorithms. In this approach, the order of a LRP and the unknown process matrices are determined based on the input and output sequences of the actual pass and the output sequence of the previous pass using parallel implementations of N4SID and MOESP identification algorithms.

Parallelization of algorithm for repetitive processes subspace identification requires creating and processing Hankel matrices. The number of processors that is simultaneously busy depends on the size of matrices storing information defining the spatial variable, and the time variable determining the position on the pass, while the length of each pass is finite [6].

This paper considers the number of processors used, speedup, efficiency, throughput and computing time of parallel implementations of subspace identification algorithms and compares them to the modified sequential versions of MOESP N4SID algorithms. They were both quantitative and qualitative indicators describing the identification repetitive processes. These algorithms are examined using the input-output data generated by the repetitive process simulator. The paper is organized as follows. Section 2 presents an introduction to the identification of discrete deterministic repetitive processes. Parallel problem identification was formulated, and its solution based on subspace algorithms is given in Sections 3 and 4. Section 5 presents the simulation results. The conclusions are given in Section 6.

II. Identification algorithms

A. Discrete linear repetitive process model

Consider the state-space model of a discrete linear repetitive process of the following form:

$$x_{k+1}(p+1) = Ax_{k+1}(p) + B_0 y_k(p) + Bu_{k+1}(p) \tag{1}$$

$$y_{k+1}(p) = Cx_{k+1}(p) + D_0 y_k(p) + Du_{k+1}(p) \tag{2}$$

where:
$0 \le p \le \alpha - 1 \in Z_+$ is the independent spatial or temporal variable,
$k \in Z_+$ is the current pass number,
$x_k(p) \in R^n$ is the state vector,
$y_k(p) \in R^l$ is the pass profile (output) vector,
$u_k(p) \in R^m$ is the input vector,
A, B, B_0, C, D, D_0 are matrices of appropriate dimensions.

To complete process description, it is necessary to specify the boundary conditions:
$x_{k+1}(0) = d_{k+1}$
$y_0(p) = f(p)$
where $d_{k+1} \in R^n$ is a vector with known constant entries and $f(p) \in R^l$
are known functions of p.

Define the following input Hankel block matrix

$$U_{0|2i-1} = \left[\begin{array}{ccc} u_{k+1}(0) & \dots & u_{k+1}(j-1) \\ y_k(0) & \dots & y_k(j-1) \\ \vdots & \vdots & \vdots \\ u_{k+1}(i-1) & \dots & u_{k+1}(i+j-2) \\ y_k(i-1) & \dots & y_k(i+j-2) \\ \hline u_{k+1}(i) & \dots & u_{k+1}(i+j-1) \\ y_k(i) & \dots & y_k(i+j-1) \\ u_{k+1}(i+1) & \dots & u_{k+1}(i+j) \\ y_k(i+1) & \dots & y_k(i+j) \\ \vdots & \vdots & \vdots \\ u_{k+1}(2i-1) & \dots & u_{k+1}(2i+j-1) \\ y_k(2i-1) & \dots & y_k(2i+j-2) \end{array}\right] = \left[\begin{array}{c} U_p \\ \hline U_f \end{array}\right]$$

$$U_{0|2i-1} = \left[\begin{array}{ccc} u_{k+1}(0) & \dots & u_{k+1}(j-1) \\ y_k(0) & \dots & y_k(j-1) \\ \vdots & \vdots & \vdots \\ u_{k+1}(i-1) & \dots & u_{k+1}(i+j-2) \\ y_k(i-1) & \dots & y_k(i+j-2) \\ u_{k+1}(i) & \dots & u_{k+1}(i+j-1) \\ y_k(i) & \dots & y_k(i+j-1) \\ \hline u_{k+1}(i+1) & \dots & u_{k+1}(i+j) \\ y_k(i+1) & \dots & y_k(i+j) \\ \vdots & \vdots & \vdots \\ u_{k+1}(2i-1) & \dots & u_{k+1}(2i+j-1) \\ hj6y_k(2i-1) & \dots & y_k(2i+j-2) \end{array}\right] = \left[\begin{array}{c} U_p^+ \\ \hline U_f^- \end{array}\right]$$

Define also the output block matrix $Y_{0|2i-1}$

$$y_{0|2i-1} = \left[\begin{array}{ccc} y_{k+1}(0) & \dots & y_{k+1}(j-1) \\ \vdots & \vdots & \vdots \\ y_{k+1}(i-1) & \dots & y_{k+1}(i+j-2) \\ \hline y_k(i) & \dots & y_{k+1}(i+j-1) \\ y_{k+1}(i+1) & \dots & y_{k+1}(i+j) \\ \vdots & \vdots & \vdots \\ y_{k+1}(2i-1) & \dots & y_{k+1}(2i+j-1) \end{array}\right] = \left[\begin{array}{c} Y_p \\ \hline Y_f \end{array}\right]$$

$$y_{0|2i-1} = \left[\begin{array}{ccc} y_{k+1}(0) & \dots & y_{(k+1)}(j-1) \\ \vdots & \vdots & \vdots \\ y_{k+1}(i-1) & \dots & y_{k+1}(i+j-2) \\ y_{k+1}(i) & \dots & y_{k+1}(i+j-1) \\ \hline y_{k+1}(i+1) & \dots & y_{k+1}(i+j) \\ \vdots & \vdots & \vdots \\ y_{k+1}(2i+1) & \dots & y_{k+1}(2i+j-1) \end{array}\right] = \left[\begin{array}{c} Y_p^+ \\ \hline Y_f^- \end{array}\right]$$

The number of block rows i should be lager than the maximum order of the LRP. Define block Hankel matrices W_p and W_p^+ consisting of U_p and Y_p and Y_p^+ ,respectively:

$$W_p = \left[\begin{array}{c} U_p \\ Y_p \end{array}\right] \tag{3}$$

$$W_p^+ = \left[\begin{array}{c} U_p^+ \\ Y_p^+ \end{array}\right] \tag{4}$$

The state-sequence matrix X_i is defined as

$$X_i = [x_{k+1}(i) \dots x_{k+1}(i+j-1)] \tag{5}$$

Define the extended observability matrix Γ_i and the reversed extended controllability matrix Δ_i:

$$\Gamma_i = \left[\begin{array}{c} C \\ CA \\ CA^2 \\ \dots \\ CA^{i-1} \end{array}\right] \tag{6}$$

$$\Delta_i = [A^{i-1}[BB_0] \dots A[BB_0]\ [BB_0]] \tag{7}$$

Assume also that the pair $\{A, C\}$ is observable and the pair $\{A, [BB_0]\}$ is controllable. Define the lower block triangular Toeplitz matrix H_i

$$H_i = \left[\begin{array}{cccc} [DD_0] & 0 & \dots & 0 \\ C[BB_0] & [DD_0] & \dots & 0 \\ CA[BB_0] & C[BB_0] & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ CA^{i-2}[BB_0] & CA^{i-3}[BB_0] & \dots & [DD_0] \end{array}\right] \tag{8}$$

The above block Hankel matrices along with the extended observability matrix, the reversed extended controllability matrix and the lower block triangular Toeplitz matrix $(3)-(8)$ play an important role in the development of subspace identification methods. Following Theorem 1 (see: [8]) , the state-space model $(1) - (2)$ can be reformulated in a matrix form:

$$Y_p = \Gamma_i X_p + H_i U_p \tag{9}$$

$$Y_f = \Gamma_i X_f + H_i U_f \tag{10}$$

$$X_f = A_i X_p + \Delta_i U_p \tag{11}$$

where $X_f = X_i$ and $X_p = X_0$.

B. Identification Problem

Given α measurements of the input $u_{k+1}(p)$ and the outputs $y_k(p)$ and $y_{k+1}(p)$ measurements generated by the LRP $(1) - (2)$ determine its order and the matrices A, B, B_0, C, D and D_0 up to within a similarity transformation.

Assuming that the augmented input $[u_{k+1}(p)\ y_k(p)]$ is persistently exciting of order $2i$ and the intersection of the row space of U_f and the row space of X_p is empty, the unknown LRP matrices A, B, B_0, C, D and D_0 can be computed based on the results of Theorem 2 (see: [8]) . It can be done in two different ways using N4SID or MOESP. In both these algorithms, the order of the process (1)-(2) can be find based on inspection of the singular value decomposition of the matrix $W_1 \theta_i W_2$

$$W_1 \theta_i W_2 = [U_1 U_2] = \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \tag{12}$$

where $W_1 \in R^{lixli}$ and $W_2 \in R^{jxj}$ are the user defined weighting matrices and θi is the oblique projection

$$\theta_i = Y_f /_{U_f} W_p \tag{13}$$

The order of LRP is equal to the number of non-zero singular values in S_1. The extended observability matrix Γ_i is computed from the the following equation

$$\Gamma_i = W_1^{-1} U_1 S_1^{1/2} T \tag{14}$$

where $T \in R^{n \times n}$ is an arbitrary non-singular similarity transformation matrix. N4SID calculates the state sequence X_i from the equation

$$X_i = \Gamma_i^{\dagger} \theta_i \tag{15}$$

where $\Gamma_i^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of the matrix Γ_i. Finally, the matrices A, B, B_0, C, D and D_0 are determined solving the following set of equations

$$\begin{bmatrix} A & [BB_0] \\ C & [DD_0] \end{bmatrix} \begin{bmatrix} X_i \\ U_{i|i} \end{bmatrix} = \begin{bmatrix} X_{i+1} \\ Y_{i|i} \end{bmatrix} \tag{16}$$

The state sequence X_{i+1} is calculated from the equation

$$X_{i+1} = \Gamma_{i-1}^{\dagger} \theta_{i-1} \tag{17}$$

where Γ_{i-1} is the extended observability matrix Γ_i without the last l rows and θ_{i-1} is the oblique projection

$$\theta_{i-1} = Y_f^- /_{U_f^-} W_p^+ \tag{18}$$

In MOESP, the unknown LRP matrices are determined in two steps. In the first step, A and C are calculated from the extended observability matrix Γ_i. In the other step, B, B_0, D, and D_0 are calculated solving a set of equations.

III. Parallel implementation of modified N4SID and MOESP algorithms

The modified N4SID and MOESP algorithms can be partitioned into the following three modules:

- Input module
- Matrix calculation modul
- Model testing modul

The data input module performs pre-processing operations in spatial context in fine-grained pipelined architecture. The detailed time analysis including technological factors can be found in [9]. The delay in the operations of fine-grained data (in cycles) can be expressed as

$$F_t = \frac{\alpha - 1}{2} + r \tag{19}$$

where r is the delay of additional variables in the passes. The transport delay can be determined from (19), on basis of the pass length. The component r corresponds to the delay resulting from the use of additional variables, synchronizing data. They are used to distribute evenly the propagation time between the elements in the matrices. The other modules are parallelized in a coarse-grained architecture. The transport delay for operation with coarse-grained data (in cycles) is

$$L_t = \alpha + r \tag{20}$$

Modules of subspace identification algorithms can been partitioned into the following classes performing parallel operations:

- Class Hankel
- Class oblique projection
- Class SVD
- Class observability
- Class controllability
- Class ABCD

The UML activity diagram (Fig. 1) presents parallelization of block Hankel matrices for subspace identification of repetitive processes.

IV. Parallelization level indicators

System architecture can be adapted to a particular computational task in itself is not a determinant of performance or efficiency.

$$y_{k+1}(p) = F^l(u_k(p))||F^{l-1}(u_k(p))||...||F^l(u_k)(p)) \tag{21}$$

$||-$ executing operations simultaneity
$F^{\sigma}-$ operator, $\sigma \in (1, .., l)$.
Many operations can be performed simultaneously at the inputs (21).

$$S_m = \frac{T_1}{T_m} \tag{22}$$

where:
S_m- the speedup ,
T_1- the execution time of the sequential algorithm,
T_m- the execution time of the parallel algorithm with m processors

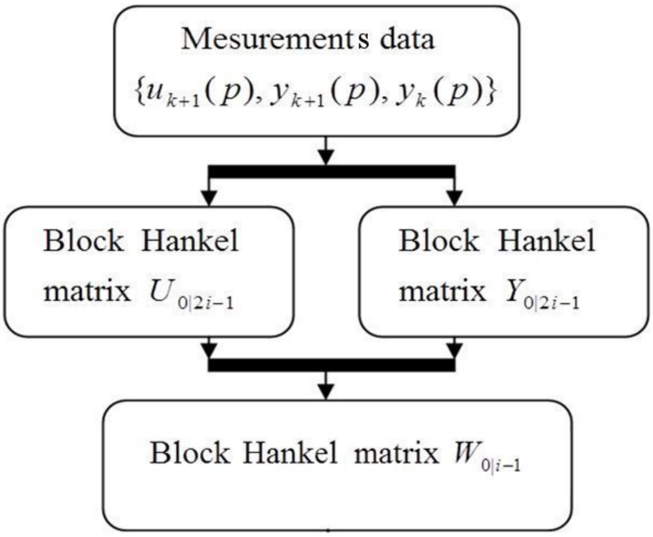

Mesurements data
$\{u_{k+1}(p), y_{k+1}(p), y_k(p)\}$
Block Hankel matrix $U_{0|2i-1}$
Block Hankel matrix $Y_{0|2i-1}$
Block Hankel matrix $W_{0|i-1}$

Fig. 1. UML activity diagram parallelized block Hankel matrices

For the evaluation of computational tasks performance in multi-core systems speedup formula was used, (22).

In multiprocessor systems, the actual value of the speedup is less than the theoretical (based on the number of processors) due to communication overheads and the need to share resources such as memory and buses. This property describes a model of efficiency, which is a measure of concurrent use of resources

$$E_m = \frac{S_m}{m} \tag{23}$$

where:
E_m- the speedup efficiency with m processors $E_m \in (0;1)$.

In the ideal case, the efficiency (23) is 1.0 which means that the speedup is proportional to the number of processor or computing elements.

In the assessment of computing activity is often used as a criterion of time needed to process a specific task or quantum computing. The concept of throughput was introduced, which corresponds to the number of data processed per time unit.

$$P = \frac{D_n}{t} \tag{24}$$

where:
$P-$ throughput,
D_n- the number of computed data,
$t-$ measurement time.

Relative to the computing system has the ability to compute the limit, usually in the long term. Throughput (24) relative to the computing system has the ability to compute the limit, usually in the long term.

V. Simulation example

Consider a simple example to illustrate the proposed approach. The following LRP of the fourth order is identified:

$$A = \begin{bmatrix} 1.516 & -0.755 & 0.125 & -0.001 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$
$$C = \begin{bmatrix} 0.048 & 0.072 & -0.006 & -0.001 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} B_0 = \begin{bmatrix} -0.571 \\ -0.142 \\ -0.453 \\ -0.258 \end{bmatrix} D = \begin{bmatrix} 0 \end{bmatrix} D_0 = \begin{bmatrix} 0.2 \end{bmatrix}$$

with zero boundary conditions $y_0(p) = 0, p = 1, ..., 4\text{x}10^5$, and the initial conditions $x_{k+1}(0), k = 1, ..., 20$, defined as a uniformly distributed pseudorandom sequence on the interval [0,2.5]. The LRP is exited with a uniformly distributed pseudorandom sequence on the interval [0,1] and its output disturbed by a pseudorandom sequence of normal distribution with mean 0 and standard deviation 0.0001. To identify selected data from two consecutive passes and N4SID and MOESP parallelized algorithms were used. Computing schems of parallellized algorithm is:
Stage A - Input-output block Hankel matrices
Stage B - SVD, observability matrix, controlability matrix, calculation of A, B, B_0, C, D, D_0 and QSR matrices
Stage C - Test module

The tables 1,2,3,4 show the comparison of the indicators of speedup, throughput, delay of parallel performance algorithm for the sequential algorithm.

TABLE I
Evaluation of the parallel algorithm—speedup for $m = 255$

Parallel algorithm			
Stage	A	B	C
	0,086[s]	0,105[s]	0.049[s]

Sequential algorithm			
Stage	A	B	C
	0,61[s]	0,42[s]	0.34[s]

TABLE II
Evaluation of the parallel algorithm—efficiency for $m = 255$

Stage	A	B	C
	0,028[s]	0,015[s]	0.027[s]

TABLE III
Evaluation of the parallel algorithm—transport delay for $m = 255$

Sequential algorithm	Parallel algorithm
1,2[s]	0,17[s]

TABLE IV
Evaluation of the parallel algorithm—throughput for $m = 255$

Sequential algorithm	Parallel algorithm
0,6[GFlop/s]	18,79[GFlop/s]

The algorithms were tested on a PC computer with Genuine Intel CPU T2300 @ 1,66 GHz, and 1536 MB RAM NVIDIA GTX 460 graphic card. Figure 2 shows the speedup of a parallelized algorithm N4SID and in Figure 3 MOESP

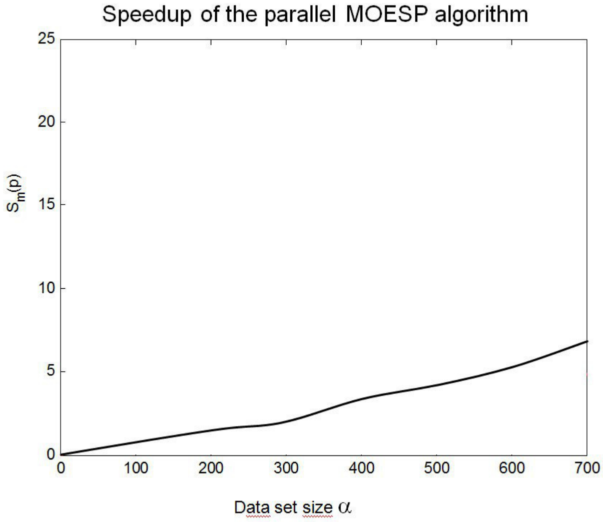


Fig. 2. Speedup of the parallel MOESP algorithm

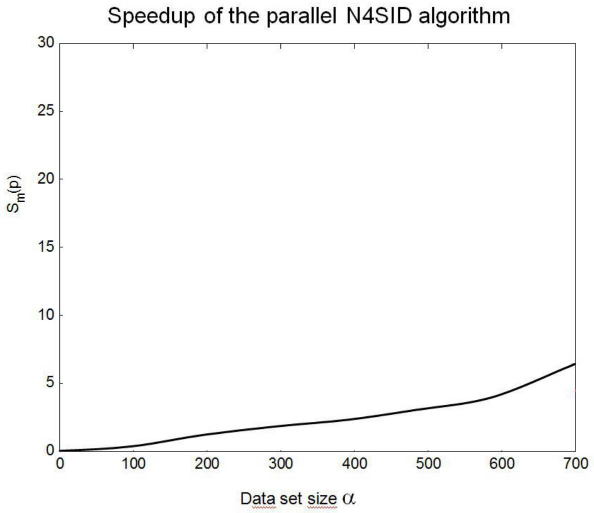


Fig. 3. Speedup of the parallel N4SID algorithm

parallelized algorithm speedup as a function of the number of measurements of α. The pass length was changed from 5 to 700 and the run time for the sequential modified algorithm N4SID and MOESP and their parallel versions was measured. Based on these measurements, the speed up was evaluated. Speedups, shown in Fig. 2 and 3, are equal to formally calculated speedups of parallel algorithms. Parallel implementations run about 5-6 times faster than their sequential versions.

VI. Conclusions

For Stages 1 and 3, the following indicators: speedup S_m, throughput P and computational efficiency E_m increase. At the same time, the transport delay L decreases. For Module 2, speedup, throughput P, and efficiency decrease while the delay transport increases. This comes from the calculation of the system order using SVD decomposition. Speedups, shown in Fig. 2 and 3, are equal to formally calculated speedups of parallel algorithms. Parallel implementations run about 5-6 times faster than their sequential versions. Transport delay is interpreted as the time of calculation. The speedup and throughput efficiency are good candidates for indices to evaluate the degree of parallelization. A parallel algorithm may be useful in selecting fast and using less resources algorithm for identification of stationary linear repetitive processes.

Acknowledgment

The author is a scholar participating in Sub-measure 8.2.2. Regional Innovation Strategies, Measure 8.2. Transfer of Knowledge, Priority VIII Regional Human Resources for the Economy Human Capital Operational Programme co-financed by the European Union Social Fund and the State Budget of Poland.

References

[1] R. Hartenstein, “Are we really ready for the breakthrough? ”, In Parallel and Distributed Processing Symposium, page 7, April 2003

[2] Z. Baker, M. Gokhale, and J. Tripp, “Matched filter computation on fpga, cell and gpu”, In 15th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, pages 207-218, USA, April 2007

[3] A. Janczak, D. Kujawa, E. Rogers, Z. Cai , “Subspace identification of process dynamica for iterative learning control”, Proceedings of the 8th International Workshop on Multidimensional Systems, Erlangen, Niemcy, 2013

[4] Tomasz Zawadzki, Dominik Kujawa, “The hybrid algorithm for procedural generation of virtual scene components”, Lecture Notes in Computer Science: 8th International Symposium, 2012

[5] A. Chiuso and G. Picci, “Some algorithmic aspects of subspace identification with inputs”, In Int. J. Appl. Math. Comp. Sci., 2001, Vol. 11, No. 1, 55-75, 2001

[6] E. Rogers, K. Galkowski and D. H. Owens “Control Systems Theory and Applications for Linear Repetitive Processes”, Springer, 2007

[7] T. Katayama, “Subspace methods for system identification”, Springer, 2008

[8] P. Van Overschee and B. De Moor, “Subspace Identification for Linear Systems”, Kluwer Academic Publishers, 2001

[9] D. Lyons and I. Giselle, “Evaluation of a Parallel Architecture and Algorithm for Mapping and Localization”, In Proceesings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, pages 546-551, June, 2007

[10] J. Stój, “Temporal aspects of redundancy in distributed real-time computer systems (original title: Wpływ redundancji na zależności czasowe w rozproszonych informatycznych systemach czasu rzeczywistego)”, PHD Thesis, Silesian University of Technology, Gliwice, Poland, 2009

DOI: 10.15439/2014F247
ACSIS, Vol. 3

Pollutant's Dispersion in the Atmosphere. Parallel Models and Applications

Marcin Majer
Faculty of Electrical Engineering, Automatic Control
and Computer Science
Opole University of Technology
Opole, Poland
m.majer@po.opole.pl

Michal Podpora
Faculty of Electrical Engineering, Automatic Control
and Computer Science
Opole University of Technology
Opole, Poland
m.podpora@po.opole.pl

Abstract—**The following article contains information about modeling of pollutants dispersion systems in the atmosphere. It also describes main types of pollutants, emission points and analysis of typical atmosphere pollutant dispersion models. It shows a way of using computer cluster systems in modeling process and proper programming libraries for parallel computing. Finally it presents a possibility of environment's protection method during the emission points' location planning with calculation of minimal destructive influence on a specified area (for example with definite restrictions of air pollution level).**

Keywords—parallel computing; distributed computing; computer cluster; atmosphere; pollutant dispersion models; infrastructure expansion planning, MPI.

I. Introduction

TO MODEL the process of pollution spreading in the natural environment, engineers and analysts extend typical atmosphere pollution monitoring systems to co-operate with computer clusters.

In a high technology age we live, single PC computers are often not enough for complicated computing problems contained in atmosphere pollutant dispersion models. It's also useful to use parallel programing methods in application creating process. There are many ready-to-use libraries for different programing languages but applying them is possible only if algorithm's model is designed for distributed architecture implementations.

The authors prepared an application to calculate hourly air pollutants concentrations in specified data points of an interest region. The application's algorithm involves solving emitter's location optimization problem in terms of fulfilling the levels of pollutants in a given area. Simulations and calculations were carried out on real input data describing specified area, including weather conditions. The following chapters describe the research and present the conclusions.

II. Pollutants Dispersion in the Atmosphere

A. Air pollution

Air pollution becomes a huge problem. Generally, there are two main methods for pollutant's dispersion systems modeling: realized with physical models and mathematical models. Both of them have the same starting point: wide knowledge about atmosphere's lower border layer characteristic. Physical models are based on executing atmospherics processes in laboratory conditions, whereas mathematical models contain mathematical description for physical and mathematical processes built on empiric measurement data. They are often called simulative deterministic models. In order to create a proper mathematical model it's important to classify substances in atmospheric air. It's not a challenge to find a number of substances in the atmosphere that are harmful for people, animals or plants. Emission of those pollutants leads to changing average established air accumulation.

TABLE I. Average Established Air Accumulation

Gas type		Content [%]
Nitrogen	N_2	78,09
Oxygen	O_2	20,95
Argon	Ar	0,93
Carbon dioxide	CO_2	0,03
Neon	Ne	$2\cdot10^{-3}$
Helium	He	$5\cdot10^{-4}$
Methane	$C H_4$	$2\cdot10^{-4}$
Krypton	Kr	$1\cdot10^{-4}$
Hydrogen	H_2	$6\cdot10^{-5}$
Nitrogen monoxide	N_2O	$3\cdot10^{-5}$
Xenon	Xe	$9\cdot10^{-6}$

The main parameters for air pollution description are: air pollutant concentration (relation of the pollutant quantity and air capacity containing it) and pollutant's stream directed to the ground (pollutant quantity ground settled per unit of area in a unit of time).

B. Air pollution spreading models

At the beginning of building atmosphere pollution monitoring system it is important to perform the parameters selection during the process of building the atmosphere's pollution state describing equation. The parameters are strongly dependent on various factors (inter alia: territorial, meteorological, and factors describing the point of pollutant's emission). The following steps are to be performed:

- collecting the data with release of dirt from different sources (for example the chimney as a punctual source),
- checking the meteorological conditions (speed and direction of wind, coefficient of diffusion, temperature, humidity),
- gathering additional information (the form of terrain, stage of afforestation).

After completing these steps it's necessary to choose one of possible air pollution spreading models.

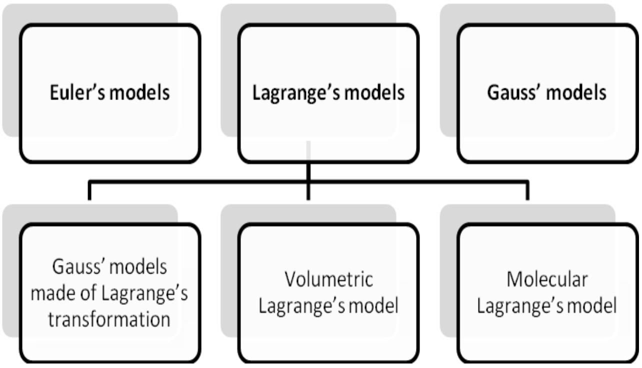

Euler's models
Lagrange's models
Gauss' models
Gauss' models made of Lagrange's transformation
Volumetric Lagrange's model
Molecular Lagrange's model

Fig. 1. Typical air pollution spreading models

There are three main types of pollution spreading models (fig. 1):

- Euler's models – the most important issue in these models is the pollution transport equation (1)[1]:

$$\frac{\partial\tilde{c}}{\partial t}+\frac{\partial(\tilde{u}\tilde{c})}{\partial x}+\frac{\partial(\tilde{v}\tilde{c})}{\partial y}+\frac{\partial(\tilde{w}\tilde{c})}{\partial z}=$$

$$D_C\left(\frac{\partial^2\tilde{c}}{\partial x^2}+\frac{\partial^2\tilde{c}}{\partial y^2}+\frac{\partial^2\tilde{c}}{\partial z^2}\right)+\tilde{S}_C \qquad (1)$$

where: $\tilde{C}$ - air pollutant concentration (relation of pollutant quantity and air capacity containing it), $\tilde{u},\tilde{v},\tilde{w}$ – direction of wind vector's $(\tilde{U})$ ingredients in square system of coordinates related with Earth along axes: OX, OY, OZ; D_C – pollution diffusion coefficient, $\tilde{S}_C$ – losses and pollutants source describing module.

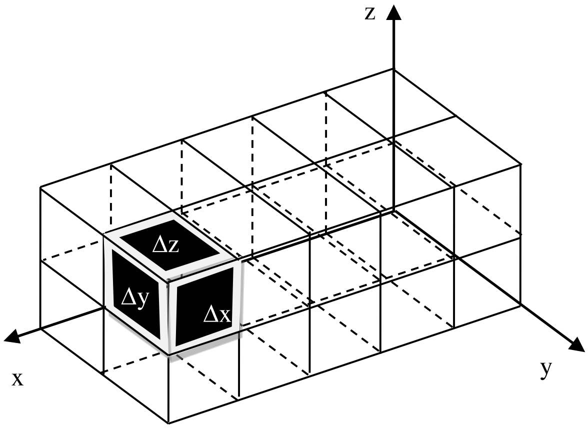


Fig. 2. *Computation's scheme grid in Euler's air pollution spreading models (moving "box" of pollutants)*

The most known Euler's formula implementations are: MISCAM (Germany 1996), MOD4 (Poland 1989), STEM-II (USA 1991).

- Lagrange's models – based on Lagrange's pollution dispersion equation (2):

$$\overline{C(r,t)} = C(r,t) =$$

$$= \int_{-\infty}^{t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(r,t|r',t')S_C(r',t')dr',dt' \qquad (2)$$

where: $C(r,t)$ - air pollutant concentration averaged in realization class in r point for time t; $S_C(r',t')$ – losses and pollutants source describing module; $p(r,t|r',t')$ - probability's density function that hypothetical air capacity that is placed in point r' at the time t' will be placed in point r at the time t .

The most known Lagrange's model implementations are: MATHEW/ADPIC (USA 1988), ARCO (Italy 1992), MDMS (USA 1983).

- Gauss models – based on either Lagrange's or Euler's pollution dispersion method with additional assumption equations for essential simplification. The main assumptions are:

- concentration's pole is stabile in time: $\frac{\partial c}{\partial t}=0$

-air moves are mostly horizontal, in OX axis direction with speed: $u \geq 0, v = w = 0$

- there are no volumetric pollution sources: $S_C = 0$

-pollutant advection's module is much stronger than turbulent module:

$$u\frac{\partial C}{\partial x} \gg \frac{\partial}{\partial x}\left(K_x\frac{\partial C}{\partial x}\right) \qquad (3)$$

After all transformations the Gauss formula takes the form as follows (4):

$$(x,y,z) = \frac{E}{2\pi\sigma_y\sigma_z\breve{u}}exp\left(-\frac{y^2}{2\sigma_y^2}\right)\left[exp\left(-\frac{(z-H_e)^2}{2\sigma_z^2}\right) + exp\left(-\frac{(z+H_e)^2}{2\sigma_z^2}\right)\right] \qquad (4)$$

where: $C(x,y,z)$ - pollutant in air concentration at (x,y,z) point; E – pollutant emission's intensity; H_e – effective emission's height.

Gauss formula, defined in bibliography as Pasquill's formula (4) is strongly recommended to be applied in cases where average wind speed $\breve{u} > 1\frac{m}{s}$. Furthermore, it is one of the most frequently realized and implemented models. The most known Gauss formula's implementations are: RAM (USA 1989), KOMIN (Poland 1998), AERMOD (USA 1995), CALPUFF (USA 1977, later modified).

III. Parallel implementation

There are plenty implementations of described models but most of them are sequential. Nowadays, extending these models to the capability of parallel/distributed implementation is highly recommended because of extremely long time of computer calculation. The best hardware solution is to use computer clusters if possible.

To implement a model in a distributed environment it is important to choose a suitable parallel programming library, and a supported programming language. Authors have chosen to adopt a sequential model by redesigning and implementing it using the LAM-MPI parallel library. The Message Passing Interface daemon executes the same code on all machines (nodes) in the same time; the nodes can compute individual parts of data and to exchange information by passing messages. The MPI specification –based libraries (e.g. LAM-MPI, MPICH2, etc.) are supposed to fully support heterogeneous computer clusters (i.e. of various operating systems and/or hardware architectures), the algorithm should be designed to support a variety of platforms (i.e. hardware configurations of a specific node, varying computational resources among the computer cluster's nodes). According to Flynn's classification [2] MPI implements MIMD model (Multiple Instruction stream - Multiple Data stream). Furthermore, MPI allows to use all the benefits of computer clusters:

- beneficial rate of price to performance,
- fault tolerance,
- high accessibility,
- hardware scalability,
- high performance.

Another way for minimizing computation time on computer clusters is by using OpenMP for additional parallelization within a specific node. OpenMP is an API that supports multi-platform shared memory multiprocessing programming in C, C++, and Fortran. An application implemented to support the hybrid model of parallel programming may be executed on a computer cluster using both OpenMP and Message Passing Interface (MPI), or more transparently through the use of OpenMP extensions for non-shared memory systems.

Supercomputers can bring different way profits. It's because of their hardware architecture. 2-level structure (figure 3) of some clusters gives multiply possibilities for parallel programming using MPI. It is a great step towards minimization of communication time. It is also possible to get additional acceleration of parallel algorithms by using simple methods like CPU gathering or shared memory utilization [3],[5].

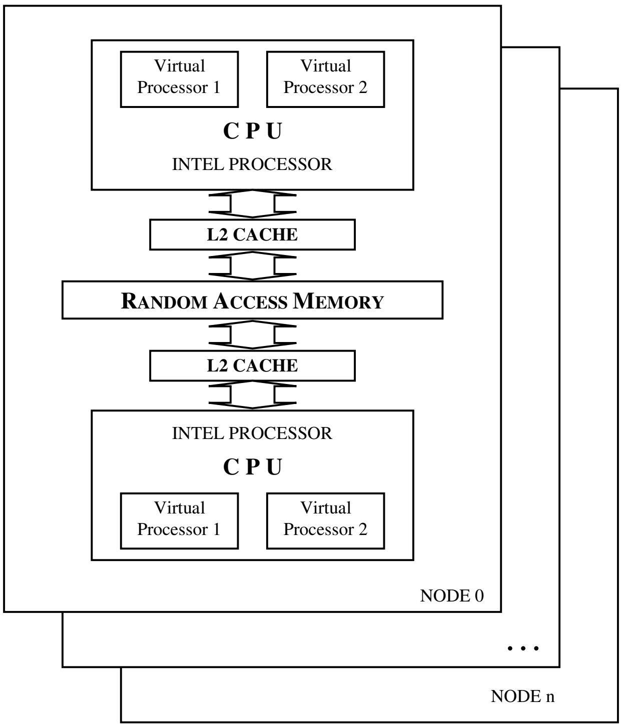


Fig. 3. *Possible cluster's node architecture*

The results of applying CPU gathering and shared memory utilization are shown in figure 4.

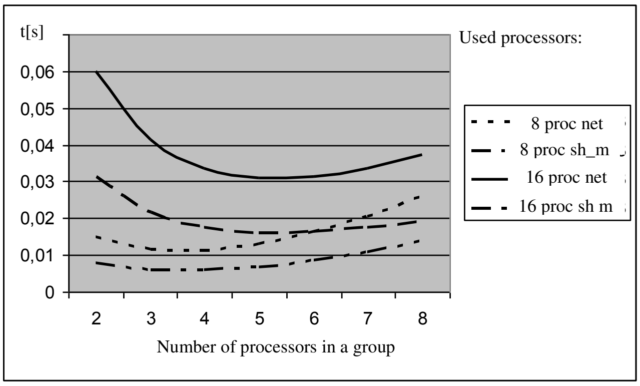
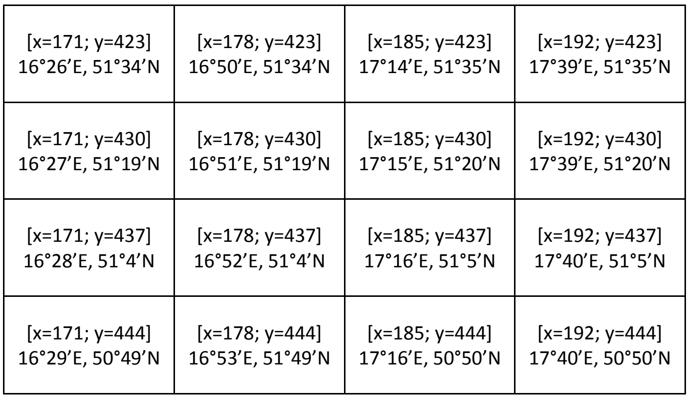


Fig. 4. Times of full data broadcast grouping processors (using shared memory and network) [own work]

IV. RESEARCH METHODOLOGY

During the application development, it was determined that the main task (further called "the solver") was the calculation of hourly air pollutant concentrations in the specified data points of the interest region.

In relation to the assumed regional reach of the system controlling the proliferation of pollutants, a field grid was established for conducting measurements of weather factors. The span for it was 84.17 km per 83.56 km, however average distance of measuring knots was 27.67 km. Spreading test points of the grid was referred to GPS coordinates as it was shown in figure 5.

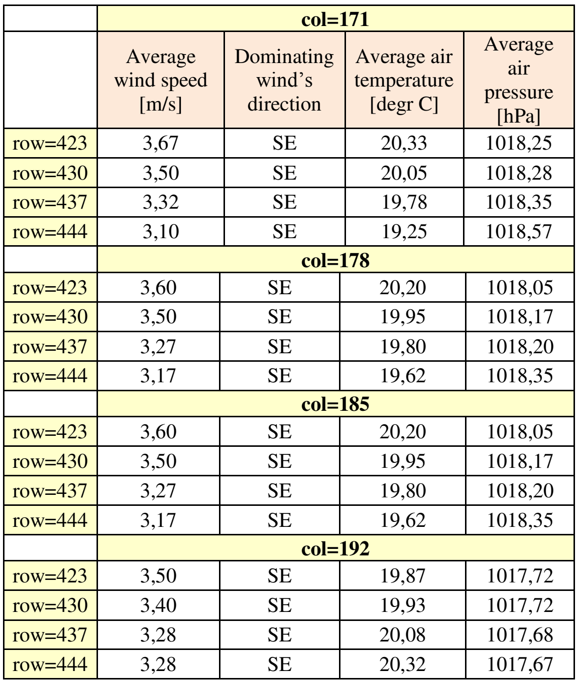


[x=171; y=423] 16°26'E, 51°34'N	[x=178; y=423] 16°50'E, 51°34'N	[x=185; y=423] 17°14'E, 51°35'N	[x=192; y=423] 17°39'E, 51°35'N
[x=171; y=430] 16°27'E, 51°19'N	[x=178; y=430] 16°51'E, 51°19'N	[x=185; y=430] 17°15'E, 51°20'N	[x=192; y=430] 17°39'E, 51°20'N
[x=171; y=437] 16°28'E, 51°4'N	[x=178; y=437] 16°52'E, 51°4'N	[x=185; y=437] 17°16'E, 51°5'N	[x=192; y=437] 17°40'E, 51°5'N
[x=171; y=444] 16°29'E, 50°49'N	[x=178; y=444] 16°53'E, 51°49'N	[x=185; y=444] 17°16'E, 50°50'N	[x=192; y=444] 17°40'E, 50°50'N

Fig. 5. GPS coordinates describing points of the weather factors grid [own work]

The factors describing weather conditions were gathered during 30 days period. After the analysis of the collected meteorological data, it was essential to average the data and to determine the dominating weather conditions in points of the measuring grid. After that process, the weather conditions data were ready to be applied to computation algorithms (fig. 6).

All the meteorological data, as it was shown above, were taken from measurements and observations. That methodology enables full result's verification and enables to examine the solver's behavior on a real model data.



	col=171			
	Average wind speed [m/s]	Dominating wind's direction	Average air temperature [degr C]	Average air pressure [hPa]
row=423	3,67	SE	20,33	1018,25
row=430	3,50	SE	20,05	1018,28
row=437	3,32	SE	19,78	1018,35
row=444	3,10	SE	19,25	1018,57
	col=178			
row=423	3,60	SE	20,20	1018,05
row=430	3,50	SE	19,95	1018,17
row=437	3,27	SE	19,80	1018,20
row=444	3,17	SE	19,62	1018,35
	col=185			
row=423	3,60	SE	20,20	1018,05
row=430	3,50	SE	19,95	1018,17
row=437	3,27	SE	19,80	1018,20
row=444	3,17	SE	19,62	1018,35
	col=192			
row=423	3,50	SE	19,87	1017,72
row=430	3,40	SE	19,93	1017,72
row=437	3,28	SE	20,08	1017,68
row=444	3,28	SE	20,32	1017,67

Fig. 6. Averaged and dominating weather factors according to field grid - used in computations [own work]

Technical parameters of the potential emitter in the form of the chimney were determined in the following way:

- height [m]: 44;
- inside diameter of the emitter's wire outlet: 1,6 m;
- speed of gasses on the exit of the emitter: 14,4 m/s
- gas temperature on the exit: 493,8 K;
- maximum emission of a gaseous substance: 43,9 mg/s.

There are also many other parameters needed for the solver, one of them is the aerodynamic coarseness coefficient of the area. It is selected according to the scheme shown in table 2.

TABLE II. THE AERODYNAMIC COARSENESS COEFFICIENT OF THE AREA SELECTION PROCESS [6].

	Type of the area's covering	**Aerodynamic coarseness coefficient of the area**
1.	Water	0,00008
2.	Meadows, pastures	0,02
3.	Farmland	0,035
4.	Orchards, thickets, groves	0,4
5.	Forests	2,0
6.	Country high-housing	0,5
7.	City up to 10000 residents	1,25
8.	City 10000 to 100000 residents	2,75
9.	City over 500000 residents	3,75

To complete all the necessary entry values for calculating model one more factor is needed - the atmosphere's balance factor (ATM). It is determined by the observations of atmosphere's stormy nature and additionally it's described by scopes of the permissible wind speeds for individual states. The principle of atmosphere's balance factor's selection is shown in the table 3.

TABLE III. THE ATMOSPHERE'S BALANCE FACTOR SELECTION PROCESS [6].

***Atmosphere's* balance factor ATM**	***Atmosphere's balance* state**	**Scope of the wind speed m/s**
1	highly unstable	1 – 3
2	unstable	1 – 5
3	medium unstable	1 – 8
4	indifferent	1 – 11
5	medium stable	1 – 5
6	stable	1 – 4

Preliminary computational tests of proper working and computer cluster's usage reasonability were investigated during Pasquill's formula model implementation in a Linux environment. The solver was implemented in C++ using an MPI library and OpenMP directives.

The computational resources for the research included the computer cluster available on the Faculty of Electrical Engineering, Automatic Control and Computer Science at Opole University of Technology. It is a processing unit with following technical parameters:

TABLE IV. CLUSTER'S TECHNICAL SPECIFICATION

Processor architecture:	x86 64 bit
Processor type and frequency:	AMD Phenom II X6 1090T s. AM3
Number of processors in a single node:	1
Number of cores in a single processor:	6
Number of nodes:	16
Number of cores (summary):	48
RAM available in a single node:	8 GB
Hard drive capacity (summary)	5 TB
Network:	2 x Gigabit Ethernet

The operating system installed in the nodes of the computer cluster is Linux. The main solver tests concerned legitimacies of the computer clusters' application in this kind of computation as a pollutant's dispersion in the atmosphere's parallel models and applications issues.

V. RESULTS

Calculating hourly concentrations of chosen pollutant is one of the examples for CPU consuming calculations benchmarking the cluster computational efficiency through a potential modeled chimney, for the domains adjacent to the chimney towards the dominating wind on the given area. The values were calculated for the points distant from oneself at 1 m in the distance of 10000 m in directions parallel to the x axis and 5000 m perpendicularly to the x axis (symmetrically towards axis).

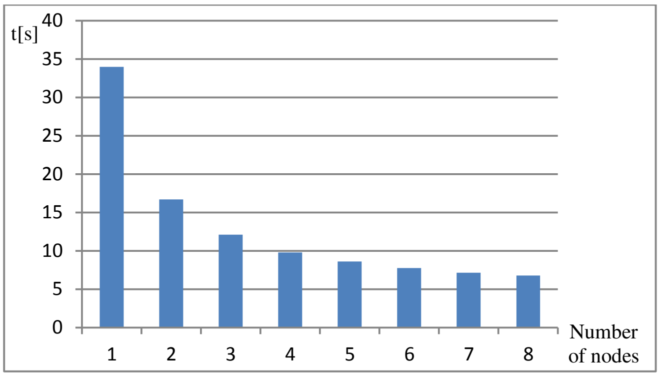


Fig. 7. Relation of the computation time of the hourly concentration of pollutants for grid of 10000 x 5000 points to the number of computational nodes [own work]

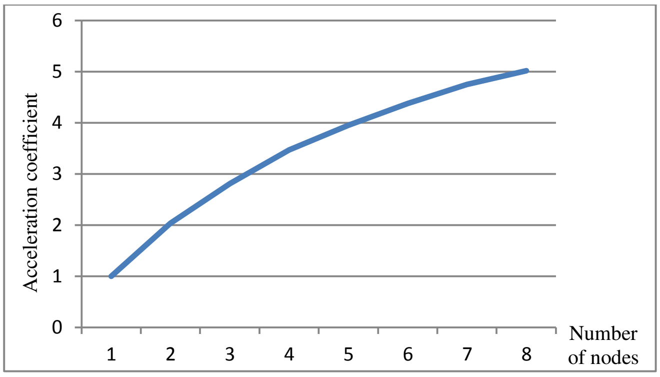


Fig. 8. Dependence of the coefficient of the acceleration on the number of computational nodes for calculation times introduced in figure 7.

In another test problem the MPI library was replaced with OpenMP directives. It required redesigning of the algorithm.

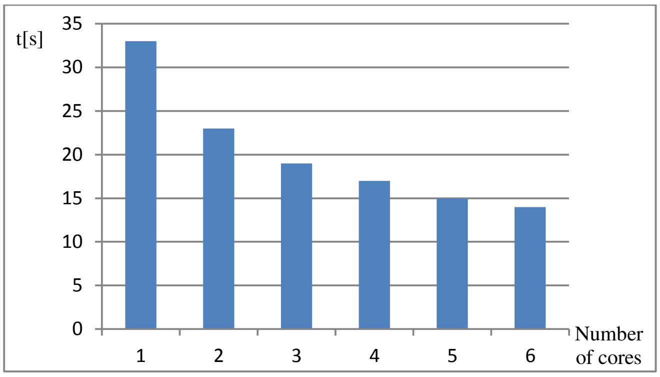


Fig. 9. Relation of the computation time of the hourly concentration of pollutants for the grid of 10000 x 5000 points to the number *of processor's cores in a single node.*

It allowed for checking the advantages of multicore processors programming on the level of single node. The test's results were shown at the following figures.

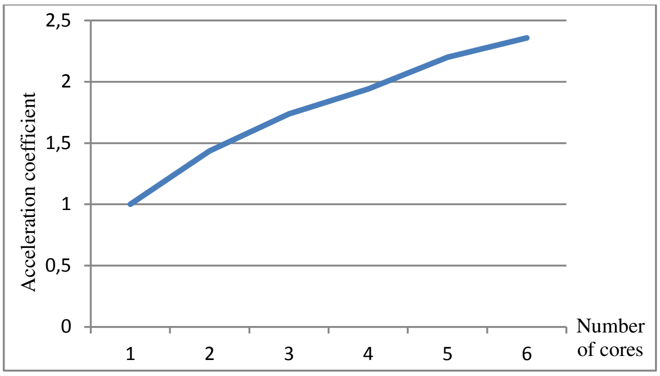


Fig. 10. Dependence of the coefficient of the acceleration on the number of *computational processor's cores for computation times introduced* in figure 9.

The following figure presents the solver's results of hourly pollutant's concentration calculation. That kind of data can be used in further computation of optimizing placing new emitter to fulfill minimal destructive influence assumptions on a specified area.

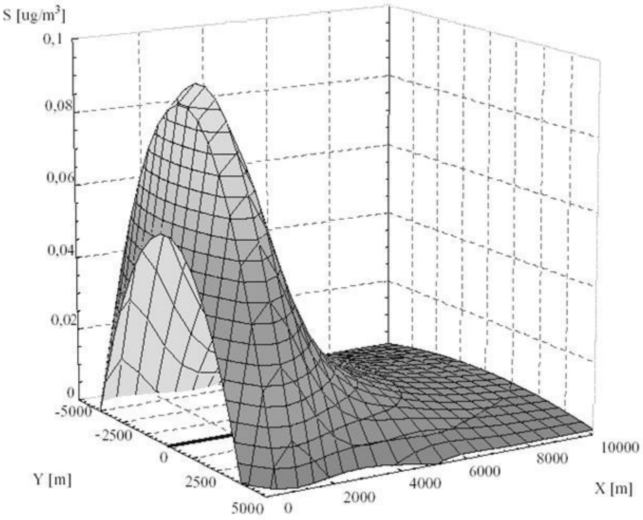


Fig. 11. *Results of hourly pollutant's concentration calculation*

VI. Summary

The specificity of computer clusters and software configuration creates a real basis for efficiency improvement of solving the computational pollution's spreading and accumulation process in the atmosphere issues. Parallel programming models can be used to effectuate a higher degree of parallelism via the simultaneous execution of separate parts of application's code on different processors, also grouping processors in calculations tasks. Parallelization of programs continues to remain a human challenge but it can bring high profits. Furthermore, this type of results of calculation can be used to emitter's location optimization problem in terms of fulfilling the levels of pollutants in a given area. It can be useful during infrastructure expansion planning process.

References

[1] Findeisen W., Szymanowski J., Wierzbicki A.: Metody obliczeniowe optymalizacji. WPW, W-wa 1973

[2] Flynn M.J.: Some Computer Organizations and Their Effectiveness, IEEE Trans. on Comp. vol C-21, No 9, 1972

[3] Kaliczyńska M., Sadecki J.: Obliczenia równoległe – klastry obliczeniowe. Elektryka z. 57, Politechnika Opolska, Opole 2006

[4] Karbowski A., Niewiadomska – Szynkiewicz E.: Obliczenia równoległe i rozproszone. OWPW, W-wa 2001

[5] Sadecki J.: Algorytmy równoległe optymalizacji i badania ich efektywności; systemy równoległe z rozproszoną pamięcią. Studia i monografie. Z.126, Politechnika Opolska, Opole 2001

[6] Dz. U. 2010 nr 16. poz 87. 2010. Rozporządzenie Ministra Środowiska z dnia 26 stycznia 2010 roku w sprawie wartości odniesienia dla niektórych substancji w powietrzu. Warszawa: Dz. U. 2010, nr 16. poz 87., 2010

2nd Workshop on Scalable Computing in Distributed Systems and 7th Workshop on Large Scale Computations on Grids

The Large Scale Computing in Grids (LaSCoG) workshop originated in 2005, and when it was created we have stated in its preamble that:

"The emerging paradigm for execution of large-scale computations, whether they originate as scientific or engineering applications, or for supporting large data-intensive calculations, is to utilize multiple computers at sites distributed across the Internet. In particular, computational Grids are collections of distributed, possibly heterogeneous resources which can be used as ensembles to execute large-scale applications. While the vision of the global computational Grid is extremely appealing, there remains a lot of work on all levels to achieve it."

While, it can hardly be stated that the issues we have observed in 2005 have been satisfactorily addressed, a number of changes has happened that expanded the world of large-scale computing. Today we can observe emergence of a much more general paradigm for execution of large-scale applications, whether they originate from scientific or engineering areas, or they support large data-intensive calculations. These tasks utilize computational Grids, cloud-based systems and resource virtualization. Here, collections of distributed, possibly heterogeneous resources, are used as ensembles to execute large-scale applications.

This being the case, we have decided to keep the LaSCoG workshop tradition alive, but to co-locate it with a conference which will have an appropriately broader scope. This is how the Workshop on Scalable Computing in Distributed Systems (SCoDiS'14) emerged.

The LaSCoG-SCoDiS'14 pair of events shares a joint Program Committee and is envisioned as a forum to promote an exchange of ideas and results aimed at addressing sophisticated issues that arise in developing large-scale applications running on heterogeneous distributed systems.

Topics

Covered topics include (but are not limited to):

- Large-scale algorithms and applications
- Symbolic and numeric computations
- High performance computations for large scale simulations
- Large-scale distributed computations
- Agent-based computing
- Data models for large-scale applications
- Security issues for large-scale computations
- Science portals
- Data visualization
- Performance analysis, evaluation and prediction
- Programming models
- Peer-to-peer models and services for scalable Grids
- Collaborative science applications
- Business applications
- Data-intensive applications
- Operations on large-scale distributed databases
- On-demand computing
- Computation as a service
- Federation of compute capacity
- Virtualization supporting computations
- Self-adaptive computational / storage systems
- Volunteer computing
- Large scale computation with GPU
- Cloud computing architectures and models
- Load Balancing in large-scale distributed systems
- Intelligent resource allocation
- Cloud Security, Privacy, Confidentiality and Compliance
- Mobile Cloud
- High Performance Cloud Computing
- Green Cloud Computing
- Economic, Business and ROI Models for Cloud Computing
- Performance, Capacity Management and Monitoring of Cloud Configuration
- Cloud Interoperability and Portability
- Cloud Application Scalability and Availability
- Big Data Cloud Service

Event Chairs

Gusev, Marjan, University Sts Cyril and Methodius, Macedonia

Paprzycki, Marcin, Systems Research Institute Polish Academy of Sciences, Poland

Petcu, Dana, West University of Timisoara, Romania

Ristov, Sasko, University Sts Cyril and Methodius, Macedonia

Program Committee

Anderson, David, University of California, Berkeley, United States

Bass, Len, NICTA, Australia

Brodnik, Andrej, University of Ljubljana, Faculty of Computer and Information Science, Slovenia

Camacho, David, Universidad Autonoma de Madrid, Spain

D'Ambra, Pasqua, ICAR-CNR, Italy

Feldmann, Anja

Filippone, Salvatore, University Rome Tor Vergata, Italy

Ganzha, Maria, University of Gdańsk and Systems Research Institute Polish Academy of Sciences, Poland

Gepner, Paweł, Intel Corporation, United Kingdom

Gordon, Minor, Software development consultant, United States

Gorgan, Dorian, Technical University of Cluj-Napoca, Romania

Goscinski, Andrzej, Deakin University, Australia

Gravvanis, George, Democritus University of Thrace, Greece

Grosu, Daniel, Wayne State University, United States

Holmes, Violeta, The University of Huddersfield, United Kingdom

Hsu, Ching-Hsien (Robert), Chung Hua University, Taiwan

Kalinov, Alexey, Cadence Design Systems, Russia

Karaivanova, Aneta, IICT-BAS, Bulgaria

Kitowski, Jacek, AGH University of Science and Technology, Department of Computer Science, Poland

Knepper, Richard, Indiana University, United States

Kranzlmüller, Dieter, Ludwig-Maximilians-Universität München (LMU), Germany

Kwiatkowski, Jan, Wrocław University of Technology, Poland

Lang, Tran Van, Vietnam Academy of Science and Technology, Vietnam

Lastovetsky, Alexey, University College Dublin, Ireland

Lee, Dongwoo, Gulf University for Science and Technology & OikoLab, Kuwait

Legalov, Alexander, Siberian Federal University, Russia

Luo, Mon-Yen, National Kaohsiung University of Applied Sciences, Taiwan

Margaritis, Konstantinos G., University of Macedonia, Greece

Milentijevic, Ivan, University of Nis, Serbia

Morrison, John, University College Cork, Ireland

Nosovic, Novica, Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina

Olejnik, Richard, CNRS - University of Lille I, France

Ouedraogo, Moussa, Public Research Centre Henri Tudor, Luxembourg

Rak, Massimiliano, Seconda Università di Napoli, Italy

Schikuta, Erich, University of Vienna, Austria

Schreiner, Wolfgang, Johannes Kepler University Linz, Austria

Shen, Hong, University of Adelaide, Australia

Song, Ha Yoon, Hongik University, South Korea

Stankovski, Vlado, University of Ljubljana, Slovenia

Talia, Domenico, ICAR-CNR and University of Calabria, Italy

Telegin, Pavel, JSCC RAS, Russia

Trystram, Denis, Grenoble Technical University, France

Tudruj, Marek, Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland

Tvrdik, Pavel, Faculty of Information Technology, Czech Technical University in Prague, Czech Republic

Vazhenin, Alexander, University of Aizu, Japan

Wei, Wei, School of Computer science and engineering, Xi'an University of Technology, China

Wyrzykowski, Roman, Czestochowa University of Technology, Poland

Xu, Baomin, Beijing JiaoTong University, China

Zavoral, Filip, Charles University in Prague, Czech Republic

DOI: 10.15439/2014F325
ACSIS, Vol. 3

A self-stabilizing algorithm for locating the center of Cartesian product of K_2 and maximal outerplanar graphs

Halina Bielak
Institute of Mathematics
Maria Curie-Skłodowska University in Lublin
pl. Marii Curie-Skodowskiej 1, 20-031 Lublin, Poland
Email: hbiel@hektor.umcs.lublin.pl

Michał Pańczyk
Institute of Computer Science
Maria Curie-Skłodowska University in Lublin
ul. Akademicka 9, 20-033 Lublin, Poland
Email: mjpanczyk@gmail.com

***Abstract*—Self-stabilizing algorithms model distributed systems and allow automatic recovery of the system from transient failures. The center of a graph is the set of vertices with the minimum eccentricity. In this paper we investigate the self-stabilizing algorithm for finding the center of Cartesian product of K_2 and maximal outerplanar graphs.**

I. Introduction

Let $G = (V(G), E(G))$ be a simple, connected graph with the vertex set V and edge set E. The distance $d(i, j)$ between nodes i and j is the length of the shortest path connecting these two nodes. The maximum distance from a given vertex i to any other vertex in the graph G is called the eccentricity $ecc(i)$ of the vertex i. The set $C(G)$ of the vertices with the minimum eccentricity is called the center of the graph G (see Fig. 1).

The Cartesian product $G_1 \square G_2$ of simple graphs G_1 and G_2 is a graph with the vertex set

$$V(G_1 \square G_2) = V(G_1) \times V(G_2)$$

and the edge set

$$\begin{gathered} E(G_1 \square G_2) = \{\{(u_1, u_2), (v_1, v_2)\} \mid \\ u_1, v_1 \in V(G_1) \wedge u_2, v_2 \in V(G_2) \wedge \\ ((u_1 = v_1 \wedge \{u_2, v_2\} \in E(G_2)) \vee \\ (u_2 = v_2 \wedge \{u_1, v_1\} \in E(G_1)))\}. \end{gathered}$$

In many cases it is important to locate the center of a graph, especially in distributed systems, where it allows placing a control center for the minimum cost of communication with peripherial nodes of the system. There are several known algorithms for locating centers in graphs. Bielak and Pańczyk [1] proposed algorithm finding weighted centroid in a tree. Farley [6] gave a linear time algorithm for vertex centers in trees. Also Hedetniemi et al. [11] gave linear time algorithm for center problems in trees. Goldman [9] and Kariv and Hakimi [12] gave an algorithm solving the center problem in networks.

A lot of research has been realized related to centers of graphs [14], [15], [3]. Distributed algorithms were also developed [2], [13]. In this paper we propose a self-stabilizing algorithm for locating the center of Cartesian product of complete graph K_2 and maximal outerplanar graph. The problem for maximal outerplanar graphs, in classical, sequential computing paradigm was solved by Farley and Proskurowski [7]. Let us define a maximal outerplanar graph as it was done in the mentioned paper [7], as a triangularization of a planar polygon (see Fig. 1). We define K_2 as a complete graph with two vertices.

In a maximal outerplanar graph $M = (V(M), E(M))$ every edge $p = \{i, j\} \in E(M)$ partitions the set of all vertices apart i and j into two distinct sets inducing connected subgraphs called sides. One of the sides may be empty. It is the case when the partitioning edge is a part of the exterior face of the graph. In fact, all the edges with one side empty form the unique Hamiltonian cycle.

Let us note that in a maximal outerplanar graph every two neighbors i and j have at most two common neighbors, each of them belonging to distinct

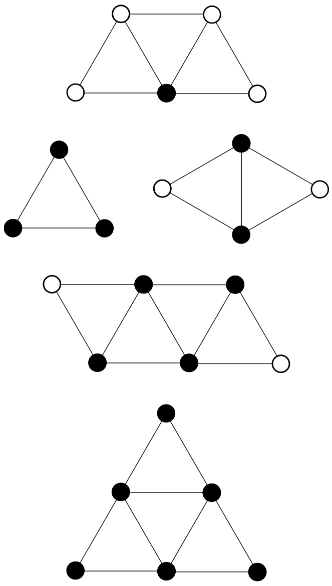


Fig. 1. Examples of centers (marked as black) in maximal outerplanar graphs.

sides of the edge $\{i, j\}$. Thus, it is sufficient to represent the side of the edge by one of the two common vertices adjacent to the edge. If a side is empty, we set $\emptyset$ as its representation.

In the Cartesian product of K_2 and any graph G, we can identify two layers of which every one is isomorphic to the graph G.

Farley and Proskurowski [7] introduced a notion of the edge eccentricity. Let us have the node i, its neighbor j and one of their common neighbors k ($\emptyset$ for the one nonexistent if applicable) in a graph G. For these three values we define $e(i, j, k)$ (edge eccentricity) in the following manner:

- the absolute value of $e(i, j, k)$ is equal to the eccentricity of the vertex i in the subgraph of G induced by $S_k \cup \{i, j\}$, where S_k is the side of the edge $\{i, j\}$ containing the vertex k,
- $e(i, j, k)$ is negative integer iff all vertices of $S_k \cup \{i, j\}$ at distance $d = |e(i, j, k)|$ from i lie at distance $d - 1$ from the vertex j.

The classical algorithm of Farley and Proskurowski [7] computes the edge eccentricity for every edge recursively using already computed values of the eccentricities for adjacent edges. It starts with outerface edges, for which the edge eccentricity (on an empty side) is equal to -1. All the details can be found in [7].

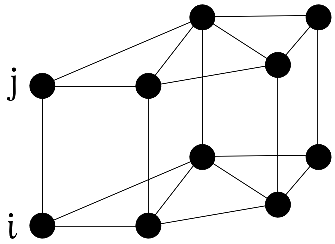


Fig. 2. An example of $K_2 \square$ maximal outerplanar graph.

In this paper we propose a self-stabilizing algorithm for finding the center in the Cartesian product of K_2 and a maximal outerplanar graph. In the following section we introduce the computational model used futher in the paper. In section III we show the notation useful in the algorithm. The algorithm is presented in section IV and its correctness and complexity is discussed in section V.

II. Computational Model

A notion of self-stabilizing algorithms on distributed systems was introduced by Dijkstra [4]. A survey in the topic can be found in the paper by Schneider [16], and further details in the book by Dolev [5]. The notions from graph theory not defined in this paper one can find in the book by Harary [10].

A distributed self-stabilizing system consists of a set of processes called computing nodes and communication links between them, which we can be modelled topologically by a graph. We assume that every node in the system runs the same algorithm and can change the state of the local variables. These variables determine the *local state* of a node. Moreover, nodes can observe the state of variables on themselves and their neighbor nodes. The state of all the nodes in the system determines the *global state*. In this paper we assume that every node has unique identificator (*id*).

Every self-stabilizing algorithm should have a class of global states defined, that are called the *legitimate states*, for which the system is stable, which means that no action should and can be done by the algorithm itself. Every other global state is called the *illegitimate* and for the algorithm to be correct there has to be some possibility to make a move in this kind of state. Every possible sequence of moves made by the algorithm must end up with the legitimate state. Indeed, it is the aim of every self-stabilizing algorithm to bring the system to the legitimate (desirable) state, either after some alteration (from the outside of the system) of variables in the nodes had been done or after the system had been started.

Generally, an algorithm consists of a set of rules. A rule has the form:

```
label: If guard
    then assignment instructions
    where definitions of objects.
```

A *guard* is a logic predicate which can refer to variables in the node itself and its neighbors. A *label* and a *where* clause are optional. We say that a rule is *active* if its guard is evaluated to true. A node is *active* if it contains any active rule. If there is no active node in the graph, we say that the system is *stabilized*. Let us note that in a stabilized system no move can be made. One of the required property of a self-stabilizing algorithm is to make a system stabilized if and only if its state is legitimate.

We assume that active rules are triggered in an arbitrary order.

III. Notation

The main difficulty in enhancing the algorithm for locating the center in a maximal outerplanar graph to the Cartesian product of a maximal outerplanar graph and K_2 is to distinguish nodes that are placed in the same layer or not. Once it is done, the basic algorithm for a maximal outerplanar graphs can be easily adapted.

Theorem 1. *For every Cartesian product $G = K_2 \square M$, where M is a maximal outerplanar graph, the graph induced by $C(G)$ is the same as the Cartesian product of K_2 and the graph induced by $C(M)$.*

Proof: Let $x \in V(G)$, then $ecc_G(x) = ecc_M(x) + 1$. Thus, the center of G is the same as the sum of the centers of both layers of graph G. ■

There are two types of neighbors (of any node $i \in V(G)$) in the graph (see Fig. 2). The first one consist of nodes in the same layer, which together form triangle faces of a maximal outerplanar graph. The second, say j, is a neighbor which belongs to the other layer, we call such two nodes i, j the *pairing* nodes.

We define the following predicate to determine whether any two neighbors are pairing nodes:

$$\text{pairing}(i,j) \;\Leftrightarrow\; j \in N(i) \wedge (N(i) \cap N(j)) = \emptyset.$$

We will use this predicate in the definition of $N_l(i)$, which we define as a set of neighbors from the same layer (maximal outerplanar graph):

$$N_l(i) = \{j | j \in N(i) \;\wedge \neg\text{pairing}(i,j)\}.$$

In the algorithm we will use the following notation:

$n(i)$ — a variable storing the set of neighbor nodes (from the same layer) for the node i. Note that $n(i)$ is a variable, whose value may be incorrect at the beginning of the algorithm run, whereas $N_l(i)$ is a set which is determinable by the node i only, based on the topology of the network by looking at connections of node i; it can be computed only by the node i. Thus the $n(i)$ variable is set to allow a neighbor to determine other neighbors of the node i.

$c(i,j)$ — a variable (stored in the node i) which stores the set of common neighbors for nodes i and j,

$e(i,j,k)$ — a variable storing (in the node i) the edge eccentricity for the edge $\{i,j\}$ and the side containing the common neighbor k (of the nodes i and j; $k = \emptyset$ for an empty side),

$opp(i,j,k)$ — a variable storing (in the node i) the representation of the side opposite to k (against the edge $\{i,j\}$),

$v(i)$ — a variable storing (in the node i) the eccentricity of the vertex i, note that it is not the *edge* eccentricity, i.e. in a legitimate state $v(i) = \max_k |e(i,j,k)|$ for any $j \in N(i)$, and $v(i) \geq 1$ for any node i.

$m(i,j,k)$ — a pair stored in the node i for inside the dual vertex $\{i,j,k\}$. After stabilization, its first element is the eccentricity of the center nodes. The second element of the pair is the direction, that the information about the eccentricity of the center comes from. If for the dual vertex $\{i,j,k\}$ the information comes from the region incident to the

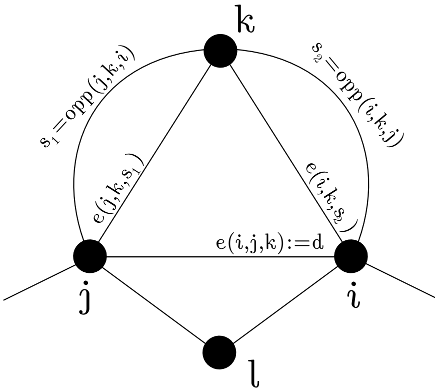


Fig. 3. Visualization of the rule 4.

edge $\{i, j\}$, then the direction is equal to $opp(i, j, k)$. In the case the information about the center eccentricity comes originally from the dual vertex $\{i, j, k\}$, then we set the direction to $\emptyset$.

IV. The Algorithm

In this section we present the rules of our algorithm (see Fig. 4). The first rule assigns the set of all neighbors of the node i in the same layer to its variable $n(i)$. Thanks to this, a node can know neighbors (in the same layer) of its neigbor, which is exploited in further rules. The second rule assigns in a node i the set of common neighbors in the same layer with node k: $c(i, k)$.

The rule 3a assigns outside edge eccentricity and sides of an edge. Rules 3b and 4 compute edge eccentricities according to [7] and both of them set $v(i)$ to proper value (see Fig 3).

The rule 5 propagates the minimum eccentricity through all the graph. The idea of the inside dual tree is used [8]. The information about the minimum eccentricity is propagated through the dual tree.

Note that in the rule 5 we used the function $MinEcc(i, j, k)$, returning the value of $m(i, j, k)$. The $MinEcc$ function is defined as follows (see Fig. 5):

Function $MinEcc(i,j,k)$

```
1  $v := v(i)$
2  $dir := \emptyset$
3  if $fst(m(k,i,j)) < v \wedge snd(m(k,i,j)) \in \{opp(k,j,i), \emptyset\}$ then
4      $(v, dir) := m(k,i,j)$
5  end if
6  if $fst(m(j,i,k)) < v \wedge snd(m(j,i,k)) \in \{opp(j,k,i),\ \emptyset\ \}$ then
7      $(v, dir) := m(j,i,k)$
8  end if
9  if $fst(m(i,j,opp(i,j,k))) < v \wedge snd(m(i,j,opp(i,j,k))) \neq k$ then
10     $v := fst(m(i,j,opp(i,j,k)))$
11     $dir := opp(i,j,k)$
12 end if
13 if $fst(m(i,k,opp(i,k,j)) < v \wedge snd(m(i,k,opp(i,k,j))) \neq j$ then
14     $v := fst(m(i,k,opp(i,k,j))$
15     $dir := opp(i,k,j)$
16 end if
17 return $(v, dir)$
```

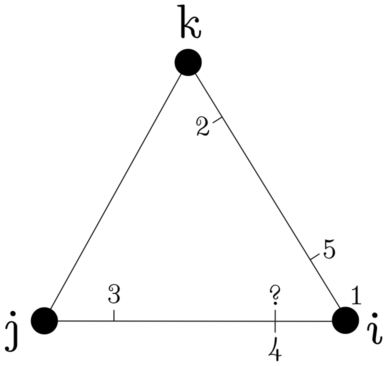


Fig. 5. The visualization of computation of the function $MinEcc(i, j, k)$. The numbers stand for the order of checking (and assigning if necessary) values of $m(\cdot, \cdot, \cdot)$. The order above is: 1. $(v(i), \emptyset)$ (lines 1–2 of the $MinEcc$ function), 2. $m(k, i, j)$ (lines 3–5), 3. $m(j, i, k)$ (lines 6–8), 4. $m(i, j, opp(i, j, k))$ (lines 9–12), 5. $m(i, k, opp(i, k, j))$ (lines 13–16). The question mark stands for $m(i, j, k)$, which is the computed value.

Two projection functions are used in the above function: $fst((a, b)) \overset{\text{def}}{=} a$ and $snd((a, b)) \overset{\text{def}}{=} b$, which take the first and second element of the pair, respectively.

The $MinEcc$ function computes the minimum value over eccentricities and the direction that it comes from for the triangle region specified by three parameters i, j, k (see Fig. 5). The first step is to consider the node i itself as a candidate with the minimum value of the eccentricity available in the neighborhood. In this case the direction would

```
1: If $n(i) \neq N_l(i)$
     then $n(i) := N_l(i)$
2: If $\exists_{k \in n(i)} c(i,k) \neq n(i) \cap n(k)$
     then $c(i,k) := n(i) \cap n(k)$
3a: If $\exists_{j \in n(i)} (|c(i,j)| = 1 \wedge (e(i,j,\emptyset) \neq -1 \vee opp(i,j,\emptyset) \neq k \vee opp(i,j,k) \neq \emptyset))$
     then $e(i,j,\emptyset) := -1$
     $opp(i,j,\emptyset) := k$
     $opp(i,j,k) := \emptyset$
     where $\{k\} = c(i,j)$
3b: If $\exists_{j \in n(i)} (|c(i,j)| = 1 \wedge (e(i,j,k) \neq d \vee v(i) \neq \max(|e(i,j,k)|, 1))$
     then $e(i,j,k) := d$
     $v(i) := \max(|e(i,j,k)|, 1)$ where
     $q = \begin{cases} -(1 + e(j,k,opp(j,k,i))) & \text{if } e(j,k,opp(j,k,i)) > 0, \\ e(j,k,opp(j,k,i)) & \text{otherwise} \end{cases}$
     $d = \begin{cases} |e(i,k,opp(i,k,j))| & \text{if } |e(i,k,opp(i,k,j))| \geq |q|, \\ q & \text{otherwise} \end{cases}$
     $\{k\} = c(i,j)$
4: If $\exists_{j \in n(i)} (|c(i,j)| = 2 \wedge \exists_{k \in c(i,j)} (e(i,j,k) \neq d \vee opp(i,j,k) \neq l \vee opp(i,j,l) \neq k \vee v(i) \neq \max(|e(i,j,k)|, |e(i,j,l)|)))$
     then $e(i,j,k) := d$
     $opp(i,j,k) := l$
     $opp(i,j,l) := k$
     $v(i) := \max(|e(i,j,k)|, |e(i,j,l)|)$
     where
     $q = \begin{cases} -(1 + e(j,k,opp(j,k,i))) & \text{if } e(j,k,opp(j,k,i)) > 0, \\ e(j,k,opp(j,k,i)) & \text{otherwise} \end{cases}$
     $d = \begin{cases} |e(i,k,opp(i,k,j))| & \text{if } |e(i,k,opp(i,k,j))| \geq |q|, \\ q & \text{otherwise} \end{cases}$
     $\{k,l\} = c(i,j)$
5: If $\exists_{j,k \in n(i)} (k \in c(i,j) \wedge m(i,j,k) \neq MinEcc(i,j,k))$
     then $m(i,j,k) := MinEcc(i,j,k)$
```

Fig. 4. The self-stabilizing algorithm for finding the center in $K_2 \square$ maximal outerplanar graph.

be $\emptyset$ as it does not come from other region. In the second and third step, the node i checks the neighbor nodes k and j as a candidates for the minimum value. If the values in the nodes k or j come from regions incident to the edges $\{i,k\}$ or $\{i,j\}$, they are not trusted. It is to ensure that no wrong value can last in the region infinitely long time (number of moves). And last two steps, the values from two neighbor regions (incident to i) are checked.

V. Correctness and complexity

Now we prove some properties of the algorithm. We assume that n is the number of nodes in a layer (a maximal outerplanar graph) of the graph.

Lemma 1. *Algorithm consisting of rules 1–4 stabilizes in $\mathcal{O}(n^2)$ number of moves.*

Proof: The stabilization of the rule 1 is obvious as the guard does not depend on variables in neighbor nodes. So the rule 1 gets inactive in finite time. The rule 2, depends only on static (after stabilizing of the rule 1) information computed by the rule 1. Hence it stabilizes in limited by a constant number of moves per node, as rule 1 does.

The same applies to rule 3a, as it depends on variable values computed by two former rules, because it is for an outerface edge (i.e. the edge belonging to Hamilton cycle), which is an initial case of the recursive classical algorithm. Once the $c(i, j)$ is properly computed in the node i, it never changes. Thus if any of the variables $e(i, j, \emptyset)$, $opp(i, j, \emptyset)$ or $opp(i, j, k)$ is in a wrong state, then all are correctly computed and also never change.

Now all the nodes have got rules 3a and 3b inactive. Then we consider the rule 4. Note that this rule is applicable only for graphs bigger than a triangle. Suppose there are two adjacent edges lying on an outerface of the graph. There has to be the third edge, which is also adjacent to them, and the edge eccentricities of this edge stabilize with rule 4. This is the first layer of proper rule 4 computation. Each next layer of proper computation of rule 4 depends on a previous layer. We have a finite graph, so the rule 4 stabilizes. As each layer of computation of rule 4 takes $\mathcal{O}(n)$ moves and there are $\mathcal{O}(n)$ layers, it all takes $\mathcal{O}(n^2)$ moves. The last layer of computation stabilizes by rule 3b, as it reaches the another outerface of the graph. ■

The following lemma describes the situation after stabilization of rules 1–4.

Lemma 2. *If the phase 1–4 has stabilized a system, then phase 5 will stabilize in $\mathcal{O}(n^2)$ number of moves.*

Proof: The pessimistic case would be if every dual vertex had wrong value for variable $v(i)$ — for example as a result of start up of the system — and having also wrong values of $m(\cdot, \cdot, \cdot)$ for every dual vertex (representing a region of the layer, i.e. maximal outerplanar graph). Let us assume that every $v(i)$ is less than the proper minimum eccentricity and there are no nodes i, j such that $v(i) = v(j)$. Then the pessimistic order of propagation of the $m(\cdot, \cdot, \cdot)$ values would be when the value $v(i)$ spreads the first (for some i) which is the biggest among all the other $v(k)$ (for all nodes k except i) but still it is less then the proper minimum eccentricity.

The above propagation takes $\mathcal{O}(n)$ moves. Note that now the dual tree is filled with improper value of some $v(i)$. But there are $n - 1$ wrong candidates of the minimum eccentricity to spread left. Once again, the pessimistic case would be when the next value to propagate was the maximum among all the candidates, which is less then the spread in the tree.

Each of these phases takes $\mathcal{O}(n)$ moves and there are $\mathcal{O}(n)$ phases, so all pases of rule 5 run in $\mathcal{O}(n^2)$ moves. ■

Now we can formulate the following theorem.

Theorem 2. *The algorithm takes $\mathcal{O}(n^4)$ number of moves to stabilize.*

Proof: The computation in each of layers is independent, so by Lemmas 1 and 2 we get the result. ■

VI. Conclusions

In this paper we proposed a self-stabilizing algorithm for finding the center of the Cartesian product of graph K_2 and a maximal outerplanar graph. We hope the method similar to the presented here can be applied to other classes of graphs. The open question is if there exists an algorithm with better complexity — number of moves bringing a system to the legitimate state.

References

[1] Bielak, H., M. Pańczyk, M.: "A self-stabilizing algorithm for finding weighted centroid in trees"; Annales UMCS Informatica, AI XII, 2 (2012), 27–37; http://dx.doi.org/10.2478/v10065-012-0035-x.

[2] Bruell, S. C., Ghosh, S., Karaata, M. H., Pemmaraju, V.: "Self-stabilizing algorithms for finding centers and medians of trees"; SIAM Journal on Computing, 29 (1999), 600–614; http://dx.doi.org/10.1145/197917.198130.

[3] Chepoi, V., Fevat, T., Godard, E.,Vaxès, Y.: "A self-stabilizing algorithm for the median problem in partial rectangular grids and their relatives"; Algorithmica, 62 (2012), 146–168; http://dx.doi.org/10.1007/s00453-010-9447-4.

[4] Dijkstra, E. W.: "Self-stabilizing in spite of distributed control"; Communications of the ACM, 17 (1974), 643–644; http://dx.doi.org/10.1145/361179.361202.

[5] Dolev, S.: "Self-stabilization"; The MIT Press (2000).

[6] Farley, A. M.: "Vertex centers of trees"; Transportation Science, 16 (1982), 265–280; http://dx.doi.org/10.1287/trsc.16.3.265.

[7] Farley, A. M., Proskurowski, A.: "Computation of the center and diameter of outerplanar graphs"; Discrete Applied Mathematics, 2 (1980), 185–191; http://dx.doi.org/10.1016/0166-218x(80)90039-6.

[8] Fleischner, H. J., Geller, D. P., Harary, F.: "Outerplanar graphs and weak duals"; Journal of the Indian Mathematical Society, 38 (1974), 215–219.

[9] Goldman, A. J.: "Minimax location of a facility in a network"; Transportation Science, 6 (1972), 407–418; http://dx.doi.org/10.1287/trsc.6.4.407.

[10] Harary, F.: "Graph Theory"; Addison-Wesley, 1972.

[11] Hedetniemi, S. M., Cockayne, E. J., Hedetniemi, S. T.: "Linear algorithms for finding the jordan center and path center of a tree"; Transportation Science, 15 (1981), 98–114; http://dx.doi.org/10.1287/trsc.15.2.98.

[12] Kariv, O., Hakimi, S. L.: "An algorithmic approach to network location problems. I: The p-centers"; SIAM J. Applied Mathematics, 37 (1979), 513–538; http://dx.doi.org/10.1137/0137040.

[13] Korach, E., Rotem, D., Santoro, N.: "Distributed algorithms for finding centers and medians in networks"; ACM Transactions on Programming Languages and Systems, 6 (1984), 380–401; http://dx.doi.org/10.1145/579.585.

[14] Laskar, R., Shier, D.: "On powers and centers of chordal graphs"; Discrete Applied Mathematics, 6 (1983), 139–147; http://dx.doi.org/10.1016/0166-218x(83)90068-9.

[15] Rosenthal, A., Pino, J.: "A generalized algorithm for centrality problems on trees"; JACM, 36 (1989), 349–361; http://dx.doi.org/10.1145/62044.62051.

[16] Schneider, M.: "Self-stabilization"; ACM Computing Surveys, 25, 1, (1993); http://dx.doi.org/10.1145/151254.151256.

DOI: 10.15439/2014F465
ACSIS, Vol. 3

Identity Providers-as-a-Service built as Cloud-of-Clouds: challenges and opportunities

Diego Kreutz
LaSIGE/FCUL, Lisbon, Portugal
Email: kreutz@ieee.org

Eduardo Feitosa
IComp/UFAM, Manaus, Brazil
Email: efeitosa@icomp.ufam.edu.br

***Abstract*—In our previous work we designed and evaluated the feasibility of highly secure and dependable identity providers (IdPs) for the increasing requirements of future IT infrastructures. In this position paper we extend our previous work by analyzing and discussing the benefits of deploying highly secure and dependable identity providers-as-a-service (IdP-as-a-Service), without compromising the confidentiality of sensitive data and operations. In order to achieve this goal, we discuss some of the forefront challenges of deploying IdP-as-a-Service as a cloud-of-clouds model to ensure important properties such as the resistance against different types of threats and attacks, arbitrary faults, and make it more realistic to improve the system availability up to the three-nines mark. Notwithstanding, the main opportunities towards IdP-as-a-Service are also analyzed. We finish the paper proposing a sustainable business model based on our previous deployments and results, showing that it can be a win-win opportunity, i.e., both IdP-as-a-Service providers and customers can benefit from it.**

Keywords*—identity providers, IdP-as-a-Service, business model and opportunities, security, dependability, high availability, cloud providers, multi-cloud, telco cloud, hybrid cloud.

I. Introduction

FUTURE IT infrastructures will further combine and foster the interoperability of several computing, storage and networking technologies, such as those driven by new concepts and architectures like software-defined networks (SDN), software-defined storage (SDS), software-defined computing (SDC) and software-defined management (SDM), orchestrated by the control elements of software-defined environments (SDE) [1], [2]. In other words, on-demand provisioning will become the rule rather than the exception in all layers of current and future IT infrastructures. In particular, SDN was one of the key missing pieces to complete the SDE puzzle, enabling on-demand provisioning of network resources [1], [3].

In this new software-defined world (SDW), comprised by a myriad of advanced technologies and world-wide interoperability/integrations (e.g., cross-domain authentication, federations of authentication and authorization infrastructures, and so forth), enterprises (from small to large) can benefit from the on-demand service and IaaS provisioning offered by flexible and dynamic software-driven IT architectures and infrastructures. Different critical sectors, traditionally more resistant to changes, have been investing on cloud-based (outsourced) IT infrastructures and services, such as oil and gas industry and banks [4], [5].

Yet, some of the major challenges are related to the need of providing and ensuring higher degrees of security and dependability on different types of systems [6], [7], [8], [9]. Strictly speaking, it is essential to have in mind that the dynamic provisioning of IT services for future environments will have to consider different crucial aspects, such as performance, high availability, confidentiality, privacy, fault tolerance, and automated security, also from a conceptual and design perspective and not only as optional bolted-on features appended to the systems in an ad-hoc mode when it is already too late, i.e., data leakage or intrusions have already happened [6], [10], [11]. In fact, security and dependability of essential services is becoming a first class concern for enterprises that depend on systems connected to the Internet. One of the reasons is that the number and criticality of security threats have been rising at the same time that attacks are getting more sophisticated and challenging to deal with. Advanced persistent threats, large scale distributed denial of service and data leakage are becoming more frequent and dangerous to the enterprise business, governments and nearly all sorts of institutions [12], [13], [14], [15], [16].

Identity providers (e.g., OpenID providers) are not an exception. Recent research indicates that there is a significant gap on the provisioning of highly secure and dependable IdPs, representing one of the top concerns for future IT infrastructures [17], [18], [19]. Therefore, to best of our knowledge, we are proposing the first IdP-as-a-Service based on a cloud-of-clouds model for deploying highly secure and dependable IdPs. This can be achieved by combining different advanced techniques from distributed systems, dependability and security. Furthermore, we have already shown how it is possible to take advantage of multi-infrastructures (e.g., data centers) to increase the robustness of the system for tolerating different types of faults and attacks [18], [19], [20].

An OpenID-as-a-service has been proposed before [21]. However, it fails at addressing several security and dependability issues of current identity providers, as we have further depicted in our previous work [18], [19], [20]. Moreover, it uses only a single cloud, based on OpenStack, to scale the OpenID service, which is susceptible to several threats and performance issues [7], [8], [22].

Our main contribution is a cloud-of-clouds IdP model for achieving the technical and financial requirements of future IT infrastructures. In other words, the model has to be capable of taking advantage of the benefits provided by diverse infrastructures and still being cost effective, i.e., represent an interesting business opportunity for both providers and customers. Therefore, in this position paper we discuss the five essential challenges for deploying IdP-as-a-Service on a cloud-of-clouds model. First, deployment and operation challenges, advantages and disadvantages on different scenarios, such as collocation, private cloud, public cloud, and telco cloud are discussed. Second, we analyze the main trade-offs of different deployment alternatives. Third, we introduce the challenges and potential solutions for ensuring the confidentiality and privacy of sensitive data in a cloud-of-clouds environment. Forth, we dissect the main challenges of small and medium enterprise on building highly secure and dependable systems for future IT infrastructures. Finally, we provide a step-by-step cost analysis, based on our previous deployments and a real IT infrastructure use case, of the IdP-as-a-Service model and the new opportunities for both providers and customers.

II. Problem Statement

IdPs are arguably important services of current and future IT infrastructures. However, existing solutions have several vulnerabilities and weaknesses, being incapable of ensuring high levels of security and dependability required by the new and dynamic world of enterprises that day-after-day depend more heavily on IT systems. For instance, currently available and deployed identity providers can be threatened by different advanced persistent threats, large scale DDoS attacks, security breaches caused by software or infrastructure vulnerabilities, and so forth [12], [13], [15], [16], [17].

RADIUS and OpenID are examples of services with different weakness regarding security and dependability [17], [18], [19], [20], as summarized in Table I. Current implementations and deployments are highly susceptible to: (i) common vulnerabilities in different parts of the IT stack; (ii) sensitive data leakage due to the fact that keys and certificates are commonly stored in the operating system's file system; and (iii) resource depletion attacks when deployed on the same physical server with current virtualization technologies (e.g., Xen hypervisor).

Table I. Vulnerabilities and properties.

Vulnerability/Support	RADIUS	OpenID
Tolerates crash faults (using back-end clusters)	Yes	Yes
Tolerates arbitrary faults	No	No
Tolerates infrastructure outages	No	No
Tolerates DDoS attacks	No	No
Risk of common vulnerabilities	High	High
Risk of sensitive data leakage	High	High
Diverse security-related vulnerabilities	Yes	Yes
Susceptible to resource depletion attacks	Yes	Yes

Another issue worth mentioning is the fact that those services are commonly deployed in a single physical infrastructure (e.g., single physical machine or multiple servers in a single data center). This is still a wide spread practice in most enterprise environments, with some forefront exceptions such as cloud providers, which have several data centers in different locations. Notwithstanding, there is also a very high inherent complexity in deploying and managing current networks and services. This fact has been one of the main propeller of initiatives towards outsourcing of middleboxes [23], routing control logic [24], among other network functionality [25], [26].

In an interconnected and interdependent world, an energy outage in one IdP could impact many uses in distinct locations. This is the case of *eduroam* [27], where an energy outage in one university can simply deny the access of thousands of users to resources geographically dispersed. Therefore, it is of paramount importance to design IdPs that can be easily deployed across multiple physical infrastructures, such as different data centers or clouds.

This scenario is becoming even more critical once identity providers are being deployed for ensuring the security (e.g., access control) of large-scale distributed virtual networks, controlling not only users but also virtual routers and virtual machines [28]. Therefore, their security and dependability is a crucial issue for ensuring different security, availability and safety properties of future IT infrastructures [6], [29].

To tackle with some of those issues regarding security and dependability of critical services, we have proposed and evaluated the feasibility of resilient and trustworthy IdP services [18], [19], [20], [30]. While these system designs and implementations solve different of the afore mentioned issues of OpenID and RADIUS-like services, they pose also additional challenges for small and medium enterprise. Highly secure and dependable services are not simple and easy to develop, deploy and operate because they require different high skilled professionals (e.g., security specialists, distributed systems experts, system operation specialists, and highly skilled network operators) that are costly to afford and maintain, in particular for enterprise that do not have IT as their main business goal. Notwithstanding, both the IT infrastructure and human resources are much more expensive for small and medium enterprise when compared to larger companies [31].

At a first glance, one could think of outsourcing identity providers to one cloud provider. But, there are also several limitations of cloud providers that can have a significant impact on the system security and reliability, such as privacy, confidentiality, product/cloud provider lock-in, possibility of data loss, application interfaces and interoperability, geographical location of data, higher levels of failure when compared to traditional in-house machines, and so forth [7], [8], [22], [32].

Therefore, in this position paper we argue for secure and dependable IdP-as-a-Service using a cloud-of-clouds model. For different applications, such as storage, database systems, and experimentation testbeds, it has already been shown that multi-cloud systems can help to free the consumer from technical and business failures that any single cloud provider might be experiencing, prevent vendor lock-in, increase security and dependability, and reduce overall service cost by choosing the most cost-effective cloud providers [33], [34], [35], [36].

III. Towards Secure and Dependable IdP-as-a-Service

A. A Resilient and Secure IdP Architecture

Previously, we depicted a functional model, system design artifacts and essential techniques required for developing and deploying more secure and dependable identity providers for future IT infrastructures [18], [19], [20]. Furthermore, we analyzed some of the main trade offs of deploying robust and resilient services on a single physical machine or on multiple physical infrastructures. Whereas the performance can be boosted when deploying the system on a single physical infrastructure (e.g., data center), a deployment on multiple physical infrastructures (e.g., three different data centers) can significantly augment the system's resistance against physical and logical failures, first class DDoS attacks and resource depletion attacks, i.e., improve in orders of magnitude the system robustness and overall availability.

Figure 1 illustrates our proposed architecture with replicated components for higher availability. The IdP service replicas are capable of tolerating arbitrary faults, i.e., any accidental or malicious faults such as those caused by different types of attacks. Compared with a tradicional identify provider architecture (e.g., OpenID), two new components, IdP gateways and secure elements, are employed to safeguard the authentication systems without compromising backward compatibility. Further details regarding the system elements, protocols and technologies can be found in our previous work [18], [19], [20].

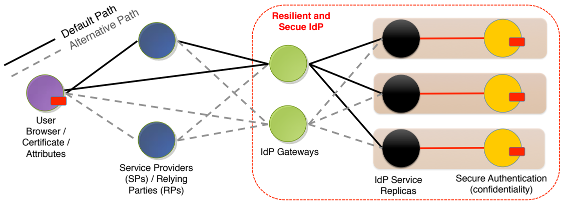

Default Path
Alternative Path
User Browser / Certificate / Attributes
Service Providers (SPs) / Relying Parties (RPs)
Resilient and Secue IdP
IdP Gateways
IdP Service Replicas
Secure Authentication (confidentiality)

Figure 1. Resilient and secure IdPs (based on the OpenID 2.0 model).

Figure 2 illustrates a deployment of our prototype in a multi-cloud enviroment. In this case, there are five clouds, being three public clouds (IdP replicas: IdP-R1, IdP-R2, IdP-R3), one private cloud (gateway GW1) and one colocation (gateway GW2). Additionally, there are also two service providers (SP1/RP1 and SP2/RP2), each of them running on a different public clouds, and end users using the IdP service to access the resources provided by SP1 and SP2, respectively.

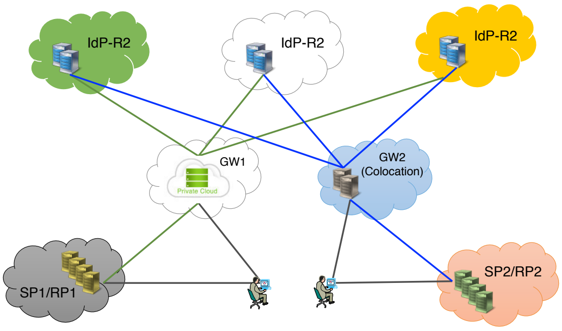


Figure 2. IdP-as-a-Service build as a cloud-of-clouds (or hybrid infra).

This deployment example is only to give an idea of the possibilites realized through our proposed architecture. In fact, we have already experienced its deployment across multiple data centers geographically distributed.

B. *Multi-environment Deployment Trade Offs*

Our resilient IdP architecture is designed to be deployable on diverse environments. A deployment in a single data center can achieve a high throughput (e.g., 2,200 authentications/s), while a multi-data center deployment can impose a higher communication latency, reducing the number of authentications/s [19], [20]. Yet, a multi-data center deployment can ensure higher levels of availability and provide protection mechanisms for resisting against different kinds of resource exhaustion attacks, large scale DDoS, and so forth [19], [20]. Therefore, decide between one or another deployment depends on the specific requirements of the enterprise, i.e., the IdP customer.

Figure 3 advances a step further our previous research and analysis [19] by extending the trade offs to multi-data center and multi-cloud environments. Here we only consider deployments targeting high availability and resistance against multiple types of threats and attacks. As we previously analyzed in details and numbers, the performance tends to degrade with the geographic distribution of machines due to (mainly) the increase in the network latency [20]. Different data centers of a single cloud provider can perform better than a multi-cloud setup if the provider has dedicated and efficient network links between the data centers. Otherwise, the performance of a multi-data center deployment will be similar to a multi-cloud environment. In this position paper, we assume only public networks in a multi-cloud configuration, which is still the most common case. Nevertheless, we are also aware that this scenario is likely going to change in a near future due to the advancements being fostered by SDN, which have already reached the backbone of telco companies such as NTT.

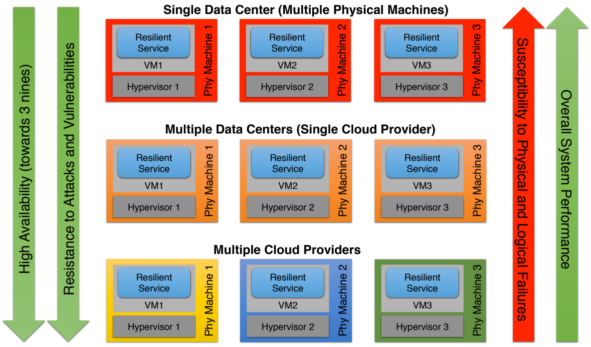

High Availability (towards 3 nines)
Resistance to Attacks and Vulnerabilities
Single Data Center (Multiple Physical Machines)
Resilient Service
VM1
Hypervisor 1
Phy Machine 1
Resilient Service
VM2
Hypervisor 2
Phy Machine 2
Resilient Service
VM3
Hypervisor 3
Phy Machine 3
Multiple Data Centers (Single Cloud Provider)
Multiple Cloud Providers
Susceptibility to Physical and Logical Failures
Overall System Performance

Figure 3. Diverse physical infrastructures: deployment trade-offs.

The susceptibility to physical and logical failures reduces with the increasing diversity of physical infrastructures. While one single data center can suffer an energy or communication blackout, it is much less likely to affect multiple data centers or cloud providers at the same time. Additionally, multi-environment configurations also increase the system availability. Moreover, multi-infrastructure deployments can give a considerable push towards achieving the "three-nines" mark of system availability [37].

Multi-infrastructure deployments also augment the system resistance against different types of attacks and vulnerabilities. Large scale DDoS attacks and common vulnerabilities (e.g., in security mechanisms, hypervisors, operating systems, physical network, physical servers, etc.) have less chance of affecting the system in a multi-cloud deployment. Each cloud provider has its own systems and protection mechanisms, making the task of any attacker much harder. In fact, some cloud providers have already shown how it is possible to protect cloud services against large scale DDoS by taking advantage of the huge amount of resources offered by different data centers and the right defense mechanisms, such as anycast methods [14], [38].

C. *Where to deploy the system elements?*

Should we use colocation, private cloud, public cloud, hybrid cloud or telco cloud? What are the most common *advantages and disadvantages of each choice?* There is no simple answer to these questions. Each environment is capable of providing different properties and benefits. Therefore, choose which one is the most adequate to deploy your IdP-as-a-Service is highly dependent on the specific requirements of the enterprise. For instance, while it is hard to ensure data confidentiality and privacy against a malicious sysadmin of a public cloud provider, to build a private cloud or rent a space in a data center for your own physical machines (colocation) will enable higher levels of protection regarding the confidentiality and privacy of sensitive data. But, at the same time, private clouds and colocation can also increase CAPEX and OPEX [39]. Differently, public clouds are less risky and more elastic in terms of scaling and costs.

There are IT companies specialized in offering colocation-as-a-service (CaaS), with data centers geographically dispersed, i.e., customers can rent spaces in different locations to increase their system availability. However, the main drawbacks of this approach is that you still need to manage your own infrastructure, as well as have a plan for scaling, which can be tricky and susceptible to the market reactions to the enterprise's products and services.

Cloud providers (e.g., Figure 4) are a much more flexible (and impose less risks) alternative to most enterprises. One can simple start with a single computing unit and dynamically scale up (on-demand) to thousands of nodes. Consequently, this reduces risks, CAPEX and OPEX. In the worst case, if the business fails, it is as simple as to scale down to zero resources. On the other hand, it would be harder get rid of the infrastructure if using colocation or building your own private cloud. Independently of the business success or not, you would still have the IT infrastructure. Eventually, if the business fails, sell it to a cloud provider at a relatively low price compared to the CAPEX and OPEX spent in the IT infrastructure.

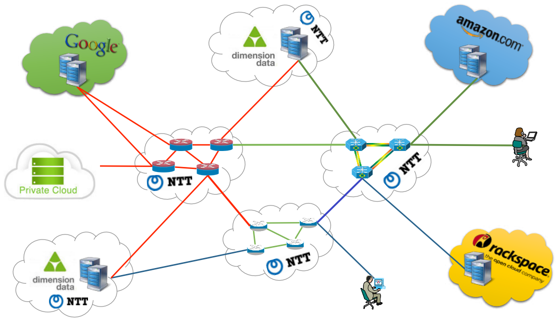


Figure 4. IaaS: private cloud, public cloud, hybrid cloud or telco cloud?

Typical cloud providers, such as Google, Amazon, and RackSpace, allow customers to choose between public and private clouds. As an example, Amazon have special contracts for companies willing to have their own private cloud, with higher security guarantees (e.g., confidentiality, privacy, and so forth). In spite of the fact that you are not the owner of the private cloud, you have contractual guarantees, i.e., the contract is the critical piece of the puzzle. Alternatively, it is also possible to build your own private cloud(s). Going a bit further, it can be interesting to consider a hybrid model, i.e., use your (potentially small) private cloud for critical and sensitive data and operations, while the less sensitive data and services of the business can be deployed on a public cloud. The main advantages of private clouds are with respect to security properties such as confidentiality and privacy, which are a bit harder to achieve or guarantee in public clouds.

Interestingly, another type of cloud model is rising, the telco cloud [40]. For a long time, telecom companies have not realized the opportunity to become big IT players (e.g., cloud providers). A typical telecom company owns large network infrastructures and has data centers in different locations. Consequently, these companies have already the main building blocks required to become cloud providers. More than that, Telcos have the network in their hands. By interconnecting their data centers through dedicated high speed links and offering on-demand network provisioning for their customers, these companies have the unique opportunity to become big cloud provider players in this competitive market. No other cloud provider (with eventual few exceptions) is capable of providing flexible, scalable and dynamic network services (inter-data centers) as telecom companies are. In fact, some global Telcos like NTT have already realized this opportunity and jumped into the cloud provider market by offering both computing and network resources provisioning as you go. For instance, NTT has more than 140 data centers spread across the world and interconnected by their own network infrastructure, providing global virtualization services [41].

Telco clouds have a clear advantage on the provisioning of network-as-a-service (NaaS), network-level security and QoS at the network level. This is particularly true between geographically dispersed data centers or cloud infrastructures. Most of the other cloud providers cannot afford to have their own network infrastructure between data centers spread across the globe. Therefore, they are tied to the contracts, services, prices and transport technologies of carrier network providers. Yet, inside a single data center, all cloud providers have the same capabilities of providing an infrastructure capable of offering NaaS and ensuring security and QoS properties.

D. Confidentiality and Privacy: more challenges ahead

Ensure the confidentiality and privacy of sensitive data and operations in public clouds is one of the major challenges for deploying IdP-as-a-Service. Therefore, multi-cloud deployments require new methods and techniques for ensuring the confidentiality of the most critical parts of the system.

In traditional deployments, server keys, session keys, authentication assertions, and so forth, can be easily compromised with access to the authentication servers. Therefore, we proposed in our previous work to split these services in two parts. The first is comprised by the traditional service/protocol itself, while the second is a special purpose component for isolating the essential operations and data required to authenticate and authorize clients. This isolation can be achieved through secure elements of different types, depending mainly on the specific requirements of the target environment [18], [19], [20]. The secure elements, proposed in our previous work, are specialized components that significantly enhance the security (in particular confidentiality) of the system without sacrificing system properties such as high availability and performance. To this purpose, we have designed a simple and common interface which can be easily implemented in different hardware or software platforms, such as grids of smart cards, tamper resistant FPGAs, virtual TPMs and highly secure hypervisors [30]. However, secure components are not quite suitable for all kinds of deployments in public or telco clouds. For instance, if you do not completely trust the cloud provider (even with a well-defined and strict contract), then you need alternative methods for protecting the confidentiality and privacy of your sensitive data and operations.

Some potential alternatives are tough available. However, each of them has certain requirements that can lead to low system performance, changes on the traditional protocol specifications or careful deployment and modification of some essential elements of our previously proposed architecture, such as the gateway. Let's briefly introduce and discuss a few possibilities and their implications, such as cryptographic cloud storage [42], shared cloud-backed file systems [35], encrypted database systems [43], [44], and secret sharing schemes [45], [46].

Cryptographic cloud storage [42] and shared cloud-backed file systems [35] can be used to provide integrity and confidentiality properties between users and the backend authentication servers. Those services offer secure and reliable storage functionality by ensuring properties such as privacy and confidentiality through multi-cloud environments and cryptographic schemes. However, to use them as an anchor of trust in an IdP-as-a-Service infrastructure, we would face at least three problems. First, shared and cryptographic storage protect only the data, but not the operations (e.g., sensitive cryptographic operations). Second, while on the client side they are simple to use (because they are designed to be Dropbox-like solutions), it is more tricky to adapt those solutions on the authentication backend components. In particular, the authentication protocols (e.g., RADIUS, OpenID) would have to be adapted. Third, these solutions require special purpose component (e.g., data processor) on the client side. This component is responsible for encrypting data (and ensuring a fair data distribution) before sending it to the multi-cloud storage system.

Encrypted databases [43], [44] can be considered as a second alternative. However, in order to effectively use an encrypted database on the backend authentication server for ensuring data and operations confidentiality and privacy, it would be necessary to change both clients and relying parties of OpenID-based architectures. Furthermore, the authentication protocol would have to be changed as well to accommodate the new authentication operations based on encrypted data.

Alternatively, secret sharing algorithms [45], [46] can be used to ensure confidentiality of sensitive data. Differently from the other two approaches, secret sharing does not impose backward compatibility problems. It poses only two challenges. First, the gateway element has to be modified in order to "join the shares" of the sensitive data (e.g., message signatures) from the different IdP service replicas. As secret sharing guarantees that with only t (threshold) shares one is able to build a valid piece of data, this would restrict the activity of any malicious cloud provider or attacker in possession of less than t shares of the secret. The second issue is that the gateway becomes now a critical element of the architecture because is assembles valid messages based on t shares received from different authentication replicas. Additionally, the gateway may also be the "dealer" of the secret sharing algorithm, i.e., the element responsible for distributing the shares to each replica or secure element of the system. Therefore, the gateway has to be protected, i.e., cannot be deployed on public or telco clouds. It should be deployed on private clouds or in collocation mode to ensure that the client (e.g., enterprise company A) is the only one with access and control over the gateway elements. This characterizes a hybrid cloud scenario.

E. Main challenges for small and medium enterprises

Secure and resilient IdPs are capable of supporting different types of threats, such as advanced persistent threats, large scale DDoS, physical and logical disruption, and ensuring critical security properties (e.g., confidentiality) [18], [19], [20]. However, to develop and deploy such kind of systems is not a simple or easy task. It requires different specialized skills in security, distributed systems, networking, systems operation, and so forth. Yet, most of small and medium enterprises are not able (or willing) to afford the cost of highly specialized IT teams to support those kind of advanced systems. In general, only large IT companies such as Google, Amazon, Rackspace, Microsoft, among others, can afford to spend considerable sums of money on highly skilled teams. Moreover, as it has already been shown, the overall infrastructure (e.g., CPU, storage, etc.) and human resources costs reduces significantly with the size of the enterprise [31]. For instance, in large scale IT providers the server admin ratio is 800 to 1k, while in small enterprises this ratio is approximately eight times smaller. This means that small enterprises have a nearly 8x higher OPEX cost considering only human resources.

Furthermore, we also need to add other variables, which add extra complexity to the provisioning of highly secure and dependable services across multiple infrastructures. Examples include the variation in cost models from provider to provider, unclear contracts (i.e., do not specify all the details and tools available to the customer), significant performance variation and quality of service guarantees between different cloud providers, diversity of technical tools and resources for deploying systems across multiple infrastructures, different levels of failures between distinct providers, and so forth [31], [47], [48]. In other words, there are too many risks and costs in building and owning their own secure and dependable identity providers for small and medium enterprises.

On the other hand, large IT providers already have large and globally spread physical infrastructures and highly skilled IT teams in various areas, such as operating systems, distributed systems, security, database systems, and so forth. Therefore, for them it is not costly or risky to provide new kinds of systems-as-a-service. On the contrary, this can add more revenue opportunities to their portfolio of products and services.

IV. Secure and Dependable IdP-as-a-Service: a win-win opportunity ahead

In this section we start by summarizing the first experimental results regarding to the performance of our prototype implementation, a multi-environment deployable secure and dependable identity provider. After that, we discuss the scaling capacity of the system based on data analysis of a real enterprise environment. Following, we analyze, discuss and propose secure and dependable IdP-as-a-Service as a viable and interesting win-win opportunity for cloud providers, new IT startups and customer, i.e., normal enterprises.

A. First experiments and results

Table II summarizes the first results of deploying our resilient and secure IdP prototype on three distinct computing environments, one single physical machine (UFAM-VMs), one single data center with multiple computing instances (Amazon-EC2, using `m3.xlarge` nodes [49]), and multiple data centers (Amazon-ECs, data centers of N. Virginia, N. California and Oregon). A complete description of the environments and results regarding fault tolerance and attacks can be found in our previous work [30], [20].

Table II. AUTHS/S WITH 20, 40, 80 AND 100 CLIENTS

Environment	20 clients	40 clients	80 clients	100 clients
UFAM-VMs	867.73	984.59	995.12	960.11
Amazon-EC2	1969.17	2166.58	2244.30	2244.04
Amazon-DCs	26.66	50.72	92.42	114.05

The best throughput is achieved by the Amazon-EC2, reaching over 2,200 OpenID 2.0 authentications/s. The main difference between UFAM-VMs and Amazon-EC2 is the computing power of the nodes, which are twice 2x more powerful in the Amazon-EC2 environment. In the UFAM-VMs we achieved a throughput of approximately 1k authentications/s. Lastly, the inter-data center deployment (Amazon-DCs), despite using the same computing nodes of the Amazon-EC2 environment, achieved the lowest performance. The main cause of this substantial drop in performance was the network latency between data centers, which was around 94x higher than the latency in the other two environments. Nevertheless, the performance (up to 114 authentications/s with 100 clients) is still enough to support the demand of enterprises with thousands of users, as shown by our statistics analysis of a real environment (see next section). Another interesting thing to mention is the fact that we have already identified several optimizations and improvements that can be done to significantly increase the system performance in all three setups [30].

B. Discussing the capacity and scaling of the system

We used the authentication statistics of a real IT infrastructure to statistically estimate the capacity of our prototype, i.e., to estimate the number of users that it can support. The reference institution has two authentication systems, multiple OpenLDAP and Active Directory (AD) servers.

Both authentication services are used by almost all the services and protocolos of the institution, such as SMTP and HTTP servers, Windows system services and components, IEEE 802.1X in wireless infrastructures, and Web content management systems. The authentication of dozens of online systems, provided by the institution to approximately 11.5k users, is also integrated through OpenLDAP. Furthermore, all logons on Windows and Linux labs, as well as other PCs, are also controlled by these two systems.

By analyzing seven days of logs from the OpenLDAP and AD servers, we identified: (a) 143,907 authentications during the worst peak hour (OpenLDAP + AD authentications per hour), which means an average of nearly 40 authentications per second; (b) 118 authentications in the worst possible case (worst OpenLDAP peak second + worst AD peak second), which in practice did not happened because the hour/second was not the same for both systems; (c) less than 102 authentications/s throughout all the seconds of the analyzed period.

Table III summarizes the capabilities of our prototype considering all three environments. Furthermore, we also estimate the number of users supported by increasing the number of instances (gateways and replicas) of the system. For instance, two gateways and eight replicas, four replicas working with each gateway, can potentially increase the system's performance by 2x (nearly linear performance). In this case, it is necessary to split the users among the instances (e.g., 50%) and/or apply load balancing techniques [50] on the gateways and database sharding techniques [51] on the replicas.

Table III. SCALING UP TO 1M USERS.

Environment	10k users	100k users	500k users	1M users
UFAM-VMs	4.16%	41.66%	208.30%	416.61%
Amazon-EC2	1.78%	17.82%	89.11%	178.22%
Amazon-DCs	35.07%	350.72%	1753.61%	3507.23%

A total of 10k users requires an average of around 40 authentications/s, considering an environment similar to the one described. Consequently, all three environments support this demand, requiring only 4.16% of the computing power of UFAM-VMs, 1.78% of Amazon-EC2, and 35.07% of Amazon-DCs. The Amazon-EC2 environment is capable of supporting an IT infrastructure with more than 500k users. This takes into consideration only the current implementation of the prototype. Though, there are different optimizations that could be done on the system and computing environment, such as (a) use the most recent version of BFT-SMaRT [52], which has several performance and durability optimization; (b) use optimized pools of thread on the gateway; (c) use multiple gateways, since the replicas are capable of processing more than 70k raw messages per second [52]; (d) use more powerful computing nodes such as `m3.2xlarge` [49], which nearly double the computing power of the nodes used in Amazon-EC2 environment; and (e) send requests in batches between the gateway and replicas. Therefore, we can arguably say that our system can be extended for supporting environments with more than 1M users and/or networked devices and systems that require more resilient and trustworthy authentication services.

C. Towards IdP-as-a-service: a win-win opportunity

Next, we provide cost estimations and discuss the possibilities of building reliable third party IdPs using a cloud-of-clouds model. Based on our previous evidence and experiments, we believe that there is a promising opportunity for the provisioning of secure and dependable IdP-as-a-Service. Our discussion and cost estimations corroborate in demonstrating the feasibility and benefits of IdPs-as-a-service.

With our Amazon-EC2 environment, we are capable of handling authentication requests of IT infrastructures with more than 500k users without any further optimizations on the system. As one Amazon EC2 `m3.xlarge` node costs 0.45 dollars an hour, and we used five of those nodes, we have a total cost of 54 dollars per day (at full operation/capacity). In one year we would have to spend $17,541.90, which is barely enough to buy the machines with the required computing power. If we look at an average salary of a *security specialist* [53] ($46,000/year), we can easily conclude that the cloud infrastructure costs is much less than the average salary of a single *security specialist*. Yet, we are not considering the infrastructure and maintenance costs (CAPEX, OPEX, TCO). Consequently, the outsourcing of critical services to specialized companies (e.g., Amazon) would be an interesting option to consider.

The infrastructure (virtual servers) at Amazon costs 17,541.90 per year. If we add to that 50% to provide the service (e.g., an OpenID provider), we come up with a total spending of 26,312.85 dollars, which is still cheaper than to have your own infrastructure and human resources. Yet, to Amazon, it could be an attractive business because the most complicated and costly part is the infrastructure, which is already provided. Therefore, as the company already has the required expertise (specialized infrastructure operation and security teams, and so forth), it is reasonable to add just another 50% to provide a IdP-as-a-Service. Consequently, both Amazon and customers would benefit from the technical and business model.

If we think about large scale demands (e.g., millions of users), things get even more interesting. For instance, with two instances of our Amazon-EC2 environment, we are capable of supporting an IT infrastructure with 1M users. In this case, the infrastructure provisioning would cost 35,083.80 dollars per year, which is not enough to pay one single *security specialist*.

Table IV gives a roughly idea of the spending costs of using secure, dependable and optimized identity providers from third parties, which could be provided as a service by companies such as Amazon, Rack Space, Google, or by a startup building a cloud-of-clouds service. We calculated the costs following the estimations presented in Table III. The infrastructure costs represents the real market practice (values) of Amazon. As an estimative, we consider that a company such as Amazon, which already provides the elastic infrastructure, is providing as well the IdP service. Thus, we added a service cost of $0.055037 per user, which we think is a reasonable value based on some OPEX costs estimation, with a good margin of net revenues. In the end, as can be observed, we have an average cost (IaaS + service) of $0.090077 per user/year. This value can arguably be considered as inexpensive for the customer and profitably for the service provider. Therefore, we could easily increase the service costs to make the business (from the service provider perspective) even more attractive. However, it is worth emphasizing that the key issue is to optimize the resources of the IdP-as-a-Service, i.e., do not over allocate resources before the demand. An IdP-as-a-Service has to be highly elastic and dynamically allocate the resources based on the business growth and customer needs.

Based on our numbers, one single customer, considering the average number of authentications/s and a demand of around 500k users, would generate an income of $90,120.80 per year (89.11% of an Amazon-EC2 like setup + service cost) for the service provider. If we extrapolate the costs and assume that a single user costs half a dollar per year, which is still extremely cheap for the customer, we would reach an astonishing 250k dollars for one single customer with 500k users, making this business highly profitable. For instance, in practical terms, a company like Facebook, which has over one billion active users [54], can reach a profit of 1.86 billion dollars only from user generated content [55]. This means that such companies would most probably be able to pay $0.090077 for having outsourced reliable and secure authentication and authorization infrastructures. Of course that a company like Facebook, due to the fact that it already owns a huge computing infrastructure and specialized human resources, could opt to have its own IdP service. However, other companies, whose main business is not IT, such as eBay, PayPal, among many others, could more likely opt for outsourced high quality authentication and authorization services. Notwithstanding, most of small and medium scale businesses would easily benefit from outsources IdP services.

Table IV. COSTS ESTIMATION FOR IDP-AS-A-SERVICE.

Cost/Users	**10k**	**100k**	**500k**	**1M**	**10M**
IaaS cost	$350.40	$3,507.65	$17,541.90	$35,083.80	$350,838.00
Service cost	$550.37	$5,503.70	$27,518.50	$55,037.00	$550,370.00
Total cost/y	$900.77	$9,011.35	$45,060.40	$90,120.80	$902,169.00

It is worth emphasizing that the service cost estimated in Table IV represents the revenue of a potential startup company specialized in building and provisioning of secure and dependable IdP-as-a-Service. In other words, a single enterprise customer (e.g., online social network system/business) with 10M users would generate an income of $902,169.00, which is reasonable enough to keep a team of specialized IT engineers.

V. FINAL REMARKS

In our previous work, we have dissected how it is possible to build more secure and dependable identity providers, which are one of the key components for ensuring the security of most IT infrastructures and systems. By developing different prototypes (e.g., resilient RADIUS, resilient OpenID) we have shown that it is possible to tolerate arbitrary faults (e.g., energy disruptions, connectivity failures, common software vulnerabilities, and so forth) and different types of attacks. However, small and medium enterprises cannot afford highly specialized IT teams to deploy and operate sophisticated systems, capable of ensuring critical properties such as confidentiality (despite eventual intrusions or malicious sysadmins), privacy and high availability (e.g., three-nines). To overcome the complexity and costs of deploying secure and dependable IdPs, we proposed IdP-as-a-Service as a viable and interesting win-win opportunity. IdP-as-a-Service can be built as a cloud-of-clouds model to achieve high levels of availability, scalability, elasticity, cost-effectiveness and robustness against a large diversity of threats and accidental or intentional problems.

Based on our first analysis, taking into account data and statistics from real environments and deployments, we discussed how IdP-as-a-Service can represent a new opportunity for cloud providers (or startups) willing to invest in this market niche. Some of the main challenges to overcome are related to identify and deploy the most appropriate system components for achieving elastic environments, high performance, and high levels of security, in particular regarding confidentiality and privacy of sensitive data and operations.

Acknowledgment. We would like to thank Vinicius Cogo and Douglas Macedo for their feedback. This work is supported by EU's 7th Framework Program (FP7), through the project SecFuNet (FP7-ICT-STREP-288349), and by CNPq/Brazil, through grants 590047/2011-6 and 202104/2012-5.

References

[1] S. Racherla, D. Cain, S. Irwin, P. Ljungstrom, P. Patil, and A. M. Tarenzio, *Implementing IBM Software Defined Network for Virtual Environments*. IBM RedBooks, May 2014.

[2] C. Dixon, D. Olshefski, V. Jain, C. DeCusatis, W. Felter, J. Carter, M. Banikazemi, V. Mann, J. Tracey, and R. Recio, "Software defined networking to support the software defined environment," *IBM Journal of R&D*, vol. 58, no. 2, 2014.

[3] D. Kreutz, F. M. V. Ramos, P. Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *ArXiv e-prints*, Jun. 2014.

[4] A. Khajeh-Hosseini, D. Greenwood, and I. Sommerville, "Cloud migration: A case study of migrating an enterprise it system to IaaS," in *CLOUD, 2010 IEEE 3rd International Conference on*. IEEE, 2010.

[5] G. Sattiraju, S. Mohan, and S. Mishra, "Idrbt community cloud for indian banks," in *ICACCI, 2013 International Conference on*, Aug 2013.

[6] D. Kreutz, F. M. Ramos, and P. Verissimo, "Towards secure and dependable software-defined networks," in *SIGCOMM HotSDN*, 2013.

[7] C. Morgan, "Cloud security concerns relating to the difference in infrastructure and data operational control," itSMF International, Tech. Rep., May 2011.

[8] L. M. Vaquero, L. Rodero-Merino, and D. Moran, "Locking the sky: A survey on iaas cloud security," *Computing*, vol. 91, no. 1, Jan. 2011.

[9] A. J. Kornecki, N. Subramanian, and J. Zalewski, "Studying interrelationships of safety and security for software assurance in cyber-physical systems: Approach based on bayesian belief networks," in *Proceedings of the FedCSIS*. IEEE, 2013.

[10] J. Torres, M. Nogueira, and G. Pujolle, "A survey on identity management for the future network," *IEEE Comm. Surveys Tut.*, 2013.

[11] P. Verissimo, N. Neves, C. Cachin, J. Poritz, D. Powell, Y. Deswarte, R. Stroud, and I. Welch, "Intrusion-tolerant middleware: the road to automatic security," *IEEE Security & Privacy*, vol. 4, no. 4, 2006.

[12] C. Tankard, "Advanced Persistent threats and how to monitor and deter them," *Network Security*, no. 8, 2011.

[13] "All about Stuxnet," 2013, http://stuxnet.net.

[14] M. Prince, "The DDoS that almost broke the internet," 2013, http://goo.gl/oeDrMY.

[15] Verizon, "Data breach investigations report," Tech. Rep., 2013.

[16] IBM, "IBM X-Force 2012 trend and risk report," Tech. Rep., 2013.

[17] S.-T. Sun, K. Hawkey, and K. Beznosov, "Systematically breaking and fixing openid security," *Computers & Security*, vol. 31, no. 4, 2012.

[18] D. Kreutz, H. Niedermayer, E. Feitosa, J. da Silva Fraga, and O. Malichevskyy, "Architecture components for resilient networks," SecFuNet.eu, Tech. Rep., 2013.

[19] D. Kreutz, O. Malichevskyy, E. Feitosa, K. R. S. Barbosa, and H. Cunha, "System design artifacts for resilient identification and authentication infrastructures," in *ICNS*. IARIA, 2014.

[20] D. Kreutz, E. Feitosa, H. Cunha, H. Niedermayer, and H. Kinkelin, "Increasing the resilience and trustworthiness of openid identity providers for future networks and services," in *ARES/ECTCM*. IEEE, 2014.

[21] R. Khan, J. Ylitalo, and A. Ahmed, "Openid authentication as a service in openstack," in *IAS*, Dec 2011.

[22] S. Wattal and A. Kumar, "Cloud computing - an emerging trend in information technology," in *ICICT*, Feb 2014.

[23] J. Sherry, S. Hasan, C. Scott, A. Krishnamurthy, S. Ratnasamy, and V. Sekar, "Making middleboxes someone else's problem: Network processing as a cloud service," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 13–24, Aug. 2012.

[24] V. Kotronis, X. Dimitropoulos, and B. Ager, "Outsourcing the routing control logic: Better internet routing based on sdn principles," in *HotNets-XI*. ACM, 2012, pp. 55–60.

[25] S. K. Fayazbakhsh, M. K. Reiter, and V. Sekar, "Verifiable network function outsourcing: Requirements, challenges, and roadmap," in *Hot-Middlebox '13*. New York, NY, USA: ACM, 2013, pp. 25–30.

[26] G. Gibb, H. Zeng, and N. McKeown, "Outsourcing network functionality," in *SIGCOMM HotSDN*. ACM, 2012, pp. 73–78.

[27] GEANT & TERENA, "eduroam," 2012, https://www.eduroam.org/.

[28] D. M. F. Mattos and O. C. M. B. Duarte, "Authentication and access control architecture for software defined networks," in *WNetVirt*, 2013.

[29] D. Kreutz, A. Casimiro, and M. Pasin, "A trustworthy and resilient event broker for monitoring cloud infrastructures," in *IFIP DAIS*, 2012.

[30] H. Niedermayer, D. Kreutz, E. Feitosa, O. Malichevskyy, A. Bessani, J. Fraga, H. A. Cunha, and H. Kinkelin, "Trustworthy and resilient authentication service architectures," SecFuNet.eu, Tech. Rep., 2014.

[31] Y. Chen and R. Sion, "To cloud or not to cloud?: Musings on costs and viability," in *2nd ACM SOCC*. ACM, 2011.

[32] A. Rot and M. Sobinska, "It security threats in cloud computing sourcing model," in *Proceedings of the FedCSIS*. IEEE, 2013.

[33] M. Correia, "Clouds-of-clouds for dependability and security: Geo-replication meets the cloud," in *Euro-Par 2013: Parallel Processing Workshops*, ser. Lecture Notes in Computer Science, 2014, vol. 8374.

[34] M. AlZain, E. Pardede, B. Soh, and J. Thom, "Cloud computing security: From single to multi-clouds," in *HICSS*, Jan 2012.

[35] A. Bessani, R. Mendes, T. Oliveira, N. Neves, M. Correia, M. Pasin, and P. Verissimo, "Scfs: A shared cloud-backed file system," in *USENIX ATC*. USENIX Association, Jun. 2014.

[36] A. Hume, Y. Al-Hazmi, B. Belter, K. Campowsky, L. Carril, G. Carrozzo, V. Engen, D. Garcia-Perez, J. Jofre Ponsato, R. Kabert, Y. Liang, C. Rohr, and G. Seghbroeck, "Bonfire: A multi-cloud test facility for internet of services experimentation." Springer, 2012, vol. 44.

[37] D. Durkee, "Why cloud computing will never be free," *Commun. ACM*, vol. 53, no. 5, May 2010.

[38] M. Prince, "Ceasefires don't end cyberwars," 2012, http://goo.gl/Vkljbi.

[39] B. Golden, "Capex vs. Opex: Most People Miss the Point About Cloud Economics," 2009, http://goo.gl/peS91k.

[40] X. Zhiqun, C. Duan, H. Zhiyuan, and S. Qunying, "Emerging of telco cloud," *Communications, China*, vol. 10, no. 6, June 2013.

[41] NTT, "Data center - nexcenter," 2014, http://goo.gl/t7uwX3.

[42] S. Kamara and K. Lauter, "Cryptographic cloud storage," in *Financial Cryptography and Data Security*, 2010, vol. 6054.

[43] R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan, "CryptDB: Protecting confidentiality with encrypted query processing," in *ACM SOSP*. ACM, 2011.

[44] Z. Dayioglu, "Secure database in cloud computing - cryptdb revisited," *International Journal of Info. Security Science*, vol. 3, no. 1, 2014.

[45] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, 1979.

[46] L. Goubin and A. Martinelli, "Protecting aes with shamir's secret sharing scheme." in *CHES*, vol. 6917. Springer, 2011.

[47] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 6, Jun. 2011.

[48] A. Khajeh-Hosseini, I. Sommerville, J. Bogaerts, and P. Teregowda, "Decision support tools for cloud migration in the enterprise," in *IEEE CLOUD*. IEEE, 2011.

[49] Amazon, "Amazon EC2 Pricing," 2014, http://goo.gl/WNEVvS.

[50] M. Das, S. Yadav, A. Kandhare, S. Malpani, R. Rathinam, and J. Thiagarajan, "Load balancing by endpoints," Sep. 14 2011, uS Patent App. 13/232,894.

[51] R. Cattell, "Scalable sql and nosql data stores," *ACM SIGMOD Record*, vol. 39, no. 4, 2011.

[52] A. Bessani, J. Sousa, and E. Alchieri, "State Machine Replication for the Masses with BFT-SMaRt," FCUL, Tech. Rep., Dec. 2013.

[53] Indeed, "Security specialist salary," 2014, http://goo.gl/RN9nR3.

[54] D. Tam, "Facebook by the numbers: 1.06 billion monthly active users," 2013, http://cnet.co/1jWlKJM.

[55] L. Fisher, "How much do social networks make from user-generated content?" 2011, http://tnw.to/1CUUS.

3rd Information Systems Education & Curricula Workshop

ISEC goal is to promote the discussion about the convergence between Computer Science and Information Systems topics, so that researchers can present a complete and detailed specification of their educational curricula by means of these two topics. We inspect papers that contribute to the better understanding of emerging and important educational fields of Computer Science (CS) and Information Systems (IS). Authors are invited to submit their papers in English, presenting the results of original research or innovative practical applications in the field.

Regarding the selection process, we plan to perform a triple review process.

Topics

Key issues in this workshop will focus on (but are not limited to):

- Convergence between IS & CS in higher Education
- Specification of IS Education Curricula
- General IS Theory
- IS Scope
- IS Educational Fields
- Student participation in research
- Adaptation to the European Higher Education
- Assessment of students
- Innovative teaching methods
- Training for career and skills development
- Definition of "knowledge" in IS
- Quality and evaluation of teaching
- Social and environmental commitment
- Curriculum organization and curriculum
- The Fundamental Concepts Underpinning IS
- Merge from IS to CS and vice versa in higher Education

Event Chairs

Fardoun, Habib M., King Abdulazziz University, Saudi Arabia

Gallud, José A., University of Castilla-La Mancha, Spain

Tesoriero, Ricardo, University of Castilla-La Mancha, Spain

Program Committee

Abou-Tair, Dhiah el Diehn, German Jordanian University, Jordan

Aknin, Noura, Université Abdelmalek Essaadi, Morocco

Arcega, Francisco, University of Zaragoza

Atzmueller, Martin, Kassel University, Germany

Bento da Silva, Juarez, Universidade Federal de Santa Catarina

Carretero González, Lorenzo, King Abdulaziz University, Spain

Cipres, Antonio Paules, University of Castilla-La Mancha

Collazos Ordoñez, Cesar Alberto, Universidad del Cauca

Corbalan, Montserrat, Technical University of Catalonia (UPC)

De la Guía, Elena, University of Castilla-La Mancha, Spain

Garcia Zubia, Javier, University of Zaragoza

Garrido, Juan Enrique, University of Castilla-La Mancha, Spain

Giménez, Rafael, Barcelona Digital Technology Centre, Spain

Igual, Raul, University of Zaragoza

Kempin, Nils, CGI, Germany

Lambropoulos, Niki, University of Patras, Greece, Greece

Llamas Nistal, Martin, Universidade de Vigo

Majchrzak, Tim A., University of Münster, Germany

Medrano, Carlos, University of Zaragoza

Mystakidis, Stylianos, University of Patras, UOC, UW, Greece

R. Penichet, Víctor M., University of Castilla-La Mancha, Spain

Restivo, Teresa, Universidade do Porto

Romero Lopez, Sebastian, PSJA Southwest Early College High School

Tambo, Erick, United Nations University, Germany

DOI: 10.15439/2014F467
ACSIS, Vol. 3

Proprietary versus Open Source Software in Support of Learning in Computer Science

Antoine Melki
University of Balamand
Department of Computer Science,
Faculty of Sciences
P.o.box 100, Tripoli, Lebanon
Email: amelki@balamand.edu.lb

***Abstract*—the topic of learning Computer Science had been the subject of many researches. From time to time, this topic flows to the surface especially when Computing instructors report students' difficulties in acquiring the necessary knowledge. On the other hand, different studies had tackled the issue of open source software in computing education from different perspective. These studies, in general, conflict with the positions taken by the different suppliers of proprietary software. This paper will investigate the contributions of both categories of software in the process of computer science learning, and then will compare these contributions to the principles of active learning in computer science, to conclude which of the 2 categories is more advantageous.**

I. Introduction

DIFFERENT studies had tackled the learning model in computer science and came with different results. Despite that, active learning finds the highest amount of support among the scholars who engaged in this line of studies. In the technology field, the proprietary software allies, who are mainly the companies that produce it, claim relying on active learning and providing the best computer science education. On the other hand, with the increase in the popularity of open source software, the allies mark the same claim. This position paper represents a descriptive study of what both software categories claim to provide to computer science education. A comparative revision concludes the paper. The value of this study lies in providing a theoretical background to support the adoption of open source software in teaching computer science.

II. Learning

A. Definition and theories

Learning is the process of acquiring modifications in existing knowledge, skills, habits, or tendencies through experience, practice, or exercise. Learning includes associative processes, discrimination of sense-data, psychomotor and perceptual learning, imitation, concept formation, problem solving, and insight learning.

Gestalt-psychology researchers drew attention to the importance of pattern and form in perception and learning, while structural linguists argued that language learning was grounded in a genetically inherited "grammar." Developmental psychologists such as Jean Piaget highlighted stages of growth in learning. More recently, cognitive scientists have explored learning as a form of information processing, while some brain researchers, such as Gerald Maurice Edelman, have proposed that thinking and learning involve an ongoing process of cerebral pathway building. Among the related topics of research is transfer of training which is defined as the activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something [1].

Learning theories tend to fall into one of several perspectives or paradigms, including behaviorism, cognitivism, constructivism, and others [1].

Cognitivism is based on the idea that the mental function can be understood and the learner is viewed as an information processor. Cognitivism focuses on inner mental activities. It is necessary to determine how processes such as thinking, memory, knowing, and problem-solving occur. People are not "programmed animals" that merely respond to environmental stimuli; people are rational beings whose action are a consequence of thinking. Cognitivism uses a metaphor of mind as computer where information comes in, is being processed, and leads to certain outcomes.

Constructivism's basic idea is that learning is an active and constructive process where the learner viewed is an information constructor. People actively construct or create their own subjective representations of objective reality. New information is linked to prior knowledge, thus mental representations are subjective.

Another concept in learning which is relevant to this study is the concept of model which is a theoretical construct or mental picture that helps one understand something that cannot easily be observed or experienced directly. Models are testable ideas created by people that capture a story about what is happening in nature [2].

B. Learning Computer Science

The literature on learning in Computer Science education is large and much diversified. What is common among most of the paradigms is the emphasis on active and collaborative learning. This is based on a common trait of almost all the educational psychology literature on learning in recent years which admits that learners construct their knowledge by interacting with their environment and other people. This is close to constructivism where some focus primarily on the individual learner and others focus primarily on the social nature of knowledge construction. In either case, the consensus is that education is not the mere transmission of knowledge from the teacher to the student but requires that students be active. Furthermore, collaborative learning is becoming a key component in many college classrooms with several benefits: improved achievement, enhanced critical thinking competencies, improved attitudes towards the subject area, and reduced student anxiety.

In the case of Computer Science, students spend the majority of their time solving problems that require them to learn skills that are applications of the concepts from the readings of manuals or classes. Neither the teacher nor the teaching Assistant is a pure lecturer, but assumes the role of facilitator in many cases, especially in Lab settings. Solutions require students to learn and practice new skills. Each problem builds on previous skills and concepts, extending the range of the students' capabilities. Generally, especially with the presence of social networks, students spend some time discussing possible solutions to the problem in groups, in pairs or individually before attempting to solve the problem. It is known that students compare results and discuss problems they encounter and solutions they discover. The transfer of learning is a major issue in learning Computer Science, since the application of learning is always a real life case; where requirements and specifications are not explicitly clears [2].

III. *Proprietary Software And Learning*

Companies providing proprietary software generate revenue by selling products in the form of software licenses and additional services and support. Accordingly, their business model is built with a very small space left for free offering. This has an effect on the companies views of education and consequently on how they see learning and strive to support it. In this section, the 3 major proprietary companies are selected as an example to support the analysis: Cisco, apple and Microsoft.

A. Proprietary Companies' Education

Cisco sees that there are pressures on education in the 21st century, mainly because education needs to change in order to adapt to the needs of the 21st century. One driver of these pressures is digital technology. Because the new technologies demand a new set of skills, digital technologies exert pressure for change but at the same time provide opportunities for transforming pedagogy because they offer access to information, networks for communication, and new means of presenting learning. A set of other factors also apply pressures on education. Globalization is one factor that exerts social and economic pressure, and provides opportunities for wider, richer learning. Other factors include economic recession, demographic pressures, and environmental stability. In responding to these pressures, educators are coming up with innovations that enrich learning and help them in dealing with specific challenges.

Meanwhile, these pressures and opportunities require people to acquire new kinds of literacy including information literacy, cross-cultural literacy, and ecological literacy. Learners are expected to be lifelong learners, because technology, politics, economics, and the environment are changing so quickly. This demands a shift away from focusing on engagement in school, to engagement in learning. It also requires an examination of what sorts of environments are most conducive to learning in the 21st century.

Cisco proposes a strategy that yields to the formation of a learning ecosystem. It claims building its strategy on the latest thinking and research about innovation, and it believes this strategy is of practical use to system leaders in education. This proposal is built on the distinction between formal learning and informal learning contexts, and education provision delivered by existing providers and new entrants. Then in another stage comes the application of new resources and new sources of insights to the education landscape, specifically, the new resource of digital technologies, which have the potential to radically transform learning, though they are not often used to their full potential. The proposal also addresses learner ownership, because when learners feel ownership of their learning, they are able to apply their own insights about how they learn best, and become "co-producers" of learning rather than just "consumers." On the other hand, system leaders need to reposition themselves so that rather than being primary providers of education, they provide a platform for a diversity of providers [3].

Apple concentrates more on hardware and emphasizes the services rendered to instructors who can customize students' devices with materials that fit their level and learning style so that the machine becomes a more powerful learning tool. Apple claims that its services allow more creativity, less time of preparation, and teaching with content from top institutions. For students who learn best by listening, the instructor is advised to download a podcast from iTunes. And for those who learn through tactile interaction, the instructor is advised to find an app. Apple also claims that the wide range of content across subjects and grades also makes it easy to tailor apple machines for students at a variety of learning levels so that the instructor can teach all the students the same lesson in different ways [4].

Microsoft focus is more on systemic education in order to produce a more efficient and effective learning environment, advocating sophisticated metrics to measure results. Microsoft looks at making the teachers relying on it better at their jobs than another and how can best practices be shared?

Technology enables analysis and is also the delivery mechanism.

Microsoft is concerned more with the science of education than the art of learning. Its products are claimed to facilitate students' ability to consume and create content and collaborate across multiple devices so that the learning process is extended beyond the classroom [5].

As it is made clear, proprietary organizations deal more with products than with learning per se. Even learning is a product. In conclusion, the principles of learning presented by proprietary organizations can be summarized as follows:

1. Emphasis is placed on altering learning patterns and models instead of moving learners
2. The level of freedom given to the users does not go beyond changing some predefined options; accordingly, the user's contribution is limited and directed
3. Technologies have predefined roles.

B. Proprietary Software and Learning Computer Science

The majority of the companies that sell proprietary software have their own professional schools. Many educational institutions contract with these companies for an exchange of technology with the curriculum. So the educational institution becomes bound to teach the curriculum supported by the corresponding technology of the partnering proprietary company. On the other hand, the company offers its product for free to the institution. Some institutions go further than that by adopting a curriculum structure analogous to that of the certificates offered by the proprietary partnering company. The result of this process is graduates who are specialized in the product of these companies.

In addition, these companies release, from time to time, offers of cheap or free software for students. All this ends in enlarging the share of the corresponding companies on the human capital that include developers, software engineers, database designers, systems analysts, information systems specialists, etc.

Further than that, proprietary companies create their own communities, competitions and events that are targeted at attracting students to their product. Examples include: Microsoft Imagine Cup and Cisco Networking Academy NetRiders.

Accordingly, the features of proprietary companies' education include:

- Content is added, edited and updated by the company
- Materials are the product of one author which is the company itself
- Releases and updates the result of a process controlled by the owning company.

In summary, the learning provided by the companies that sell proprietary software is a controlled learning.

C. Open Source Software and Learning

The prevalence of the Internet as of the early nineties has facilitated the collaboration between software programmers from different poles of the world, allowing an easy distribution of their production. In addition to that, the distinct advantages offered by Free/Libre and Open source software (FLOSS) resulted in an increasing recognition and adoption of this category of software.

The impact of FLOSS is discussed in many studies. International Data Corporation (IDC) gives an estimate of 22.4% growth in revenues in 2013, reaching 8.1 billion dollars. Comparisons show an increase in the rate of adoption of FLOSS as compared to proprietary software in operating systems and web servers, while proprietary, mainly Microsoft and Apple, are keeping the lead in desktop usage [6].

The studies describing the use of FLOSS in education show a set of principles [7]:

- Content is not something static but dynamic
- Learning resources are manifold
- Users are also active creators
- Support and learning resources are closely connected
- Open and transparent structures foster re-use and discourse, but also continuous improvement and evolutionary growth
- Existence of a wide range of possible activities to engage at around the core product
- Self-studying and learning from what others did are the pre-dominant form of learning.

D. Open Source Software in Computing Education

The penetration of open source software in computing schools and departments is natural, and so it is increasing, with some of the major players in the software market increasing the amount of revenue they make from activities that use FLOSS. A non-exhaustive summary of the advantages of FLOSS in teaching Computer Science can be:

- Lower costs for both the university and the students
- Projects are more beneficial for research since the user of FLOSS is free to get the source and implement new ideas.
- FLOSS is valuable for teaching as it offers students the opportunity to share their contributions to projects [8].

The main disadvantage in using FLOSS is the lacks of hardware drivers especially for wireless, graphic cards and suspend/sleep functionality in laptops. On the other hand, studies show that students' feedback on using FLOSS is mixed. While many students get excited to open source tools, others consider that learning a proprietary tool gives a better chance to get a job.

As for the instructors, studies showed that adopting FLOSS is a challenge due partly to some misconceptions. A non-exhaustive list of the issues that have to be overcome in teaching FLOSS can include [9]:

1. Misconceptions concerning the adoption of FLOSS, like:

a. Only hobbyists use FLOSS.

b. It is a niche market with limited diffusion.

A number of studies showing current FLOSS penetration rates in the market prove the fallacy of these misconceptions.

2. Lack of managerial support at the level of department or school, which is usually overcome by a decision to substitute some or all proprietary products by open source.

3. A misconception concerning the quality of FLOSS is that proprietary software is better and students learn more by using better software. This is not true as some FLOSS is among the leaders in their market segment as shown by reviews from publications in the field. On the hand, from an educational point of view, there should not be any difference in what can be taught using FLOSS or proprietary software, as concepts are the same [10].

On the other hand open source in computing education show that a learning environment is formed around communities. While this applies more to informal education settings, schools and departments can still participate and contribute, as this belongs to the core principles of FLOSS. These communities can be described as follows [11]:

- Dealing with up to date content where everyone can add, edit and update the content
- Materials are usually the product of many authors with many contributions from people other than original authors
- Relying on frequent releases and updates where product features and community structures are the result of a continuous process of renegotiation and reflection within a continuous development cycle
- Reusing of prior learning outcomes and processes, as these are systematically available through mailing lists, forums, commented code and further instructional materials
- The community members represent a large support network that voluntarily function in a collaborative manner
- New solutions are adapted early by the community.

IV. Active Learning and Computer Science

In Computer Science, it is crucial to create an active learning environment to improve students' comprehension and retention of material, allow students to take control and regulate their own learning, and eventually empower them with necessary skills to solve problems outside of the classroom. Over the years, various strategies have been developed by Computer Science instructors to promote active learning [12].

A literature review would find many definitions of active learning; however, several essential components are common between the different definitions:

1. Active learning is not the passively listening to lecture, where students apply material to "real life" situations [13].

2. Usually, an active learning environment allows students to talk about what they are learning, write about it, relate it to past experiences, and apply it to their daily lives, "they must make what they learn part of [14].

3. Supporting active learning is necessary because "what students learn is greatly influenced by how they learn, and many students learn best through active, collaborative, small group work inside and outside the classroom" [15].

4. As Briggs points out, in a rapidly changing field as Computer Science, "students tend to be active and visual learners" and it is beneficial to provide a learning environment where students can interact with the material. "Active learning is especially effective for CS students who tend to be visual/intuitive learners." [16]

5. Active learning can incorporate many instructional methods, such as collaborative/ cooperative, project/ problem-based learning, role play, debates, etc. and even the use of functional emerging instructional technology teaching tools. Lindquist et al. [17] demonstrates how mobile phones can be used to "broaden and enhance the use of active learning in large classrooms."

V. An Assessment of The Contributions

The goal of this position paper is achieved through the comparison of the support provided by proprietary software to active learning in Computer Science to the support offered by open source software. For this reason, the following 6 characteristics of active learning are derived from the literature review

a. Students applying material to "real life" situations: The scope of open source real life situations for Computer Science students is much larger than that of proprietary software. In the latter case, an internship or a project at an organization requires that this organization employs the same tools and software as those used at school; otherwise a proprietary training is needed. In the case of open source, both parties, the organization and the trainee, are not bound to any specific software. In case of any restriction, training and related materials are usually free and available through many communities.

b. Relating what is learned to past experiences: The concept is so similar to software reuse, which is possible in the realm of open source at a much wider scale than that of proprietary. In open source software, reuse can be at all levels, from operating systems to interfaces, which does not hold true in the case of proprietary.

c. Visual/intuitive learning: Both proprietary and open source software can offer the same learning. There is not enough literature to support the superiority of any category over the other.

d. Collaborative/cooperative learning: The community nature of open source gives it advantage in the issue of collaboration and cooperation. As mentioned before, proprietary companies form their own communities, but that lack the level of freedom and wide coverage existing in the open source communities. Project/ problem-based learning: There is no literature that gives advantage to any category over the other in problem-based learning. As for project learning, the facts verify that more initiatives and projects can be initiated in open source. A clear example is the Linux operating system.

e. Role playing: The level of freedom and independence provided by open source software gives it an advantage over proprietary software in the case of role playing. The possibilities of accessing the source codes, studying the designs, and the option to apply operations like reverse engineering, allow the learners to play all the roles played in the life cycle of software.

VI Conclusion

In closing, open source software backed by communities is clearly a better supporter of active learning than proprietary software backed by companies. By accepting that active learning is an effective model to learn Computer Science, it can be concluded that open source software provides more opportunities to learn computer Science, than what proprietary software provides.

References

[1] D. H Schunk. Learning theories: an educational perspective. Boston: Pearson, 2012, p. 8.

[2] O. Hazzan, T. Lapidot, and N. Ragonis, Guide to Teaching Computer Science: An Activity-Based Approach, Springer-Verlag London Limited, 2011, pp. 35-45.

[3] Cisco. The Learning Society. San Jose: Cisco, 2010, http://www.innovationunit.org/sites/default/files/The%20Learning %20Society%20White%20Paper.pdf, retrieved on 22/5/2014.

[4] Apple Inc.Apple and Education. http://www.apple.com/education retrieved on 22/05/2014.

[5] Microsoft Inc. Microsoft in Education. http://www.microsoft.com /education/ww/products/Pages/Products.aspx. Retrieved on 22/05/2014.

[6] D. Lipsa and R. Laramee, "Open Source Software in Computer Science and IT Higher Education: A Case Study", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.1, January 2011.

[7] A. Meiszner, "Learning the Open Source Way: FLOSS as a learning environment", OpenLearn 2007 conference, Milton Keynes, UK, 2007.

[8] R. Bosch and S. Vila-Marta, "Current Trends in Free Software Research", Research report LSI-08-36-R. 2009, retrieved from http://www.lsi.upc.edu/dept/techreps on 22/5/2014.

[9] Morelli, A. Tucker, N. Danner, T. de Lanerolle, H. Ellis, O. Izmirli, D. Krizanc, and D. Parker, "Revitalizing Computing Education Through Free and Open Source Software for Humanity", CACM, 52, 8, (2009). Retrieved from http://doi.acm.org/10.1145/ 1536616.1536635 on 22/05/2014.[doi:10.1145/1536616.1536635]

[10] A. Cerone, and S. Sowe, "Using Free/Libre Open Source Software Projects as E-learning Tools", Proceedings of the Fourth International Workshop on Foundations and Techniques for Open Source Software Certification (OpenCert 2010), Electronic Communications of the EASST, Volume 33, 2010.

[11] S. Sahami, A. Aiken, and J. Zelenski, "Expanding the Frontiers of Computer Science: Designing a Curriculum to Reflect a Diverse Field", SIGCSE'10, Milwaukee, Wisconsin, USA, March 10–13, 2010, retrieved from http://portal.acm.org on 10/03/2014. [doi=10.1.1.157.8431]

[12] D. Schweitzer and W. Brown. Interactive visualization for the active learning classroom. ACM SIGCSE Bulletin, 39(1), 2007, pp. 208–212. [doi=10.1.1.108.247]

[13] D. Paulson and J. Faust. Active learning for the college classroom. 2002. Retrieved from http://chemistry.calstatela.edu/Chem&Bioch/active/ main.htm on 22/05/2014.

[14] A. Chickering and Z. Gamson. Seven principles for good practice. American Association Higher Education Bulletin, 39, 1987, pp. 3–7.

[15] T. Briggs. Techniques for active learning in CS courses. Journal of Computing Sciences in College, 21(2), 2005, pp. 156–165.

[16] D. Lindquist, T. Denning, M. Kelly, R. Malani, W. Griswold, and B. Simon. Exploring the potential for mobile phones for active learning in the classroom. Proceedings of the 38th SIGCSE Technical Symposium on Computer SCIENCE Education, 2007, pp. 384–388. [doi:10.1145/1227310.1227445]

Emerging Aspects in Information Security

ADMITTEDLY, information security works as a backbone for protecting both user data and electronic transactions. Protecting the communication and data infrastructure of an increasingly inter-connected world has become vital nowadays. Security has emerged as an important scientific discipline whose many multifaceted complexities deserve the attention and synergy of the computer science, engineering, and information systems communities. Information security has some well-founded technical research directions which encompass access level (user authentication and authorization), protocol security, software security, and data cryptography. Moreover, some other emerging topics related to organizational security aspects have appeared beyond the long-standing research directions.

The Emerging Aspects in Information Security (EAIS'14) workshop focuses on the diversity of the information security developments and deployments in order to highlight the most recent challenges and report the most recent researches. The workshop is an umbrella for all information security technical aspects. In addition, it goes beyond the technicalities and covers some emerging topics like social and organizational security research directions. EAIS'14 is intended to attract researchers and practitioners from academia and industry, and provides an international discussion forum in order to share their experiences and their ideas concerning emerging aspects in information security met in different application domains. This opens doors for highlighting unknown research directions and tackling modern research challenges. The objectives of the EAIS'14 workshop can be summarized as follows:

- To review and conclude researches in information security and other security domains, focused on the protection of different kinds of assets and processes, and to identify approaches that may be useful in the application domains of information security
- To find synergy between different approaches, allowing to elaborate integrated security solutions, e.g. integrate different risk-based management systems
- To exchange security-related knowledge and experience between experts to improve existing methods and tools and adopt them to new application areas
- To present latest security challenges, especially with respect to EC Horizon 2020

Topics

Topics of interest include but are not limited to:

- Biometric technologies
- Human factor in security
- Cryptography and cryptanalysis
- Critical infrastructure protection
- Hardware-oriented information security
- Social theories in information security
- Organization- related information security
- Pedagogical approaches for information security
- Individual identification and privacy protection
- Information security and business continuity management
- Decision support systems for information security
- Digital right management and data protection
- Cyber and physical security infrastructures
- Risk assessment and risk management in different application domains
- Tools supporting security management and development
- Emerging technologies and applications
- Digital forensics and crime science
- Misuse and intrusion detection
- Security knowledge management
- Data hide and watermarking
- Cloud and big data security
- Computer network security
- Security and safety
- Assurance methods
- Security statistics

Event Chairs

Awad, Ali Ismail, Luleå University of Technology, Sweden

Bialas, Andrzej, Institute of Innovative Technologies EMAG, Poland

Program Committee

Banerjee, Soumya, Birla Institute of Technology

Bun, Rostyslav, Lviv Polytechnic National University

Clarke, Nathan, Plymouth University, United Kingdom

Cyra, Łukasz, European Commission - Joint Research Centre Institute for the Protection & Security of the Citizen

Dworzecki, Jacek, Police Academy in Szczytno

Fernandez, Eduardo B., Florida Atlantic University, United States

Furnell, Steven, Plymouth University, United Kingdom

Furtak, Janusz, Military University of Technology, Poland

Geiger, Gebhard, Technical University of Munich, Faculty of Economics

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Hämmerli, Bernhard M., Hochschule für Technik+Architektur (HTA), Switzerland

Harnesk, Dan, Luleå University of Technology

Hasssaballah, M., South Valley University, Egypt

Kalbarczyk, Zbigniew, University of Illinois at Urbana-Champaign

Kapczynski, Adrian, Silesian University of Technology, Poland

Klamka, Jerzy, Polish Academy of Sciences

Kosmowski, Kazimierz, Gdansk University of Technology

Mahmoud Mohamed, Ehab, Osaka University, Japan

Mamojka, Mojmír, Police Academy in Bratislava

Pańkowska, Małgorzata, University of Economics in Katowice, Poland

Rot, Artur, Wroclaw University of Economics, Poland

Soria-Rodriguez, Pedro, Atos Research & Innovation

Stokłosa, Janusz, Poznań University of Technology

Suski, Zbigniew, Military University of Technology

Szmit, Maciej, Orange Labs Poland, Poland

Thapa, Devinder, Luleå University of Technology

Yen, Neil, The University of Aizu, Japan

Zamojski, Wojciech, Wrocław University of Technology

Zieliński, Zbigniew, Military University of Technology

DOI: 10.15439/2014F439
ACSIS, Vol. 3

ICT Security Risk Management: Economic Perspectives

Gebhard Geiger
Technical University of Munich.
Faculty of Economics
Arcisstrasse 21
80333 München, Germany
Email: g.geiger@ws.tum.de

***Abstract*—ICT security incidents are frequent and prevalent. They not only pose threats to the integrity and operation of the critical infrastructures of modern society, but also tend to cause enormous damage to the public economy. In addition, ICT security management can be very costly, while the success of security technologies and strategies often remains uncertain. From the economic perspective, this situation requires various basic research approaches to be taken to increase ICT security. These include an improved understanding of the far-reaching, highly interconnected consequences of ICT security incidents, development of methodologically sound measures of ICT security risk as well as approaches to measure the effectiveness and cost-efficiency of the ICT security management.**

I. Introduction

MODERN societies almost completely depend on electronic information and communications services and systems that are susceptible to *cyber*-attack. This susceptibility also characterises socio-economic organisations that provide society, in ICT-dependent ways (information and communications technology, ICT), with vital services such as administration, telecommunications, energy, transport, and financial services ("Critical Information Infrastructures", CII).

While in the engineering sciences conventional risk and vulnerability analyses of complex systems have largely focussed on problems of technological safety, research on ICT systems is also concerned with the security of critical infrastructures *vis-à-vis* information- and network-based threats and attacks. Threat thereby means intentional, including preparatory, individual or organised social action directed against the integrity and availability of data, ICT systems and infrastructures. Violation of ICT security, on its part, usually entails consequences that go far beyond the technological and operational problems of systems engineering and management. When successful, *cyber*-attacks tend to cause enormous damage to the public economy. In addition, ICT security management can be very costly, while the success of security technologies and strategies often remains uncertain. From the economic perspective, this situation requires various basic approaches to be taken to improve ICT security. These include an improved understanding of the far-reaching, highly interconnected consequences of ICT security incidents (systems analysis), methodologically sound measures of ICT security risk (quantitative risk analysis) as well as measures of the effectiveness and cost-efficiency of the ICT security management (quantitative risk assessment).

The present paper provides selected theoretical and methodological perspectives on ICT security management and some of its basic economic requirements and implications. It is in part adapted from earlier contributions by the author to security and risk management theory [1], [2].

II. ICT and CII Systems Analysis

From the methodological point of view, two basic questions arise in the economic analysis of ICT and CII security risk management. First, how much effort does modern society have to invest into CII security provisions to keep potential losses from *cyber*-attacks within the boundaries of acceptable risk? Secondly, how can potential losses be measured or estimated, given the fact that the infrastructures under attack are normally highly interconnected and the consequences of an attack are widely distributed and uncertain?

A suitable conceptual and methodological framework for attack consequence analysis is provided by modern systems research. In the social and engineering sciences, systems-theoretical concepts and methods have been developed and applied primarily to model, design and control complex socio-technological interactions. Their suitability for interdisciplinary research rests on their capacity to describe adequately the structure and behaviour of both physical systems and the "intentional" attributes of social action such as interests, preferences and purposes. Accordingly, conventional engineering approaches to systems safety can be directly extended to cover aspects of ICT security by combining, in systematic ways, models of systems behaviour, quality management, robust systems design, and failure response and recovery with threat and vulnerability analyses and strategies of ICT security management.

Using the conceptual frameworks of theoretical and applied systems analysis, four key approaches to ICT infrastructure protection can be identified. The first is the application of models and methods of operational research (OR) such as exercises, simulations, games of strategy and scenarios to information-based social systems. OR experiments and scenario techniques are particularly suitable for exploring the vulnerability of information infrastructures under realistic conditions, especially incident response and recovery of large-scale information-based systems under computer network attack. The results obtained from OR experiments can be further analysed and refined by combining them with conventional fault tree (event tree, etc.) representations of IT security incidents. Scenarios established in this way can provide a detailed overview of possible incidents across the entire threat spectrum, which, even in the absence of probability estimates, admits realistic predictions of the effects particular incidents and responses to them will have. Alternatively, OR experiments simulating IT security threats can be carried out to generate the statistical data-bases required for the design and optimisation of risk and security management strategies.

The second key approach is based on the concept of systems vulnerability adapted from industrial safety engineering. ICT systems and infrastructures are said to be vulnerable to failures (risks, threats) to the degree to which they (will likely) lose their operability if one or more of their components fail (a safety or security incident occurs). Related concepts are the robustness, resilience, reliability and availability of systems. Vulnerability analysis is concerned with the interactions between system components, and the interrelationships between system design and function. Once combined with suitable OR techniques, vulnerability analyses can provide insight into the overall performance, degradation or robustness, of large-scale information infrastructures in the event of an IT-based attack. The political and economic significance of such insight would seem obvious.

The third key approach to the security of ICT systems and infrastructures to be included here elaborates upon modern business process simulation techniques. The significance of these approaches arises from the fact that the thrust of the economic damage caused by *cyber*-attacks is very likely not triggered by the physical destruction of the ICT systems themselves. It will arise rather from the delay or disruption of remote and ramified, computer-based economic transactions that are affected, for instance, the economic consequences of the disruption of international logistic chains, financial transactions or e-commerce.

Finally, advanced approaches to theoretical and applied risk analysis must be included here. ICT risk research is broadly concerned with the causes, frequency, consequences and management of damage that may arise from the operation of ICT systems. The public policy implications of IT security management (economic costs, limits of acceptable risk, comparative risk assessments, etc.) often require quantitative risk assessments. These could in principle be obtained using statistical methods and familiar models of rational planning and decision making under risk. Unfortunately, ICT security risks involve threats of purposeful, covered action rather than measurable system failure rates with recurrent, identifiable causes. Primarily because of this dependence on strategic as opposed to probabilistic uncertainty, however, ICT security risks are hard to assess in statistical terms.

III. Economics and Security Risk Assessment

In view of growing threats to public ICT security and increasing budgetary restraints, risk management in government, industry and business must be both effective and cost-efficient. To this goal, recent advance in the econometric and operational sciences must be exploited to develop and apply a generic quantitative risk assessment methodology as a security planning and management device to protect public infrastructures and large-scale ICT systems. The concept of quantitative risk assessment thereby means the coherent intrinsic, or “fair”, pricing of risks. It implies considerably more than risk measurement in the sense of statistical risk analysis (“intrinsic” refers to risk quantification within a given accounting system rather than to risk prices extrinsically determined by the market for risky goods or services). It is evident that the practical use of a coherent approach to measure the intrinsic value of any given risk would be considerable. It could help to determine, in a realistic and systematic way, the amount of risk reduction achieved per euro invested in technologies and management efforts to prevent safety and security incidents in CII, or mitigate the damage arising from such incidents. As for security management, this is exactly what is otherwise known (though badly missing in practical applications) as calculating the Return on Security Investment (ROSI).

IV. Quantitative Risk Assessment

Risk management has long been suffering from the fact that risk is an elusive concept. Correspondingly, existing methods to assess risks and risk reduction measures tend to be ambiguous and controversial, if not manifestly inconsistent, for one of the following two reasons. They are either *ad hoc* rather than systematic, meaning that they lack theoretical coherence, or hard to operationalise. In either case, they may not provide the reliable information decision makers need to solve their problems.

The following situation provides an instructive example. Although attempts have been reported to estimate probabilities of (e.g., terrorist, *cyber*-war) attacks in order to quantify security risks [3], [4]., the probability of such an event is generally not a well-defined concept, at least not in the strict sense of mathematical model building. In fact, security incidents imply planned, purposeful human action and, therefore, are quite the opposite of random events. When modelled in mathematical terms, they have to be conceptualised as “games of strategy” rather than “games of chance” [5]. To the extent that a terrorist´s plan of an attack is unknown to the operator of the system threatened, the attack (time, place, technology employed, etc.) involves uncertainty, but not probability. Accordingly, ICT security incidents must be modelled by “What-if” scenarios (uncertainty arbitrarily removed, probability of occurrence put equal to 1) and concentrate on their probabilistic damage consequences. The scenarios are based on the assumption that effective protection technologies will constrain the actions of the attackers and thus help to mitigate the consequences of the attacks or prevent them entirely.

Security risks are accordingly conceptualised as probability distributions of the amounts of loss or damage to be prevented or incurred in a security incident. In other words, “What-if” scenarios can be used as reference cases relative to which the probabilistic attributes of security incidents can be analysed in a definite way. For example, whether a *cyber*-attack may be successful or not may depend on whether the ICT assurance technology built into the system under attack has been updated recently or not – the classical case of risk reduction in the sense of incident consequence mitigation by means of physical protection.

V. The Econometric approach to risk assessment

Advance has recently been made on the basis of novel methodological approaches to economic utility theory and the statistical foundations of quantitative risk assessment. The methodology for optimal, cost-efficient risk and security management employed in these approaches involve concepts of “generalised expected utility” that have been demonstrated to be able to admit coherent, explicit numerical representations of risk preferences, while accommodating basic empirical, individual and social attitudes towards risk. Most importantly, however, they have proven to be sufficiently simple for operational use in applied risk research. In this context, it is also important to note that “utility” has nothing to do with naïve views of “degree of individual satisfaction”, “desirability” and the like: it is a technical term simply meaning a behavioural risk preference score.

The core concept of quantitative risk assessment is the pricing of risk. Risks can be formally represented as probability functions $f(x)$ of the likely gains or losses x (in monetary terms or otherwise) obtained from safety or security incidents with uncertain consequences. A real number $c(f)$ is called the *certainty equivalent of the risk* $f(x)$, if $f(x)$ and the certain amount $c(f)$ of gain or loss are indifferent in preference terms from the perspective of the planner or decision maker. The certainty equivalent of a given risk can accordingly be viewed as the fair, or “intrinsic” price of that risk, considering that f and $c(f)$ are equal in preference. In practice, it can be explicitly calculated for every given probability function f.

Fig. 1 illustrates important realistic features of the quantitative account of risk assessment. One such feature is the marked deviation of the fair price (curved line in Fig. 1) from the probabilistic mean value of a risk (straight line), thus expressing widely observed, non-neutral human attitudes towards risk. Another feature is the capacity of the present approach to accommodate patterns of variability of risk attitude across various dimensions of risk. Finally, this simple and straightforward concept of intrinsic pricing of risks provides a powerful management tool, admitting direct assessments to be made of the effectiveness and cost-efficiency of planning and decision-making under risk.

VI. Effectiveness of ICT Security Risk Management

Real systems can generally be assumed to be operated with larger or smaller risk management effort. Two risks f and g linked to the effort aiming to mitigate them can be estimated, considering the likely consequences of security incidents affecting any such system considered. Furthermore, the risk prices $c(f)$ and $c(g)$ of the risks with and without appreciable risk management arrangements, respectively, can be calculated and compared. For example, the comparison $c(f) \geq c(g)$ shows the effectiveness of the measures planned or taken to reduce the risk g to f. In this example, the price difference $c(f) - c(g)$ is positive. It measures the Return on Security Investment (ROSI) that can be gained when the system changes from the risky state g to the less risky state f. If, on the other hand, the difference $c(f) - c(g)$ is small or even turns out negative, the risk management proves ineffective.

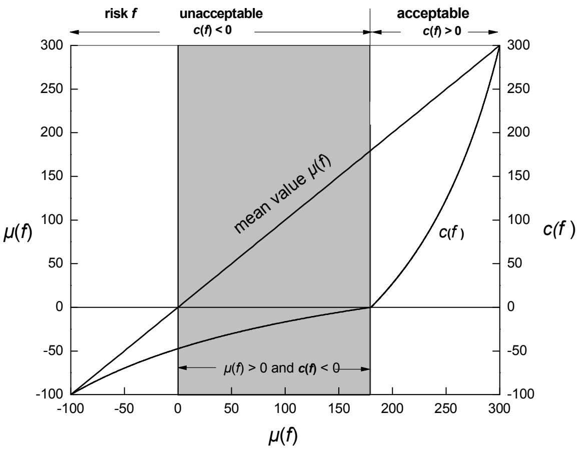


Fig. 1 Certainty equivalent $c(f)$ and mean value $\mu(f)$ of probability function f. Example after [1].

VII. Cost-Efficiency of Security Risk Management

Let $k(f, g)$ be the cost incurred by security managers to reduce the risk g to f. The ratio of ROSI to cost of the security arrangements made gives the amount of risk reduction per euro invested. It measures the cost-efficiency of the risk reduction achieved. Risk management is optimal if for given “*status quo* risk” g, the target risk level f is chosen so that the cost-efficiency ratio is at maximum within a given set of alternative risk mitigation choices.

A hypothetical numerical example is shown in Fig. 2. In the example, q is the rate at which a firewall technology detects hacker attacks and malware programmes of given types directed against a privately owned computer network. Without the firewall in operation, x is the amount of economic damage incurred or prevented with probability $g(x)$ (e. g., Euros, in monetary terms) if an attack occurs. The equivalent number of Euros saved or lost increases from the *status quo* with $c(g) = 0$ and $q = 0\%$ to $c(f)$, if money is invested to adjust q optimally. Clearly, it reaches its maximum for $q = 100\%$. The function $k(q)$ gives the buying price and operational cost (per unit time) of the firewall system. It shows the typical effects of “economies of scale” and increasing and decreasing marginal costs known from managerial economics [6]. The fact that for large q-values the k-curve is steeply rising results from disproportionate increases in expenditure for large values of q. High levels of security may then become unaffordable. The cost-efficiency ratio $c(f)/k$ reaches its maximum at approximately q = 22% in this example. Such a low value of the cost efficiency ratio means that the firewall technology is very expensive, while the network-based applications involved are of moderate or little economic value.

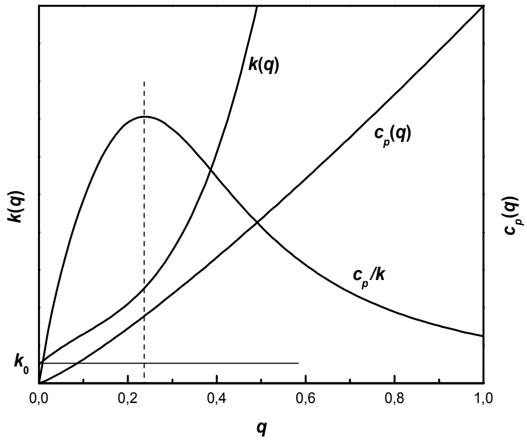


Fig. 2 Risk price $c_p(q)$, cost $k(q)$ of security technology and cost-efficiency ratio c_p/k as functions of the attack detection rate q achieved

The difference between an ideal optimum at q = 100% and another, arguably more realistic one at q = 22% reflects the impact of high costs of risk reduction on security: if risk management is expensive, the most efficient solutions may be ineffective, and conversely, if risk management is affordable, security solutions may appear cost-efficient, but ineffective. Unfortunately, this discrepancy between what is desirable and what is affordable in security corresponds to all too realistic, virtually daily experience. As the example demonstrates, however, this discrepancy can at least be understood very much in detail in systematic, quantitative ways, using a suitable, quantitative approach to ICT security economics.

VIII. Process Modelling and ICT Security Management

Safety and security planning in large-scale CPT and CII systems can be made very effective by combining scenario-based computer simulations of systems and processes (e. g., Monte Carlo simulations) with numerical estimates of damage probabilities in simulated security incidents. The effectiveness and cost-efficiency of technical, organisational and procedural risk management provisions can thus be assessed quantitatively prior to their implementation. Risk and security management as well as attacks can be modelled as processes. A process model may, in turn, help to identify all the relevant risks attached to a process itself or any further actions triggered by it. In ICT security analyses, it is therefore important to develop a generic process model of attacks against the systems considered. This can be done, in principle, using the systems modelling and simulation techniques indicated above.

IX. Conclusion

From the economic point of view, quantitative risk assessment is central to ICT and infrastructure security management for at least two reasons. First, disruptions of complex systems tend to affect large and diverse areas of public life, simultaneously involving many different individual needs, interests and preferences. Secondly, optimisation of risk management under constrained resources must accordingly be based on quantitative cost and cost-efficiency estimates as well as trade-offs between competing values and preferences. The problems arising here are highly significant for key risk management activities such as resource allocation, the prioritisation of competing or even mutually exclusive management goals, optimal planning and decision-making, and effective and cost-efficient organisation. While these problems have been widely discussed in the classical economic and management literature, risk-based solutions are still rare which are methodologically coherent (i.e. systematic rather than *ad hoc*), operational and broadly applicable at the same time.

As planning and decision support devices, the methods outlined above are suitable for government agencies and public safety and security services, operators of large-scale systems and for the security management of public infrastructures. They offer advantages especially for cost-efficient ICT security planning and procurement.

References

[1] G. Geiger, “Economic Perspectives on Security Management,” *European CIIP Newsletter*, vol. 8, No. 1, pp. 17–19, March – July 2014.

[2] E. Petzel, R. Czaja, G. Geiger, and C. Blobner, „Does lift of liquid ban raise or compromise the current level of aviation security in the European Union? Simulation-based quantitative security risk analysis and assessment,“ presented at the 22nd SRA-E Conference, Trondheim, NO, June 17–19, 2013.

[3] T. Aven and O. Renn, “The Role of Quantitative Risk Assessments for Characterizing Risk and Uncertainty and Delineating Appropriate Risk Management Options, with Special Emphasis on Terrorism Risk,” *Risk Analysis*, vol. 29, pp. 587-600, 2009. DOI 10.1111/j.1539-6924.2008.01175.x.

[4] G. G. Brown and L. A. Cox, “How probabilistic risk assessment can mislead terrorism risk analysts,” *Risk Analysis*, vol. 31, pp. 196–204, 2011. DOI 10.1111/j.1539-6924.2010.01492.x.

[5] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior,* 2nd ed. Princeton, NJ: Princeton University Press, 1947.

[6] C. Thomas and S. C. Maurice, *Managerial Economics*, 10th ed. Boston, MA: McGraw-Hill/Irwin, 2010, ch. 10.

Position papers of the 2014 Federated Conference on
Computer Science and Information Systems pp. 123–130

DOI: 10.15439/2014F460
ACSIS, Vol. 3

Achieving Software Security for Measuring Instruments under Legal Control

Daniel Peters, Ulrich Grottker, Florian Thiel
Physikalisch-Technische Bundesanstalt,
Germany
Email: {daniel.peters, ulrich.grottker, florian.thiel}@ptb.de

Michael Peter, Jean-Pierre Seifert
Security in Telecommunications,
Technische Universität Berlin,
Germany
Email: {peter, jpseifert}@sec.t-labs.tu-berlin.de

***Abstract*—In recent years measuring instruments have adopted general-purpose operating systems to offer the user a broader functionality that is not necessarily restricted towards measurement alone. Additionally the trend to the internet of things from which measuring instruments are not immune, e.g. smart meters and traffic enforcement cameras just to name a few, brings forth security questions.**

In this paper, a flexible software system architecture that can be constructed out of freely available open source software is presented which addresses these challenges within the framework of essential requirements laid down in the Measuring Instruments Directive of the European Union. The system architecture is based on a modular design assuring correct collaboration between modules by encapsulating them in different virtual machines and supervising their communication.

I. Introduction

According to estimations about four to six percent of the gross national income in industrial countries is accounted for by measuring instruments which are subject to legal control [16], e.g. electricity meters, gas meters, etc. and their related measurements. In Germany alone, this corresponds to an amount of 104 to 157 billion Euros each year [16]. Hence, manipulations of measuring instruments' software could have far-reaching financial consequences. Clearly, special measures should be considered to secure such instruments.

In the light that around 98% of the world's computer systems are embedded devices [4], security in embedded systems will inevitably play an important role in computer science. This fact is further supported by the tendency of these systems to become more and more connected over insecure networks, e.g. the internet. Challenges in constructing a secure embedded system arise due to the increase in complexity, which is driven by the growing demand for improved capabilities, the digitization of manual and mechanical functions, and interconnectability.

Nowadays, most of the manufacturers of measuring instruments prefer building their software stacks on general purpose operating systems (GPOSs), like Linux and Windows, due to the wide availability of device drivers, software infrastructure, and applications. These systems were not created with high security-awareness in mind. Their total embedded software content often exceeds 10 million source lines of code (SLOC). The alarming danger becomes clear when knowing that tests place the number of severe bugs in well-written open source software code at the rate of about one per 2 000 SLOC [3]. Such a high amount of erroneous code consequentially increases the vulnerability of these systems to attackers because just one bug could be so severe that a sophisticated attacker is able to run arbitrary code on the measuring instrument. The National Vulnerability Database[1], for example, reveals such bugs weekly.

First of all, one should have a look at the basic lawful requirements a measuring instrument in legal metrology has to meet with respect to security. Here, consumer protection and the certainty of a correct measurement are most important. A consumer must be sure that, for example, a fuel dispenser is not manipulated to charge more than what was fuelled. Hence in Europe, member states denominate institutions, called *notified bodies*, which are responsible to review the measuring instruments before commissioning. Current scrutiny in the laboratory concentrates on the validation of correct measurements from the hardware parts, e.g. physical sensors. Software analysis is hampered by obstacles like proprietary software, where source code cannot be checked. The approval for commission is often given after a "black box" validation of the sensors, by application of some test-vectors at the user interface and the sealing of as many interfaces as possible. The aforementioned fuel dispenser is stated to be not manipulated as long as no seal is broken. This assumption is too optimistic, considering that just one open interface, e.g. an USB-port or WLAN, can allow a sophisticated attacker to run arbitrary code by exploiting a single vulnerability.

The remainder of the paper is organized as follows. Section II provides an introductory part about legal metrology. Section III describes how virtualization can help to construct a suitably secure embedded system for measuring instruments. Section IV together with Section V describe our framework before it is analysed in Section VI. In Section VII, we give a conclusion and describe further work to be done for the final implementation.

II. Legal Metrology

Legal metrology comprises measuring instruments that are employed for commercial or administrative purposes or for

[1] Catalogue of software bugs published by the U.S. National Institute of Standards and Technology, and the U.S. Department of Homeland Security's National Security Cyber Division

measurements which are of public interest. More than 100 million legally relevant meters are in use in Germany [16]. The majority of them are used for business purposes, in particular they are commodity meters for the supply of electricity, gas, water or heat. Other classical measuring instruments, with which the end user comes into contact, are e.g. counters in petrol pumps or scales in the food sector. Measuring instruments are to a large extent also required in the public traffic system. Examples are speed or alcohol meters. The commonality of all these applications is that the person executing or being affected by an official measurement cannot check the determined result, the parties concerned must rather rely on the accuracy of the measurement. Hence, the central concern of legal metrology is to protect and ensure that trust. In this context, legal metrology does a lasting contribution to a functioning economic system by simultaneously protecting the consumers.

The *International Organization of Legal Metrology (OIML)* was set up to assist in harmonising such regulations across national boundaries to ensure that legal requirements do not lead to barriers in trade. Software requirements for this purpose are formulated in the *OIML D 31* document [21].

WELMEC is the European committee to promote cooperation in the field of legal metrology, for example by establishing guides to help notified bodies (responsible for checking the measuring instruments) and manufacturers implement the Measuring Instruments Directive described below.

A. Measuring Instruments Directive

Directive 2014/32/EU of the European Parliament and of the Council [20], which is based on Directive 2004/22/EC [19], known as the *Measuring Instruments Directive (MID)*, are directives by the European Union to establish a harmonized European market for measuring instruments, which are used in different member states. The aim of the MID is to protect the consumer and to create a basis for fair trade and trust in the public interest. The directive is limited to ten types of measuring instruments that have a special economic importance because of their number or their cross-border use. These are: water meters, gas meters and volume conversion devices, active electrical energy meters, heat meters, measuring systems for the continuous and dynamic measurement of quantities of liquids other than water, automatic weighing instruments, taximeters, material measures, dimensional measuring instruments, and exhaust gas analysers. The MID defines basic requirements for these measuring instruments, e.g. the protection against tampering and the display of billing-related readings.

Each measuring instrument manufacturer themselves decide which technical solutions they want to apply. Nevertheless, they must prove to a notified body that their instrument complies to the MID requirements. The notified bodies that must be embraced by the manufacturers are denominated by the member states. In Germany, for example, the *Physikalisch-Technische Bundesanstalt (PTB)* is such a notified body. The PTB is furthermore the German national metrology institute providing additional scientific and technical services, which is why it achieves the demanded technical expertise needed. In general, the combination of technical expertise related to the measuring instruments, competence for the assessment, monitoring of product related quality assurance systems, and experience with European regulations, are required. Additionally, it is of particular importance that the notified body is independent and impartial.

B. WELMEC

WELMEC is the European cooperation responsible for legal metrology in the European Union and the European Free Trade Association (EFTA). Currently, representative national authorities from 37 countries are part of the WELMEC Committee.

WELMEC Working Groups (WG) are established by the WELMEC Committee for the detailed discussion of issues of interest and concerns to WELMEC Members and Associate Members. Currently, there are eight active Working Groups and one of them (WG7) is solely responsible for software questions and issues the *WELMEC 7.2 Software Guide*. As of this writing its current version is WELMEC 7.2 Issue 5 [24], with Issue 6 near its completion. The WELMEC 7.2 Software Guide provides guidance to manufacturers and to notified bodies, on how to construct or check secure software for measuring instruments. Although it is based on the MID and its addressed instruments, its solutions are of general nature and may be applied beyond. The document states that by following this guide, a compliance with the software-related requirements contained in the MID can be assumed.

Before constructing a secure measuring instrument software architecture, it is important to clarify legally relevant parts, because only these parts are critical, while of course, ensuring that non-legally relevant parts do not effect legal ones. According to WELMEC 7.2 all modules are *legally relevant* that make a contribution to or influence measurement results. These modules facilitate auxiliary functions like *displaying* data, *protecting* data, *saving* data, *identifying* the software, executing *downloads*, *transferring* data, and *checking* of received or stored data.

III. Creating a Secure System

An operating system is the essential component for hardware abstraction in a software system. It acts as an intermediary between programs and the hardware and, from the security perspective, plays the most important part in constructing a secure system. Generally, operating system architectures are subdivided into two main designs, the monolithic kernel and the microkernel system architecture[2] shown in Figure 1.

In a monolithic kernel system architecture, the entire operating system is working in privileged mode sharing a single memory space with the system software, such as file systems and complex device drivers with direct access to the hardware -

[2] Often another architecture is mentioned, e.g. Windows is sold as a hybrid kernel architecture, combining aspects of a microkernel and a monolithic kernel architecture. We consider this kernel to be a "smaller" monolithic kernel, because in contrast to a microkernel many (nearly all) operating system services are in kernel space, like in a monolithic kernel.

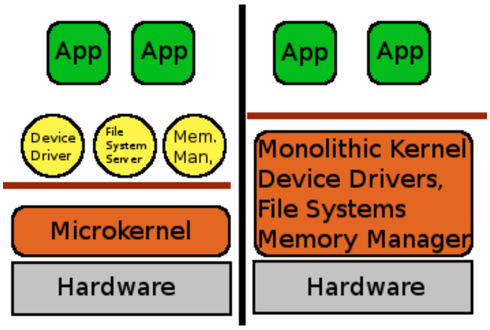


Fig. 1. Comparison between a microkernel (left) and a monolitic kernel design (right)

in Figure 1 the privileged mode is under the horizontal lines. The advantage of this architecture is performance, because user applications are able to access most services, e.g. I/O devices and TCP/IP networking, with a simple and efficient system call. The disadvantage of this approach is the resulting large *Trusted Computing Base (TCB)*. The TCB refers to those parts of a system (software and hardware) which are needed to ensure that it works as expected. Therefore, it must be trustworthy.

In the microkernel design, the microkernel is the only software executed at the most privileged level. Hence, in contrast to a monolithic design, services are implemented in separate processes - in Figure 1 represented as yellow circles. The motivation to place as much functionality as possible in separate protection domains, not running in privileged mode, is to gain stability because, for example, a crash in the network stack that would have been fatal for a monolithic system is now survivable. Consequently, the goal of this architecture is to keep the TCB small and under control as even well-engineered code can have several defects per thousand SLOC [3]. Hence, a bigger system has inherently more bugs than a small system and often a bigger attack surface. For comparison, modern microkernels have around 15K SLOC and less, the monolithic kernel of Linux (version 3.6) at least 300K SLOC to a maximum of 16M SLOC, depending on the configuration.

A. Virtualization

A major drawback in constructing a new software system on a microkernel is that drivers and software libraries available for known GPOSs, e.g. Linux which uses a monolithic kernel design, need to be ported, or in the worst case, completely rewritten. Virtualization seems to be the right solution to incorporate the best implementations of both architectures in a single system. Virtualization can be divided into two main approaches [13]. *Pure virtualization* - sometimes also referred to as *faithful* or *full virtualization* - supports unmodified guest operating systems, running atop another kernel, safely encapsulated. The advantage of this approach is that closed-source operating systems are directly executable. Commodity processors often do not have adequate support for pure virtualization, requiring complex technologies, such as binary translation, to be used [1]. For the second approach, called *para-virtualization*, the guest operating system is presented with an interface that is similar but not identical to the underlying hardware to make virtualization possible. To improve performance the guest operating system is often modified. The approach used depends on the guest operating system that should be employed and the hardware features available.

In the past, embedded systems used to be relatively simple devices and their software was dominated by hardware constraints, which made virtualization unattractive. Nowadays, most embedded systems have the characteristics of general purpose systems and increasingly have the power to actually run applications built for PCs. This power together with the low development costs for a board combined with virtualization technologies like ARM TrustZone [5] makes virtualization in the embedded world - and therefore in measuring instruments - attractive.

A strong motivation for virtualization is security. By running an operating system in its own environment safely encapsulated - a so called *virtual machine (VM)* - the damage of an attack is restricted to this virtual machine because access to the rest of the system can be prohibited, as shown in Figure 2.

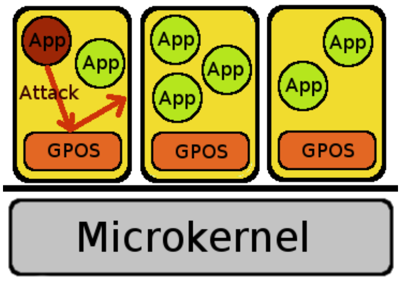


Fig. 2. Left VM's general-purpose operating system (GPOS) is compromised by a pernicious application. Due to isolation the other VMs are not vulnerable

The access is not just restricted to the VM, the whole hardware access can be redirected through the underlying kernel through virtualized device drivers, preventing direct communication with the hardware or any hardware access at all. This isolation can only be assumed if the privileged software managing the virtual machines, called *virtual machine monitor (VMM)* or *hypervisor*[3], is correctly implemented. A microkernel, because of its minimality principle, seems to be a good choice for implementing a hypervisor [9, 17, 15, 10, 8, 22].

B. Policies of a Secure System

Multiple Independent Levels of Security/Safety (MILS) is a high-assurance security architecture based on the concepts of *separation* and *controlled information flow*. The foundation of the system is a small kernel - as used in our design

[3] In this paper we do not differentiate between a VMM and a hypervisor as sometimes done. Both refer to the underlying (micro-)kernel.

- implementing a limited set of critical functional security policies. This special kernel, often called *separation kernel* or *partition kernel*, implements the policies for *information flow control*, *data isolation*, *damage limitation*, and *periods processing* [2].

- *Information flow control* ensures that information cannot flow between partitions unless explicitly permitted by the system security policy
- *Data isolation* ensures a partition is provided with mechanisms whereby isolation within it can be enforced
- *Damage limitation* ensures that a bug or attack damaging a partitioned application cannot spread to other applications
- *Periods processing* ensures that information from one component is not leaked into another one through resources, which may be reused across execution periods

As stated in Section III, modern monolithic kernels and whole GPOSs consist of tens (sometimes hundreds) of millions of SLOC. Hence, they are too difficult and expensive to evaluate. The separation kernel, which, with sometimes no more than 10K SLOC, is small enough to be thoroughly evaluated and mathematically verified for the highest assurance level. The applications managing sensitive data are then built on top of the secure separation kernel. An advantage of this approach is its modularity, allowing software of varying security demands to run on the same microprocessor by means of software partitioning through the kernel. This way the MILS security policies are also stacked, meaning that a module layered atop the separation kernel cannot circumvent the enforced restrictions which the separation kernel defines for it.

IV. Our Framework

Our proposed framework, shown in Figure 3, consists of three parts. The big block on the right is the VM for the non-legally relevant software (N) and the block next to it is the VM for the legally relevant software (L). In the L VM all the computations that are needed for the measurement procedure are executed, e.g. image processing in a traffic enforcement camera. The N VM is only allowed to run software that has no measurement purpose, e.g. like showing the manual or starting a calculator. This strict separation ensures that non-legally relevant software has no effect on the legally relevant one, as postulated in the MID. On the left hand side the smaller blocks form our framework. Their general purpose is to supervise the communication between the L/N VM and the hardware. These modules can be native microkernel processes or VMs themselves. In our opinion the VM approach seems to be the better one, because communication between the individual VMs can take place through network protocols already implemented in the GPOS and the GPOS device drivers can be used [6, 10]. Through the implemented network stacks *virtual private networks (VPNs)* can be created to encrypt communication. The VM approach even allows modules to be transferred out to different computers, creating a distributed system over a network. A disadvantage of this concept is the invalidation of the minimal implementation principle, which should be counteracted by using minimal configurations for the GPOS.

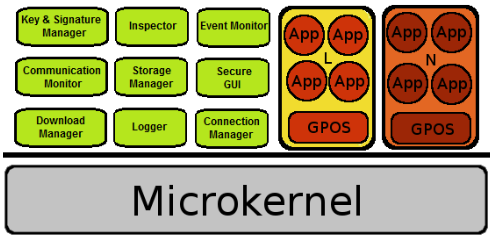


Fig. 3. Framework

Our framework consists of inclusively legally relevant modules, fulfilling legally relevant functions, as demanded in the WELMEC 7.2 Guide and mentioned in Section II-B. The mapping of the functions to the modules is as follows:

- Displaying data: *Secure GUI*
- Protecting data: *Key & Signature Manager*
- Saving data: *Storage Manager*
- Identifying the software: *Inspector*
- Executing downloads: *Download Manager*
- Transferring data over network: *Connection Manager*
- Recording modifications: *Logger*

The *Communication Monitor* redirects queries from and to the I/O devices, e.g. sensors and keyboard. Finally the *Event Monitor* is a watchdog, running availability checks of the software to dynamically spot manipulations.

A. System Requirements

The confidence of a user in a modular system is based on their confidence in the individual system components. Thus an important aspect of the system is the necessity of a secure boot mechanism, where all modules are checked for authenticity and integrity. Only starting from an untampered kernel on an untampered central unit, the kernel can check the other modules for authenticity. In the case of measuring instruments, where seals are used to detect hardware manipulation, the boot-loader should lie on a separate tamper-proof sealed storage unit, that is not readable/writeable for the other system components. The first check then starts at the boot process where, for example, the hash value, e.g. SHA-2, of the kernel binary is calculated and checked with the pre-calculated value stored in the storage unit of the boot-loader. If the two values are identical, the kernel can be loaded and starts with similar tests on the individual modules.

After a successful boot process confidential conversation channels between the VMs must be established. The most important system requirement is a correct microkernel / VMM, that enforces isolation of the VMs and has no covert communication channels. It must be impossible to subvert the VMM or to attack other VMs through a compromised VM. The microkernel needs to assign unique unchangeable identifiers to the virtual network interface controllers of the VMs and must create buffers for every virtual connection the system needs. The buffers are not directly accessible by the VMs, they only simulate a network transmission.

There must be a mechanism to switch from legally relevant to non-legally relevant mode. A hardware switch accessible by the VMs would be an example. If the switch is set to legal mode, every input from devices that can be used for non-legally relevant tasks, e.g. a keyboard, would be redirected to the non-legally relevant VM, and to the legally relevant VM, the other way around. Another method would be to use a touch screen that is divided into legally relevant parts and non-legally relevant ones.

The scheduler implemented must ensure fixed runtimes for every VM to ensure worst-case response times and to minimize the potential for denial-of-service attacks. A measuring instrument is a real time system, meaning that, if within a maximum time frame measuring tasks are not completed, failure has occurred. Therefore, absolute *worst-case execution times (WCET)* for the VMs must be guaranteed. An advantage of our system and of embedded systems in general is their static behaviour. The amount of needed VMs remains constant over the whole execution time, therefore a static schedule can be declared from the beginning. A *temporal partition schedule* should be applied [11], assuring that VMs do not starve, i.e. do not get execution time. Each VM is provided a *window of execution* within the repeating timeline. In our framework some VMs, e.g. the Connection Manager that is responsible for redirecting interrupts, could be split into more than one window of execution. In this way, the system can react faster to input devices.

B. Distributed System View

As already mentioned, we have built a virtual distributed system in which the VMs communicate through a virtual network. The microkernel ensures that the network is reliable, hence for the transport layer protocol we can use *UDP*. For security reasons the network layer protocol used when a VM communicates with the Connection Manager should be *IPsec* in *Transport Mode*. To encrypt the packets IPsec uses the mechanism *Encapsulating Security Payload (ESP)*, encrypting the payload by the symmetric encryption algorithm AES-CBC with the keys, that are being managed by the Key & Signature Manager. Communication between other VMs is not that critical and, therefore, does not need encryption.

For the application layer a protocol must be defined that encapsulates the commands and data. Thereby every VM could use its own protocol or interpretation of payload, e.g. the Storage Manager accepts requests like storing data to disk or getting data from disk, which other VMs do not need. Hence, every VM needs to know the structure of the protocols which other VMs use, if they want to communicate with each other.

Each VM has a server application that listens to a predefined port to receive and afterwards respond to the queries. In our architecture, the VMs fulfil client and server duties because they redirect tasks to and process tasks for other VMs, which makes them so called *servents*. The VMs can be divided into three core layers: the *user interface layer*, the *processing layer*, and the *data layer*. Tasks for the *user interface* are executed by the Communication Manager that redirects I/O from and to devices, the Secure GUI, which displays the graphical user interface (GUI), and the Connection Manager communicating to the outer world. The *processing* is done by the L/N VMs, the Inspector, the Event Monitor, and the Download Manager. Finally the *data* is managed by the Key & Signature Manager, the Logger, and the Storage Manager. If the scope is to construct a real distributed system, the only modules needed on the measuring instrument are the user interface layer VMs and the Event Monitor, all the other modules can be outsourced to other machines.[4]

V. Description of the Individual Modules

By dividing the framework into modules, it can be tailored individually for every measuring instrument. For example, if a measuring instrument does not need a download mechanism, the Download Manager should be removed and if it does not need network access, the Connection Manager is unnecessary. Besides the L VM, the modules every measuring instrument needs, are the Key & Signature Manager, the Inspector, the Logger, the Communication Monitor, and the Event Monitor. A closer look at the individual modules is given below.

A. Event Monitor

The Event Monitor is an autonomous watchdog timer that ensures correctness in the event of a delay that could harm the measurement. In extreme cases, it even deletes or marks measurements as invalid. Additionally, the Event Monitor can advise the Inspector to automatically check the system for integrity, which then reports errors to the Logger or advises the Event Monitor to shut down the system. Every module announces to the Event Monitor in fixed intervals that it is sane. If a module is not responding, the Event Monitor can restart it. For restarting VMs or shutting down the system, the Event Monitor needs special access rights that no other module has.

B. Key & Signature Manager

The Key & Signature Manager is responsible for assuring confidentiality and integrity by managing the public keys of the VMs which want to communicate to the outside world over the Connection Manager. The Key & Signature Manager serves as a certification authority (CA), assigning and dispensing public keys to the corresponding VMs, which are in turn used to negotiate a symmetric key. The manager should have, as every module has, exclusive rights to a portion of the storage device to hold sensitive information. Every VM has its own key database holding the symmetric keys. If a public key gets changed, the manager informs the other VMs to renegotiate a symmetric key. If a key is compromised, the certificate can be invalidated and the respective VM can be prompted to regenerate a key pair and to transmit the new public key

[4] When using an open network reliability is not ensured, therefore TLS should be used as the transport layer protocol, which the Connection Manager should enforce for network communication.

to the manager. Only an authorized entity should be able to command the Key & Signature Manager in this way, even a sealed hardware switch could be possible that needs to be broken to allow the reassigning of keys.

Furthermore, the manager incurs the protection of legally relevant data by holding the keys for file system data encryption and the hash values of the SoftwareIDs for integrity checks at boot and runtime.

C. Connection Manager

The Connection Manager is the only VM with physical network access. All data transmitted from and into the network goes through this module. Hence, the Connection Manager is critical from a security point of view. The manager is a firewall for the system, analysing received data and, according to its rules, redirecting the packets to the appropriate VM. For this purpose, a well-defined protocol must be established. If a packet does not conform to the protocol or the firewall rules, it is discarded.

Another one of its duties is the encryption of legally relevant data in transit, by building up a VPN to only trusted end-points. Data coming from other VMs is itself already encrypted by the symmetric keys the individual VMs have pre-negotiated with the end-points, and cannot be read out by the Connection Manager. It can only be redirected. In this way a compromised Connection Manager is not able to modify data. Non-legally relevant data can be send unencrypted, but firewall rules for outgoing packets should be obeyed.

D. Inspector

The Inspector serves as a remote attestation server for market surveillance. A measuring instrument shall be designed to allow surveillance control by software after the instrument has been placed on the market and put into use. Hereby, software identification shall be easily provided by the measuring instrument. To achieve these requirements the Inspector module is indirectly accessible through the network. The Connection manager redirects the network packets after checking if an authorized person is connected to the device. After connection establishment the Inspector module can advise the Storage Manager to check the file-system and the individual measurements for integrity, to transmit and check the identifications of all modules, to advise the Logger to print out the logging, to check for enough storage capacity and if needed to check the other modules for malware. Finally, it can advise the Key & Signature Manager to invalidate and incur public keys.

E. Download Manager

In legal metrology, legally relevant software can only be updated if the software update was checked prior to download on the device, and afterwards by breaking a seal. Downloads for non-legally relevant parts are allowed without new checking. In our system architecture this is no problem due to the strict isolation. Before legally relevant software is updated, the Download Manager checks if the sealed hardware switch for downloading legally relevant software is set. If this hardware switch is set, measuring must be disabled. For non-legally relevant updates it only must be ensured that the computational time the download mechanism needs, does not disturb correct measuring. Therefore the Download Manager should get a minimum running time, if a non-legally relevant software download is performed.

An upload can take place through different interfaces. If the download is started through the ethernet interface, the Connection Manager receives the request and redirects the download to the partition of the Download Manager through the Storage Manager. If another interface is used, e.g. USB, the data goes through the Communication Monitor. After the download is finished the respective module informs the Download Manager that checks the update on its partition for authenticity and integrity through hash values. If everything is correct the Download Manager advises the Storage Manager to copy the update to the legally relevant or non-legally relevant partition, respectively. Afterwards, the Download Manager reports the download to the Logger and advises the Event Monitor to restart the legally and/or non-legally VM.

F. Communication Monitor

The Communication Monitor supervises the queries from and to the I/O devices. This module ensures that user input and software cannot influence measurement data in an unwanted way, that peripheral devices can only be accessed by legally relevant software, and that the transmission of data from the legally relevant VM to the non-legally relevant one is licit. If an input device is allowed to send data to the non-legally relevant VM, a switch must be set as described in Section IV-A.

This Monitor serves as a kind of firewall for internal communication of the legally relevant VM, blocking packets that do not conform to well-defined rules and checking that confidential data is not transmitted to the non-legally relevant VM. The communication monitor is the only VM that has direct access to the peripheral devices besides the physical network card (accessible by the Connection Manager),the display (accessible by the Secure GUI) and the storage device (accessible by the Storage Manager). Other modules can just communicate with each other through their virtual network cards. An interrupt from an input device is first directed to the Communication Monitor, which in turn translates it to a network package and sends it to the legally relevant VM.

G. Secure GUI

Non-legally relevant output must be unambiguously marked to distinguish it from legal one. The Secure GUI supervises this output to the screen and is the only VM that can write directly to the screen. One possible way is to define selected parts of the screen to the legally relevant software that cannot be changed by the non-legally relevant VM. Another solution is to change the running mode via a hardware switch. Hence, when the switch is set non-legally relevant software cannot write to the screen. In either case the Secure GUI module conducts the buffering for both VMs. For that purpose the legally relevant and the non-legally relevant VMs could each have their own virtual video card driver which redirects requests to the Secure GUI, which in turn visibly separates the output for the user. Another solution is to communicate the screen output through the virtual network cards, by using a well-defined protocol to winnow the screen output data from other message data. By using this solution, the Secure GUI communicates the same way with the VMs as every other module and no extra driver must be written, respecting the minimality principle.

H. Storage Manager

The Storage Manager is the only VM with access to the storage device. Every other module that wants access to its storage partition must send the read and write commands through network packets. The Storage Manager assures strict isolation of the individual module's stored data in use. It checks the module's file permissions, making sure that no illicit manipulation of data takes place. It is also responsible for the encryption of data-at-rest, making the file-system data unreadable if the key is unknown. In the case of errors, the Storage Manager informs the Logger and in extreme cases, e.g. no storage capacity and nothing can be deleted, the Event Monitor to shut down the system.

I. Logger

The Logger is responsible for tracking interventions. On interventions and errors the other modules inform the Logger, by sending him a network packet. This massage is then concentrated to an aligned format and transmitted to the Storage Manager. On demand the Logger sends its log-file to the Inspector, which in turn may send it to the Secure GUI, the Connection Manager or the Communication Manager to transmit it to other devices.

VI. Analysing the System

We analyse the system by showing that the policies that should be implemented according to MILS, to be specific, *data isolation*, *information flow control*, *damage limitation* and *periods processing*, are upheld.

A. Data Isolation

Data isolation is generally enforced by the *component architecture* principle and by using a theoretical verified separation kernel as our VMM. Hence, data isolation between the VMs can be presumed. As we also enforce the *least privileged* principle, our untrusted VMs running the variable software, i.e. the L and N VMs (see Section IV) have no direct access to the outside world. Their software must be checked by an *independent expert validation*, as done by the notified bodies, to ensure data isolation inside the VMs. Data-at-rest security is managed by the Storage Manager, which is responsible for the isolation and encryption of data on the storage device.

B. Information Flow Control

Every VM is provided with a virtual network interface for communication. The separation kernel ensures that no other communication is possible. Information from the outside world, i.e. from peripheral devices and network, go through the Communication Monitor and the Connection Manager, respectively. These VMs, in turn, communicate with the other VMs after carefully checking conformity to a well-defined protocol, as mentioned in Section IV-B, and checking access permissions enforcing the *least privileged* principle. Through unique network card numbers (MAC-address) the communication partners are known and cryptography protects the confidentiality and integrity of their communication.

C. Damage Limitation

The directly exposed VMs are the Communication Monitor and the Connection Manager, because they are the only VMs accessible through peripherals from which attacks could be mounted (NIC, USB, SD, ...). These, in turn, have no direct access to the storage device and no access to measurement data as the *least privileged* principle is applied. To prevent damage, the *minimal implementation* principle must be followed, hence the modules, which are not just processes but VMs with GPOSs, must have minimal configurations. A verified network stack and network interface card driver would drastically reduce the attack vectors.

As described in section III-A, virtualization adheres to the *component architecture* principle, limiting the damage to the respective VM in which it occurs. Additionally, *independent expert validation* by the notified bodies ensures damage limitation in the legally relevant VM.

Another measure taken is monitoring performed by the Event Monitor. If a VM does not conform to its predefined rules, the Event Monitor notices the misbehaviour and takes action, e.g. reloading and restarting the module. In the worst case, the Event Monitor can shut down the system.

D. Periods Processing

The separation kernel must ensure that no hidden channels, through which information could leak out, are present. These could arise due to scheduling because the VMs run consecutively often using the same resources, e.g. shared caches.

Another way of getting secret information is to exploit variations for covert channels or side channel attacks. In general, side channel attacks benefit from variations, e.g. in timing, power consumption, electromagnetic emanation and temperature, to gain information of a cryptosystem. A high-awareness security system should be tested for side channel attacks and should take sophisticated attacks into account, e.g. cache-based side channel analysis [18]. A covert channel exploits the same variations as a side channel attack but is used by malicious processes to exchange information. For example, a VM could try to reduce or extend its execution time, which then can be analysed by the ensuing VM to gain information.

Most of the counter-measures are taken by the separation kernel or can be taken care of by special hardware. Covert channels that arise due to the scheduling policy must be considered separately. As mentioned in Section IV-A, a *partition schedule* should be employed to eradicate timing variations in the scheduling process because every VM has its fixed running window, which cannot be released. To ensure that a VM cannot delay the beginning of the execution time of the ensuing VM, e.g. by a system call before the end of its execution window, a time buffer is needed between the VMs.

VII. Conclusion and Future Work

In this paper, we presented a framework for a new secure system architecture for measuring instruments in legal metrology. We constructed our architecture by analysing the requirements for measuring instruments demanded in the MID and the WELMEC 7.2 Software Guide, combined with methodologies and concepts from high-assurance software systems, i.e. MILS. To harness device drivers and network stacks of general purpose operating systems we came to the conclusion that virtualization is the right solution to combine security with usability. We took a three-pronged approach. First, we separated the legally relevant parts from the irrelevant ones by putting them in different virtual machines. Second, we made sure that their virtual machines have no direct access to I/O devices. Lastly, we constructed a secure framework which provides services to these VMs. This framework, also consisting of separated VMs, monitors the information flow, correctly delegates requests from and to I/O devices, and helps control agencies to verify instruments in commission.

A. Future Work

To show the feasibility of our approach, we have started to build a system atop a L4-microkernel. In our opinion, the L4-microkernel family is a good choice because it is widely used and consists of third generation microkernels. One of these microkernels (seL4) is even fully verified [12] inferring that classical security threats against operating systems, like buffer overflows, null pointer dereferencing, arithmetic overflows, arithmetic exceptions, pointer errors and memory leaks, are not possible. Another positive aspect of many L4-microkernels is that a binary compatible para-virtualized Linux is available. For our demonstrator (PandaBoard Rev. A3) we used L4Linux running atop the open source Fiasco.OC L4-microkernel which yields good results even for real-time applications [14].

Our main goal is to construct a configurable framework, applicable for every measuring instrument under legal control.

References

[1] K. Adams and O. Agesen. A Comparison of Software and Hardware Techniques for x86 Virtualization. *SIGARCH Comput. Archit. News*, 34 (5):2–13, Oct. 2006. ISSN 0163-5964. doi: 10.1145/1168919.1168860.

[2] R. W. Beckwith, W. M. Vanfleet, and L. MacLaren. High Assurance Security/Safety for Deeply Embedded, Real-time Systems. *Embedded Systems Conference*, 2004.

[3] B. Chelf. Measuring Software Quality - A Study of Open Source Software. Coverity, 2011.

[4] C. Ebert and C. Jones. Embedded Software: Facts, Figures, and Future. *Computer*, 42(4), April 2009. ISSN 0018-9162. doi: 10.1109/MC.2009.118.

[5] T. Frenzel, A. Lackorzynski, A. Warg, and H. Härtig. ARM TrustZone as a Virtualization Technique in Embedded Systems. *Proceedings of the Twelfth Real-Time Linux Workshop, Nairobi*, 2010.

[6] H. Härtig, J. Loeser, F. Mehnert, L. Reuther, M. Pohlack, and A. Warg. An I/O Architecture for Mikrokernel-Based Operating Systems. *TU Dresden technical report TUD-FI03-08, Dresden*, July 2003.

[7] H. Härtig, M. Hohmuth, N. Feske, C. Helmuth, A. Lackorzynski, F. Mehnert, and M. Peter. The Nizza secure-system architecture. *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, 2005. doi: http://doi.ieeecomputersociety.org/10.1109/COLCOM.2005.1651218.

[8] G. Heiser. The Role of Virtualization in Embedded Systems. In *Proceedings of the 1st Workshop on Isolation and Integration in Embedded Systems*, IIES '08, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-126-2. doi: 10.1145/1435458.1435461.

[9] G. Heiser, V. Uhlig, and J. LeVasseur. Are Virtual-machine Monitors Microkernels Done Right? *SIGOPS Oper. Syst. Rev.*, 40(1):95–99, Jan. 2006. ISSN 0163-5980. doi: 10.1145/1113361.1113363. URL http://doi.acm.org/10.1145/1113361.1113363.

[10] M. Hohmuth, M. Peter, H. Härtig, and J. S. Shapiro. Reducing TCB Size by Using Untrusted Components: Small Kernels Versus Virtual-machine Monitors. In *Proceedings of the 11th Workshop on ACM SIGOPS European Workshop*, EW 11, New York, NY, USA, 2004. ACM. doi: 10.1145/1133572.1133615.

[11] T. Kerstan, D. Baldin, and S. Groesbrink. Full virtualization of real-time systems by temporal partitioning. *Proceedings of the Sixth International Workshop on Operating Systems Platforms for Embedded Real-Time Applications, Brussels*, 2010.

[12] G. Klein, K. Elphinstone, G. Heiser, J. Andronick, D. Cock, P. Derrin, D. Elkaduwe, K. Engelhardt, R. Kolanski, M. Norrish, T. Sewell, H. Tuch, and S. Winwood. seL4: Formal Verification of an OS Kernel. In *Proceedings of the ACM SIGOPS 22Nd Symposium on Operating Systems Principles*, SOSP '09, pages 207–220, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-752-3. doi: 10.1145/1629575.1629596. URL http://doi.acm.org/10.1145/1629575.1629596.

[13] A. Lackorzynski, A. Warg, and M. Peter. Virtual Processors as Kernel Interface. *Proceedings of the Twelfth Real-Time Linux Workshop, Nairobi*, 2010.

[14] A. Lackorzynski, J. Danisevskis, J. Nordholz, and M. Peter. Real-Time Performance of L4Linux. *Proceedings of the Thirteenth Real-Time Linux Workshop, Prague*, 2011.

[15] M. Lange, S. Liebergeld, A. Lackorzynski, A. Warg, and M. Peter. L4Android: A Generic Operating System Framework for Secure Smartphones. In *Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '11, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-1000-0. doi: 10.1145/2046614.2046623.

[16] N. Leffler and F. Thiel. Im Geschäftsverkehr das richtige Maß. In *Schlaglichter der Wirtschaftspolitik*, Monatsbericht November, 2013.

[17] A. S. Liebergeld, M. Peter, and A. Lackorzynski. Towards Modular Security-Conscious Virtual Machines. *Proceedings of the Twelfth Real-Time Linux Workshop, Nairobi*, 2010.

[18] M. Neve and J.-P. Seifert. Advances on Access-driven Cache Attacks on AES. In *Proceedings of the 13th International Conference on Selected Areas in Cryptography*, SAC'06, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-74461-0. URL http://dl.acm.org/citation.cfm?id=1756516.1756531.

[19] *DIRECTIVE 2004/22/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*, March 2004. Official Journal of the European Union.

[20] *Directive 2014/32/EU of the European Parliament and of the Council*, February 2014. Official Journal of the European Union. doi: 10.3000/19770677.L_2014.096.eng.

[21] *General requirements for software controlled measuring instruments*, 2008. OIML D 31.

[22] M. Peter, H. Schild, A. Lackorzynski, and A. Warg. Virtual Machines Jailed: Virtualization in Systems with Small Trusted Computing Bases. In *Proceedings of the 1st EuroSys Workshop on Virtualization Technology for Dependable Systems*, VDTS '09, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-473-7. doi: 10.1145/1518684.1518688.

[23] F. Thiel, U. Grottker, and D. Richter. The challenge for legal metrology of operating systems embedded in measuring instruments. In *OIML Bulletin*, 52 (LII). OIML Bulletin, 2011.

[24] *WELMEC 7.2 Issue 5 Software Guide*, March 2012. WELMEC European cooperation in legal metrology.

DOI: 10.15439/2014F490
ACSIS, Vol. 3

Measuring Security: A Challenge for the Generation

Janusz Zalewski
Dept. of Software Engineering
Florida Gulf Coast University
Ft. Myers, FL 33965, USA
zalewski@fgcu.edu

Steven Drager
William McKeever
Air Force Research Lab
Rome, NY 13441, USA
Steven.Drager@us.af.mil
William.McKeever.1@us.af.mil

Andrew J. Kornecki
ECSSE Department
Embry-Riddle Aeronautical Univ.
Daytona Beach, FL 32114, USA
kornecka@erau.edu

***Abstract*—This paper presents an approach to measuring computer security understood as a system property, in the category of similar properties, such as safety, reliability, dependability, resilience, etc. First, a historical discussion of measurements is presented, beginning with views of Hermann von Helmholtz in his 19-th century work "Zählen und Messen". Then, contemporary approaches related to the principles of measuring software properties are discussed, with emphasis on statistical, physical and software models. A distinction between metrics and measures is made to clarify the concepts. A brief overview of inadequacies of methods and techniques to evaluate computer security is presented, followed by a proposal and discussion of a practical model to conduct experimental security measurements.**

I. Introduction

When Henry I, the King of England, decreed in the first half of the XII-th century that a yard shall be "the distance from the tip of the King's nose to the end of his outstretched thumb", neither he nor any of his subjects realized that the first standard of measuring length was introduced over the ages [1]. The standard of measuring length (distance) has significantly evolved, from the ancient Egyptian cubit to the one based on physical properties, as captured in a diagram presented in Figure 1.

The current definition of the standard unit of length, a *meter*, involves the speed of light and reads as follows [2]: "the length of the path travelled by light in vacuum during a time interval of 1/299,792,458 of a second." The historical evolution of the humankind's understanding of the unit of length, pictured in Figure 1, shows an amazing path, which led us from a very vague concept to an extremely precise definition based on the speed of light, we have now. It must be noticed, however, that it took us nearly 800 years to straighten the concept, which we now take for granted.

It is the conjecture of this paper that at current stage of understanding how to measure security as a system property, we are at the point comparable to the early days of attempting to measure length. All methods we have are as vague as the one applied by Henry I to defining the unit of length. In this view, the rest of the paper is devoted to clarification of basic concepts of measurement and how they can be applied to building a model of security as a system property that could be used to measuring security.

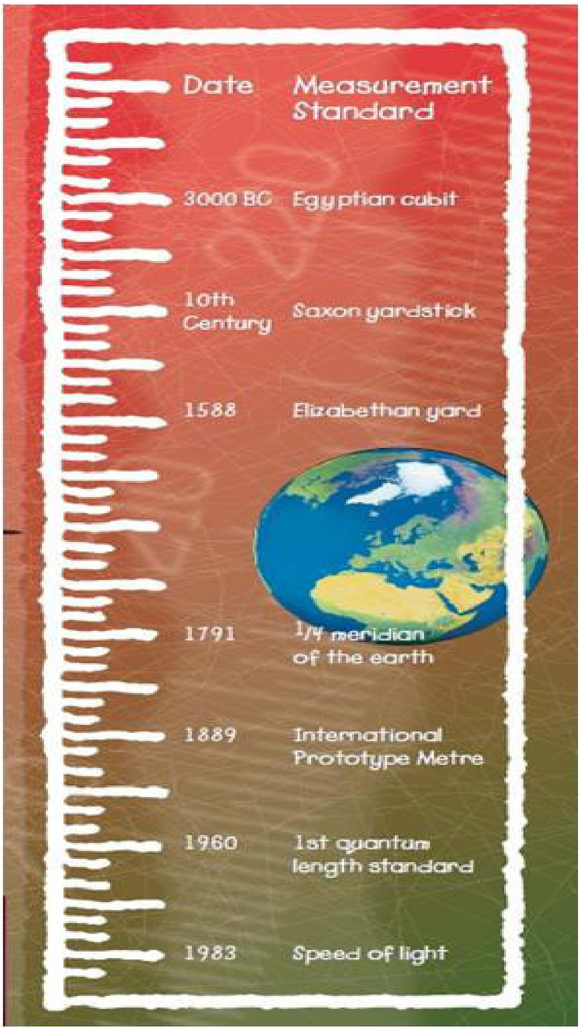


Fig. 1 Evolution of the concept of unit of length [1]

This work was supported in part by the 2014 Visiting Faculty Research Program at the Information Institute of the Air Force Research Labs, in Rome, New York. Distribution A: approved for public release.

II. What is a Measurement?

A. Hermann von Helmholtz Concept of Measurement

Although there are several concepts of measurement, they all seem to converge to the idea formulated in the 19-th century by Herman von Helmholtz, in his groundbreaking work "Zählen und Messen" [3], in which Helmholtz says:

> "The special relation which can exist between the attributes of two objects and which is designated by us by the name equality is characterized by […] Axiom I: If two magnitudes are equal to a third, they are equal to each other."

This statement, which may seem trivial from today's perspective, actually is very constructive and quite distinctly sets the stage for conducting measurements in a way that it determines the following:

- a *property* (called an attribute) of a object to be measured;
- a *standard*, that is, in Helmholtz' words, the third magnitude, to which others are compared; and
- an existence of a *procedure* used to make the comparisons between magnitudes.

This procedure is further characterized by von Helmholtz in the same work, as follows:

> "The procedure by which we put the two objects under proper conditions in order to observe the stated result and to be able to establish its occurrence or its non-occurrence, we shall designate as the method of comparison."

Defining measurement procedure as a method of comparison, von Helmholtz gives several examples of physical quantities that can be measured, by comparison with a standard, including distance, time, brightness, pitch of tone and weight, measured with the use of scales, for which he explains the measurement principle further:

> "… the bodies the weights of which we compare can consist of the most different materials and can be of different form and volume. The weight which we call equal is only an attribute of these bodies discriminated by abstraction."

To summarize, the contribution of von Helmholtz was to make a clear distinction between three factors necessary for a measurement to make sense: a property to be measured, a standard against which comparisons are made, and a procedure to determine how exactly make the comparisons. In modern terms, the standard can be viewed as a *metric*, and measurement procedure relates to a *measure*, that is, measuring instrument.

Overall, von Helmholtz' contribution to measurement theory is much broader than that, and as one of the investigators of his work states, "Zählen und Messen" is "commonly regarded as a turning point between an older concept of measurement in which quantity precedes number and the present concept in which quantity and number are defined separately" [4].

B. Statistical Approach to Measurements

The contribution of von Helmholtz is significant, in terms of the logic of measurement and the associated theory. However, without questioning his work, newer theories treat the measurement processes as statistical in nature. The principal assumption of the statistical approach to measurements is that due to the inherent uncertainties in the measurement process, the result of a measurement always consists of two numbers: the value of the measured quantity and the estimation of the measurement uncertainty with which this value has been obtained (error).

With this view, it is easy to recognize that even the most common notion of measuring time results in two values. When we ask "What time is it?", we obtain a single value, say, 5:30pm, which just happens to be indicated on a watch, but with an implicit understanding that the accuracy of this time value is one minute.

To illustrate the significance of the implications of this concept, one can show an apparently trivial example of measuring the resistance of a DC battery [5]. With a simple battery model consisting of an ideal battery (with zero resistance) and an ideal resistor connected to it in series, the actual measurement circuit will need to have several sources of noise, representing uncertainty. In particular, given some simplifying assumptions, such as linear and time-invariant circuits and neglecting temperature effects, among the factors that cannot be ignored are the following:

- noise caused by battery voltage fluctuations and thermal effects from the resistor
- noise from the voltmeter used in the measurement and its calibration error
- load resistance, including input impedance of the voltmeter.

Combining all these factors leads to a rather significant complication in calculating the battery resistance, making it a non-linear computation of what looked like a simple application of Ohm's Law. Consequently, taking into account uncertainties in the measurement process turns out to be crucial in providing the quality of measurement values.

C. Lessons from Measurements in Physics

To help realize the challenge of measuring properties, one can look closer at the extreme of measuring strictly physical properties (quantities). In addition to length, mentioned above, among physical properties we are most familiar with are time and mass.

The current definition of a second, a metric (unit) of time, involves atomic radiation and reads as follows [2]: "the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom." It must be noticed that this definition, just like the one of a unit of length, quoted in Section I, evolved historically from much less precise definitions and understanding of respective quantities. A historical background can be found at [2].

The metric of mass (its unit), a kilogram, is currently the only physical unit that officially remains defined based on a physical artifact, an international prototype stored in the International Bureau of Weights and Measures, near Paris. However, there is a substantial push towards defining it more precisely, using the number of atoms in a silicon 28 crystal [6]. Developing this new definition has not been fully successful, yet, but (in the context of considering definition of security) it is worth mentioning, why this is so: "The measurement uncertainty is 1.5 higher than that targeted for a kilogram redefinition […]. The measurement accuracy seems to be limited by the working apparatuses." Clearly, any measurement of security must involve the use of measuring devices and assessment of their accuracy.

It may be further argued that security is not a physical property and cannot be measured directly, so even considering such measurements would make little or no sense. In physics, however, there are examples of quantities, which do not measure directly certain properties of matter. One such prominent example is temperature, which is essentially a quantity corresponding to and measuring kinetic energy.

It is clear from these lessons that several points have to be taken into consideration, if one is to develop scientifically based security measurements:

- the process of designing a validated metric of security may take years, if not decades;
- any measures of security must be treated as (physical or mental) measurement devices (instruments), to which regular statistical measurement theory applies
- security is likely to be measured only indirectly, possibly via its inherent components.

D. Software Measurements

With all that has been said in the subsections above, software measurements cause a particular challenge. First of all, software is not a physical quantity, so the question arises can we really distinguish some meaningful software attributes that would have significance regarding the estimation of software quality? In other words, "Analogous to physics, there is the idea whether we can compare a software quality attribute to a norm" [7].

This dilemma has been resolved in two ways. First, we apply a concept of a latent variable, to represent a property that cannot be measured directly but can be estimated using observable attributes (or respective variables representing them) [7]. Second, being aware of our imperfection in approaching the measurements of software, similarly to the evolution of a concept of measuring length and time, we relax the requirement about ultimate quality of software measurements by adopting the rule: "For software then, like time, we want measures that are practical and that we expect will evolve over time to meet the need of the day" [8].

The first publication adopting concepts of measurement theory to software measurements, and comparing them, appears to be [9]. Among the major factors that attention should be paid to in software measurements, the authors list *uncertainty* of the measurement, stating that "improvements in the maturity of software engineering as a truly engineering discipline require for software measurements to include the evaluation of measurements uncertainty whenever measurement results are expressed" [9]. However, they further apply measurement concepts to the function-point analysis, which is a method estimating development effort not the quality of software itself.

III. Can Security Be Measured?

A. Overview

There have been numerous publications in the last decade on security assessment, including books [10-11], research and engineering papers [12-13], government reports [14-16], and Internet sources [17-18], all of them discussing security metrics. However, a vast majority of them deal with metrics at the management level and have very little to do with measurement in a scientific sense of the term, as developed in measurement theory [5,7-8].

What is meant by security metrics in these publications is primarily adherence to standards, whether established industry standards [19-21] or internal company standards [22-23], leading to the assessment of how security policies are executed, for example, by implementing respective processes and auditing them. As one paper defines it [24], security metrics mean "the measurement of the effectiveness of the organization's security efforts over time." While this way of security assessment is beneficial and productive, measuring security as a property of a computing system or software is not particularly well developed.

What is of specific interest in the current paper is not security at the enterprise or the organization level, but rather how security as a computer system property or software property can contribute to protecting information and other resources during system's operation. In this regard, security can be viewed as one specific aspect of system's dependability, the other two aspects being safety and reliability, with one of the earliest papers addressing this issue published over twenty years ago [25].

Such focus on quantitative assessment of operational aspects of security has become more popular in recent years. A thorough survey has been published in 2009 [26], covering quantitative representation and analysis of operational security since 1981, and addressing the question whether "security can correctly be represented with quantitative information?" The major finding of this study was that "there exists significant work for quantified security, but there is little solid evidence that the methods represent security in operational settings." This brings us to the question "Is security measurable?" Before that, it would be even more important to answer a more fundamental question: "Why do we measure?"

B. Why Do We Measure?

There is an often quoted and famous statement by Lord Kelvin [27] that "when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind". Similar motivations guided generations of physicists who gave us all the discoveries thanks to which we are now able to define the basic metrics of physical quantities so precisely. Despite a different nature of software, which is not a material entity, this view of measurement can be also pursued.

Software engineering, being a young discipline, does not have its Lord Kelvin, yet, but one name is certainly worth mentioning. Watts Humphrey deserves quoting, having said [28] that "quality management is impossible without quality measures and quality data. As long as software people try to improve quality without measuring and managing quality, they will make little or no progress." This is the main premise why measurements are critical for any software controlled system. Introduction of rigorous processes based on measurements allows software organizations improve their products, reaching higher capability maturity levels.

For a complete picture, it is worthwhile including a comment from an electrical engineer, published in a systems engineering magazine [29]. After outlining significant deficiencies in current approaches to security and pointing to successes of engineering disciplines, which base their designs on scientific measurements, Fred Cohen writes:

> "As systems engineers, it would be nice to be able to use the same sorts of notions of design for information security as we use for other sorts of design. It would be nice to be able to have standard units of measurement against which we could test things. It would be nice to be able to develop tools for measurement that could be calibrated against the standards, to have a theoretical basis for developing a mathematics and testing it, and then to be able to build up a systems engineering approach to information security like we do in other engineering fields. But first, we need to be able to make meaningful measurements."

With these three sample views, coming from a physicist, a software engineer and an electrical/systems engineer, it becomes quite obvious that the measurements are necessary to improve decision making. In engineering, we have to say it even more strongly, that we measure properties to receive adequate information to determine system's behavior and be able to better control system's parameters. Thus, what has been also expressed in the most recent security research quite clearly [16,30-31], we want to measure security to predict system's behavior and better respond to potential threats or, at least, estimate the associated risks. As one author stated it rather bluntly [32]: "And until we can measure security, we can't improve it."

C. Measurable or Not?

As the quoted author stated in [32], and several other publications expressed as well [33-36], there are significant concerns about the feasibility of security assessment, with some authors even arguing that security as a system property is not measurable [37-38]. In particular, [38] presents a view that any security metric must be a *computable function* mapping a set of features of systems, subject to security concerns, into the real numbers. Under this assumption, introducing a system model with an owner, its adversaries, and an observer, it is claimed that security is non-measurable for the combination of the following three reasons:

- the set of unmitigated weaknesses (vulnerabilities) is not measurable by anyone, including the owner of the system;
- the set of weaknesses (vulnerabilities) known to the observer is not known by the owner of the system and thus is not measurable by the owner; and
- no system owner can know the totality of his adversaries.

Other authors are less skeptical, advocating respective developments [39] and even outlining a number of reasons why measuring security is hard but feasible, including [40]:

- impossibility of testing all security requirements
- interactions between measurements and security
- changes in the environment imposed by adversaries
- subjectivity of the evaluators.

In addition, the same authors also offer some guidance, which are mainly considerations on what should be included in security measurement to make it "more accurate and useful." Among those suggestions several are worth mentioning [40]: (a) building adequate models; (b) using a set of metrics as opposed to a single metric; (b) use different metrics for different purposes; (c) embrace uncertainty.

In the editorial introduction to the special issue of IEEE Security and Privacy Magazine, on the Science of Security [41], the guest editors also express skepticism about measurability of security properties, and anticipate a rough road to reaching this goal, saying that: "We're a long way from establishing a science of security comparable to the traditional physical sciences, and even from knowing whether such goal is even achievable."

The same authors, in another article for this issue of IEEE Security and Privacy [42], referring to "Lord Kelvin's oft-repeated maxim," argue that the essential issue in making progress in security measurement is the existence and usefulness of respective tools. They offer a tip to pursue security metrics saying that two types of metrics can and need to be pursued: "either analytical or experimental."

As pointed out in the aforementioned editorial, we should aim at making the security measurement process comparable to those used in physical sciences. Let's look, then, into the ways the values of security can be assessed using scientific methods, similar to those of measuring physical quantities.

IV. Model for Security Assessment

A. Scientific Approaches to Measurement

Following the observation from [42], for assessment of value of a system property, where there is no science or theory developed, one could try conducting measurement experiments. Nevertheless, if experimental assessment of a system property quantitatively is impossible or difficult, one can also apply simulation. As Glimm and Sharp, for example, point out [43]: "It is an old saw that science has three pillars: theory, experiment, and simulation." This principle is broadly applied in physics, the mother of modern sciences, but it has been also adopted in various ways in computing [44-45].

A closer look at selected computing disciplines reveals that, knowingly or not, this principle has merit, for example, in computer networks. Analytical modeling of network traffic is usually done using queuing theory, measuring network parameters, such as throughput and latency, is done via experiments, and computer simulations use combined computational models to accomplish what cannot be done with theory or live experiments.

However, before any theory, experiment or simulation is developed, putting cards on the table is necessary by developing an initial model of the phenomena whose properties are to be measured. This is the critical first step to conduct the measurement.

B. General Modeling Objectives

Summarizing the discussion thus far, the critical elements in measurements of any property are the following:

1) Clearly identify the *property* to be measured. It is at this point where building a model of the phenomenon is necessary. We use the term "property", although in measurement theory [46], it is called *measurand*.
2) Establish a *metric* to quantitatively characterize the property. Ideally, this would be a unit of measurement, but for vaguely defined properties it can be just a standard against which measurements are applied, or a scale against which the property can be evaluated.
3) Develop a *measure*, which would apply the metric to related objects under investigation. Ideally, this is just a measuring instrument, but for vaguely defined metrics it can be a formula or any other mental device to apply a metric. One important characteristic of a measure should be its linearity, that is, any two identical changes in the property value should be reflected as two identical changes in the measure.
4) Design the *measurement* process to deliver results. An important part of this process is calibration[1] of the measuring device [46], an activity almost never thought of in soft sciences. Another crucial component of this process is the collection and availability of data.
5) Make sure that each instance of a measurement delivers a result composed of the *value* of the measurement and the estimate of its accuracy (an *error*). Alternatively, and consistently with current views in measurement theory, it could be a rage of values designating one value as "measured quantity value" [46].

So knowing all this, now the question is, are we able to develop a model for security measurement? It should embrace all important factors regarding this phenomenon.

C. Architectural Model for Security Assessment

Various types of mathematical models exist to depict physical and mental phenomena, all forming the basis of modern science and engineering. Some of them are continuous, for example, differential equations, but most of those used in computing are discrete, such as queuing theory, finite state machines, network and graph models (Bayesian networks, Petri nets, Markov chains), rule-based systems, etc., including what is called formal methods.

An interesting approach to modeling measurement processes is presented in [9] and involves the IDEF0 process notation specified in the Federal Information Processing Standard [47]. This model is shown in Figure 2 and includes the phenomenon being measured, shown as a *process*, and the *control* unit representing an entity receiving *measurement* results and taking respective *actions*. A number of additional inputs to both the process and the control unit are considered as well.

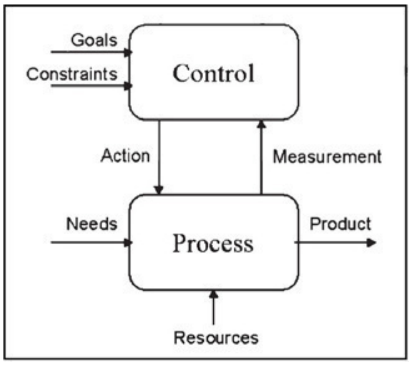


Fig. 2 Modeling of measurement activities according to [9]

We propose the adaptation of this model, making it closer to those used in control theory, which can reflect an impact of external circumstances on computer system's security. Taking the analogy with control engineering, one would only keep interfaces relevant to security during system's operation and, as a result, derive a model of an embedded

[1] The International Vocabulary of Metrology [46] defines *calibration* as "operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication."

controller (or more broadly, a cyberphysical system) subject to security threats as shown in Figure 3.

The diagram shows that multiple controller interfaces to the process, the operator, the network, and the database, are all subject to security threats, forming the attack surface. More importantly, to take the analogy further, just like control theory assumes that the controlled process (a plant) is subject to disturbances, security theory, if one is developed for this model, could assume that known or unknown *threats* play the role of disturbances to the controller. While the control theory can make usually realistic assumptions about the statistical nature of disturbances (e.g., Gaussian noise), it would be challenging – but not impossible – to try and develop a statistical model for threats.

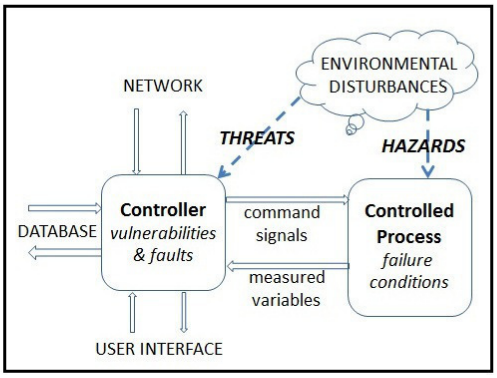


Fig. 3 Generic view of an embedded controller with security threats

In this model, *vulnerabilities* affecting the controller are understood as an "asset or group of assets that can be exploited by one or more threats" [48] or as a "weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat source" [49], while a *threat* can be defined as "a state of the system or system environment which can lead to adverse effects" [50]. Consequently, the disturbances in Figure 3 are an abstraction incorporating all threats relevant to security and play a role in assessing security.

This is our generic model of a cyberphysical system subject to security threats. It has internal vulnerabilities and an attack surface composed of four interfaces. It is a precondition to meet objective (1) from Section IV-B. Now the question is how to define its security property?

D. Definition of the Term

From what has been written in general literature on security measurements, cited earlier in this paper, it is not a simple and unique property, which could be easily identified and defined. Literature on cyberphysical systems is already big and exponentially growing, but is relatively silent on the issue of security measurement [51-52]. We are, therefore, proposing our own approach, which is based on a multifaceted view of security and its measurement.

Looking at definitions of security in established standard glossaries, such as [49] or [53], it becomes immediately clear that in none of these documents security is defined as a system property. For example, one of several definitions in [53] reads as follows: "Protection of information and data so that unauthorized persons or systems cannot read or modify them and authorized persons or systems are not denied access to them" and a corresponding one in [49]: "A condition that results from the establishment and maintenance of protective measures that enable an enterprise to perform its mission or critical functions despite risks posed by threats to its use of information systems."

These are both good definitions, but not for our purposes, because they both refer to security as a *state*, as opposed to *ability*. A definition of security as a system property must imply that one wants to measure it. In this regard, just like for several other properties, the definition should include a phrase "the extent to which" or "the degree to which." Consequently, we propose adopting the definition of security from [53], to read as follows:

> *security*. The extent to which information and data are protected so that unauthorized persons or systems cannot read or modify them and authorized persons or systems are not denied access to them.

What is additionally important and captured well in [53] is the fact that the secure system must be not only protected against threats but also accessible to those authorized.

Having the definition in place, one needs to figure how to assess "the extent" or "the degree" to which the conditions spelled out in the definition are met? The community has adopted several ways to do it. One view, which gained especially wide popularity, is called C-I-A triad, where the acronym comes from the first letters of, what are called, Confidentiality, Integrity, and Availability [54]. The assessment of the degree to which a system is secure is based on meeting the three criteria of the C-I-A triad.

Another broadly adopted view to assess security is based on the STRIDE threat model, which determines the security of the system based on how well it is protected against the following six specific threats: Spoofing Identity, Tampering with Data, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege [55].

We tend to agree with these multifaceted views of assessing security. To use a trivial comparison, measuring security is like assessing patient's health. It is necessary for a doctor to look at more than one parameter to determine a proper diagnosis or to discover a potential disease. Analogically, from the security perspective, we are looking for system health involving multiple indicators, not just one. Additionally, we must take into account that security situation changes over time [56], so the system is dynamical and the security assessment must be continuous.

This merely concludes meeting objective (1) outlined in Section IV-B and gives a background to meet objective (2).

V. Practical Considerations

A. Outline of Establishing a Security Measurement Process

Thus far, we have determined the model for security assessment for one particular class of systems, cyberphysical systems, and defined security as a term. What is necessary in the next step is developing the measurement process (with metrics and measures) for measuring security in the proposed context. This is, of course, an open question and a tremendous challenge.

The model of Figure 3 forms the basis for building a case study for security assessment, by analyzing threats and vulnerabilities. The traditional way of determining and investigating threats is done using attack trees, supported with methods like STRIDE or DREAD as tools for general security analysis [57-58]. In this paper, because of the need for more quantitative approach, an alternative method is suggested, based on assessing the vulnerabilities as per the Common Vulnerability Scoring System (CVSS) [59-60].

To recap what we are looking for, let's repeat that items (2)-(4) from Section IV-B have to be addressed: a *metric*, which for CVSS is a continuous numerical scale; a *measure*, which for CVSS is a set of integrated formulas; and the *measurement* process, which in this case relies on applying the measures to continuously collected data. With these assumptions, the data can be obtained by online checking of the subject entity (embedded device, server, cyberphysical system, etc., for which security is being measured) for known vulnerabilities, as per the Common Vulnerability Exposure (CVE) database [61]. Then calculating the security score based on the CVSS can be accomplished. Several authors have proposed similar methodologies to use CVE/CVSS data [62-63] for security measurement purposes, although without actual theoretical underpinning.

The challenge is the unpredictable nature of threats. Even if one can design countermeasures for existing threats and assess those, there is high likelihood that new, unknown, threats will appear, so one has to design the security system for the unknown, as well as include this type of unpredictability in the computational model for security assessment. The lack of sufficient information for calculating security values suggests building a model based on one of the theories, which deal with uncertainty, for example Bayesian belief networks [64], Dempster-Shafer theory [65], fuzzy sets [66] or rough sets [67].

B. Overview of a Case Study in Aviation

The aircraft internal networks tied with air traffic management and airline operations bring security to the forefront, because they may adversely affect flight safety. This would fit in the model presented in Figure 3. However, the existing aircraft system safety guidance does not address airborne networks and data security issues.

Even though the RTCA committee on Aeronautical Systems Security, SC-216, completed Airworthiness Security Process Specification guidance, DO-326/ED202, in 2010 [68], its work focuses on processes, methods and considerations, staying away from engineering and scientific approach based on measurements and analyses. Often the terminology used in the documents contradicts that used by scientific community. As an example, the aviation community uses term "measures" to represent the procedures, approaches, and tools used to mitigate the security threat (which in common language are "mitigation measures" or "countermeasures").

There is an evident challenge to quantitatively characterize the security properties. Nevertheless, there is a significant practice, established in the safety domain, to use a metric based on ranking applied on an ordinal scale. Clear and unambiguous determination of the metric's scale categories (with assigned ranks) would allow developing effective measures leading to modeling of security for specific assets. However, the measurements would need to be based on the developers' experience and collection of well scrutinized historical data. The resulting measurement (rank or category) would be representing the value, while the accuracy is defined by the category boundaries. Just like in the case described in previous subsection, due to the subjective nature of assessment and lack of sufficient information, it might be useful to explore the application of theories dealing with uncertainty [64-67].

Security property is often assessed indirectly, in terms of risk. Similar to the safety domain, where risk is defined as a combination of probability of hazard and severity of the potential consequences, the security domain also uses this concept. The metrics used for assessing such security aspects as attacker profile, vulnerabilities, operational conditions, or threat conditions, are defined in terms of likelihood (or probabilities). Again, these metrics are more ordinal than numerical. Metrics such as likelihood of attack, impact of a successful attack, level of exposure (vulnerability), are very subjective, ill-defined, and collecting data for them is an obvious challenge. The typical categorization of the attack likelihood is presented below:

- Frequent – anticipated to occur routinely in the life of each asset.
- Probable – unlikely to occur during a routine operation but may occur a few times in the life of an asset.
- Remote – unlikely to occur during its total life but may occur several times in the total life of an entire group of this type of assets.
- Extremely Remote – occurrence not anticipated during its total life but may occur a few times in the total life of entire group of this type of assets.
- Extremely Improbable – occurrence not anticipated during the entire operational life of all assets of this type.

The obvious question is what does it mean "routinely", "unlikely", "not anticipated"? How much is "few" or "several"? There is no agreement on specific numerical values and assessment of these likelihoods is difficult. Similarly, typical categorization of a successful attack's impact or consequence is:

- Catastrophic – loss of system (occurrence of multiple fatalities).
- Hazardous – large reduction in safety margins or functional capabilities (potential serious or fatal injury).
- Major – significant reduction in safety margins or functional capabilities.
- Minor – slight reduction in safety margins or functional capabilities.
- No Safety Effect – no impact on the operational capability of the system.

Again, the questions are: what is "slight", "significant" or "large"?

Using similar categories we can classify vulnerability level of the asset (e.g., highly vulnerable, vulnerable, marginally vulnerable, not vulnerable) and the effectiveness of the applied countermeasures (e.g., highly effective, effective, marginally effective, not effective).

The current trend in aviation security [68] is to use the term "characteristics" to denote "property" used in this paper. The aviation community agrees on the following set of parameters defining security property (S) under specific operational conditions (indicated as O):

- A - likelihood of attack
- V - level of asset vulnerability
- E - effectiveness of applied countermeasures
- I - level of impact upon successful attack.

There has been little discussion on how these parameters should be measured, less even what models are reflecting their interrelations. Considering the discrete and ordinal nature of the above parameters, there is a possibility to create mathematical model of security S in a form of a discrete function:

$$S = f(A, V, E, I, O)$$

Evidently, higher ranks of parameters A, V, and I would have a negative impact and thus decrease the security value, while higher rank of parameter E would have positive impact on security as the system property. Based on historical data and actual assessment of security an attempt can be made to identify the $f()$ function.

VI. Conclusion

This paper presents a view on addressing an enormous challenge of measuring computer security as a system property. Guided by principles of measurements introduced in the 19-th century by Hermann von Helmholtz, as well as by the statistical nature of measurements, and facing some fundamental questions whether security is a measurable property, a high-level model for security assessment is proposed. This model is built exploiting an analogy with a control system, treating threats as disturbances to the controller. The proposed model requires identifying measured property, establish appropriate metric, developing measure and the measurement process, and finally present the results in form of a value with an associated accuracy.

This model can be only as good as the data set to which it can be applied. With a chronic lack of reliable data related to security threats and vulnerabilities, it is proposed to use the National Vulnerability Database [61] and apply to it the Common Vulnerability Scoring Systems (CVSS) [59-60], to derive security assessment using computational methods dealing with uncertainty. Comparing the process of security assessment to the development of measurement standards and processes for physical quantities, such as length or time, it is anticipated that refining and adjusting the concepts of computer security assessment may take decades and in fact is a challenge for the entire generation.

References

[1] National Physical Laboratory, *History of Length Measurement*. Teddington, Middlesex, United Kingdom. URL: http://www.npl.co.uk/educate-explore/posters/history-of-length-measurement/

[2] National Institute of Standards and Technology, *Definitions of the SI Base Units*, Gaithersburg, Maryland, USA. URL: http://physics.nist.gov/cuu/Units/current.html

[3] H. von Helmholtz, Zählen und Messen: erkentnisstheoretisch betrachtet. In: *Philosophische Aufsätze: Eduard Zeller zu seinem fünfzigjährigen Doctorjubiläum gewidmet*. Leipzig, Germany: Fues Verlag, 1887, s. 17-52. English translation: *Counting and Measuring*, New York: Van Nostrand, 1980.

[4] O. Darrigol, "Number and measure: Hermann von Helmholtz at the crossroads of mathematics, physics, and psychology," *Studies in History and Philosophy of Science*, vol. 34, pp. 515–573, 2003.

[5] R.W. Potter, *The Art of Measurement: Theory and Practice*. Upper Saddle River, NJ: Prentice Hall PTR, 2000.

[6] B. Andreas et al., "Counting the Atoms in a ^{28}Si Crystal for a New Kilogram Definition," *Metrologia*, vol. 48, pp. S1-S13, 2011.

[7] H. Zuse, *A Framework of Software Measurement*. Berlin and New York: Walter de Gruyter, 1998.

[8] L.M. Laird and M.C. Brennan, *Software Measurement and Estimation: A Practical Approach*, Hoboken, NJ: Wiley & Sons, 2006.

[9] P. Carbone et al., "A Comparison between Foundations of Metrology and Software Measurement," *IEEE Trans. Instrumentation and Measurement*, vol. 57, no. 2, pp. 235-241, February 2008.

[10] D.S. Herrmann, *Complete Guide to Security and Privacy Metrics: Measuring Regulatory Compliance, Operational Resilience and ROI*. London: Auerbach Publications, 2011.

[11] W.K. Brotby and G. Hinson, *Pragmatic Security Metrics: Applying Metametrics to Information Security*, Boca Raton, FL: CRC Press, 2013.

[12] A. Atzeni and A. Lioy, "Why to Adopt a Security Metric? A Brief Survey." *Quality of Protection: Security Measurements and Metrics*, D. Gollmann, F. Massacci and A. Yautsiukhin, Eds. New York: Springer-Verlag, 2006, pp. 1-12.

[13] J. Bayuk and A. Mostashari, "Measuring Systems Security," *Systems Engineering*, vol. 16, no. 1, pp. 1-14, 2013.

[14] E. Chew et al., *Performance Measurement Guide for Information Security*. NIST Special Publication 800-55 Rev. 1. National Institute of Standards and Technology, Gaithersburg, Maryland, 2008.

[15] W. Jansen, *Directions in Security Metrics Research*, Report NISTIR 7564, National Institute of Standards and Technology, Gaithersburg, Maryland, April 2009.

[16] R. Barabanov, S. Kowalski and L. Yngström, *Information Security Metrics: State of the Art*. Swedish Civil Contingencies Agency, DSV Report No. 11-007, March 2011.

[17] *A Community Website for Security Practitioners*. URL: http://www.securitymetrics.org

[18] G. Hinson, *Seven Myths about Security Metric*. 2006. URL: http://www.noticebored.com/html/metrics.html

[19] *Department of Defense Trusted Computer Systems Evaluation Criteria* (aka *Orange Book*), DoD 5200.28-STD, Washington, DC, December l985.

[20] *Common Criteria for Information Technology Security Evaluation, Parts 1-3*. Documents No. CCMB-2012-09-001, 002 and 003, September 2012. URL: http://www.commoncriteriaportal.org/cc/

[21] ISO/IEC 15408 Information Technology – Security Techniques – Evaluation Criteria for IT Security – Part 1: Introduction and General Models, Geneva, 2009.

[22] N. Bartol et al., *Measuring Cyber Security and Information Assurance. State of the Art Report*. Information Assurance Technology Analysis Center (IATAC), Herndon, VA, 2009.

[23] *Software Security Assessment Tools Review*. Booz Allen Hamilton, McLean, VA, 2009.

[24] D.A Chapin and S. Akridge, "How Can Security Be Measured?" *Information Systems Control Journal*, vol. 2, 2005.

[25] Littlewood, B. et al., "Towards Operational Measures of Computer Security," *Journal of Computer Security*, vol. 2, no. 2-3, pp. 211-229, June 1993.

[26] V. Verendel. "Quantified Security Is a Weak Hypothesis." *Proc. NSPW'09 New Security Paradigms Workshop*, Oxford, UK, 8-11 September, 2009. New York: ACM, New York, 2009, pp. 37-50.

[27] W. Thompson – Lord Kelvin, "Electrical Units of Measurement," Lecture at the Institution of Civil Engineers, London, 3 May 1883, *Popular Lectures and Addresses*, vol. 1, pp. 73-136, 1889.

[28] W.S. Humphrey, *The Watts New? Collection: Columns by the SEI's Watts Humphrey*, Special Report CMU/SEI-2009-SR-024, Software Engineering Institute, Pittsburgh, Penn., November 2009.

[29] F. Cohen, "How Do We Measure Security?" *INCOSE Insight*, vol. 14, no. 2, pp. 30-32, July 2011.

[30] R.M. Savola, "Quality of Security Metrics and Measurements," *Computers & Security*, vol. 37, pp. 78-90, 2013.

[31] J.L. Bayuk, "Security as a Theoretical Attribute Construct," *Computers & Security*, vol. 37, pp. 155-175, 2013.

[32] S.M. Bellovin, "On the Brittleness of Software and the Infeasibility of Security Metrics," *IEEE Security and Privacy*, vol. 4, no. 4, p. 96, July/August 2006.

[33] M.D. Aime, A Atzeni and P.C. Pomi. "The Risks with Security Metrics." *Proc. QoP'08, 4th ACM Workshop on Quality of Protection*, Alexandria, VA, October 27, 2008. New York: ACM, 2008, pp. 65-69.

[34] J. Rosenblatt. "Security Metrics: A Solution in Search of a Problem." *EDUCAUSE Quarterly*, vol. 31, no. 3, pp. 8-11, July 2008.

[35] W. Jansen, W., *Directions in Security Metrics Research*. Report NISTIR 7564. National Institute of Standards and Technology, Gaithersburg, Maryland, 2009.

[36] T. Sree Ram Kumar, A. Sumithra and K. Alagarsamy. "The Applicability of Existing Metrics for Software Security," *Intern. Journal of Computer Applications*, vol. 8, no. 2, pp. 29-33, October 2010.

[37] R. Savola, "On the Feasibility of Utilizing Security Metrics in Software-Intensive Systems," *Intern. Journal of Computer Science and Network Security*, vol. 10, no. 1, pp. 230-239, January 2010.

[38] M.D. Torgersen, "Security Metrics for Communication Systems," *Proc. ICCRTS'07, Intern. Command and Control Research and Technology Symposium*, Newport, RI, June 19-21, 2007.

[39] W.H. Sanders, "Quantitative Security Metrics: Unattainable Holy Grail or a Vital Breakthrough within Our Rich?" *IEEE Security and Privacy*, vol. 12, no. 2, pp. 67-69, March/April 2014.

[40] S.L. Pfleeger and R.K. Cunningham, "Why Measuring Security is Hard?" *IEEE Security and Privacy*, vol. 8, no. 4, pp. 46-54, July/August 2010.

[41] S. Stolfo, S.M. Bellovin and D. Evans, "Measuring Security." *IEEE Security and Privacy*, vol. 9, no. 3, pp. 60-65, May/June 2011.

[42] D. Evans and S. Stolfo, "The Science of Security," *IEEE Security and Privacy*, vol. 9, no. 3, pp. 16-17, May/June 2011.

[43] J. Glimm and D.H. Sharp. "Complex Fluid Mixing Flows: Simulation vs. Theory vs. Experiment." *SIAM News*, vol. 39, no. 5, June 12, 2006.

[44] G. Dodig-Crnkovic, "Scientific Methods in Computer Science." In *Proc. PROMOTE IT 2002, 2nd Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden*, Skövde, Sweden, April 22-24, 2002, pp. 126-130.

[45] R.W. Longman, "On the Interaction Between Theory, Experiments, and Simulation in Developing Practical Learning Control Algorithms." *Intern. Journal of Appl. Math. Comput. Sci.*, vol. 13, no. 1, pp. 101-111, January 2003.

[46] *International Vocabulary of Metrology – Basic and General Concepts and Associated terms (VIM)*. 3rd Edition, Joint Committee for Guides in Metrology, 2012.

[47] *Integration Definition for Function Modeling (IDEF0)*, Draft FIPS Publication 183, National Institute of Standards and Technology, Gaithersburg, MD, December 1993.

[48] *ISO/IEC 27005:2011 Information Technology - Security Techniques - Information Security Risk Management*. International Organization for Standardization, Geneva, 2011.

[49] *National Information Assurance (IA) Glossary*. CNSS Instruction No. 4009. Committee on National Security Systems, 26 April 2010.

[50] *ISO/IEC/IEEE 24765a:2011 Systems and Software Engineering -- Vocabulary*. International Organization for Standardization, Geneva, 2011.

[51] *Foundations for Innovation in Cyber-Physical Systems*. Report of the Workshop held in Rosemont, Ill., March 13-14, 2012. Energetics Inc., Columbia, MD, January 2013.

[52] Steering Committee for Foundations in Cyber-physical Systems. *Foundations for Innovation: Strategic R&D Opportunities for 21st Century Cyber-physical Systems*. National Institute of Standards and Technology, Gaitherburg, MD, January 2013.

[53] *IEEE Software and Systems Engineering Vocabulary*. IEEE Computer Society, Washington, DC, URL: http://computer.org/sevocab

[54] *Standards for Security Categorization of Federal Information and Information Systems*. FIPS Publication 199, National Institute of Standards and Technology, Gaithersburg, MD, February 2004.

[55] M. Howard, D.C. LeBlanc, *Writing Secure Code. Second Edition*, Microsoft Press, Redmond, Wash., 2003.

[56] R. Bejtlich, *The Practice of Network Security Monitoring: Understanding Incident Detection and Response*, No Starch Press, San Francisco, Calif., 2013.

[57] J.A. Ingalsbe, L. Kunimatsu, T. Baten, and N.R. Mead, "Threat Modeling: Diving into the Deep End." *IEEE Software*, vol. 25, no. 1, pp. 28-34, January/February 2008.

[58] D. Dhilon, "Developer-Driven Threat Modeling: Lessons Learned in the Trenches." *IEEE Security and Privacy*, vol. 9, no. 4, pp. 41-47, July/August 2011.

[59] P. Mell, K. Scarfone, and S. Romanosky (Eds.) *CVSS – A Complete Guide to the Common Vulnerability Scoring System. Version 2.0*. 2007. National Institute of Standards and Technology, Gaithersburg, Maryland. URL: http://www.first.org/cvss/cvss-guide

[60] *Common Vulnerability Scoring System Support V2*. National Institute of Standards and Technology, Gaithersburg, Maryland. URL: http://nvd.nist.gov/cvss.cfm/

[61] *National Vulnerability Database Version 2.2*. National Institute of Standards and Technology, Gaithersburg, Maryland. URL: http://nvd.nist.gov/

[62] J.A. Wang et al., Security Metrics for Software Systems, *Proc. ACM-SE '09, 47th Annual Southeast Regional Conference*, Clemson, SC, March 19-21, 2009, Article No. 47.

[63] A. Tripathi and U.K. Singh, On Prioritization of Vulnerability Categories Based on CVSS Scores, *Proc. ICCIT, 6th Intern. Conference on Computer Sciences and Convergence Information Technology*, Seogwipo, South Korea, November 29 - December 1 2011, pp. 692-697.

[64] F.V. Jensen, *An Introduction to Bayesian Networks*, London, UK: UCL Press, 1996.

[65] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press, 1976.

[66] L. Zadeh and J. Kacprzyk (Eds.), *Fuzzy Logic for the Management of Uncertainty*, New York: Wiley & Sons, 1992.

[67] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publishers, 1991.

[68] RTCA DO-326 Airworthiness Security Process Specification, December 2010.

Frontiers in Network Applications, Network Systems and Web Services

SYMPOSIUM SoFAST-WS focuses on modern challenges and solutions in network systems, applications and service computing. The Symposium builds upon the success of Frontiers in Network Applications and Network Systems (FINANS'2012) and 4th International Symposium on Web Services (WSS' 2012) held in 2012 in Wroclaw, Poland. These two events are now integrated into one event to fully exploit the synergy of topics and cooperation of research groups.

The topics discussed during the symposium include different aspects of network systems, applications and service computing. The primary objective of the symposium is to bring together researchers and practitioners analyzing, developing and administering network systems, with particular emphasis on Internet systems. Authors are invited to submit their papers in English, presenting the results of original research or innovative practical applications in the field.

Topics

Topics include (but are not limited to):

- Architecture, scalability and security of Open API solutions,
- Technical and social aspects of Open API and open data,
- Service delivery platforms - architecture and applications,
- Telecommunication operators API exposition in Telco 2.0 model,
- The applications of intelligent techniques in network systems,
- Mobile applications,
- Network-based computing systems,
- Network and mobile GIS platforms and applications,
- Computer forensic,
- Network security,
- Anomaly and intrusion detection,
- Traffic classification algorithms and techniques,
- Network traffic engineering,
- High-speed network traffic processing,
- Heterogeneous cellular networks,
- Wireless communications,
- Security issues in Cloud Computing,
- Network aspects of Cloud Computing,
- Control of networks,
- Standards for Web services,
- Semantic Web services,
- Context-aware Web services,
- Composition approaches for Web services,
- Security of Web services,
- Software agents for Web services composition,
- Supporting SWS Deployment,
- Architectures for SWS Deployment,
- Applications of SWS to E-business and E-government,
- Supporting Enterprise Application Integration with SWS,
- SWS Conversational Protocols and Choreography,
- Ontologies and Languages for Service Description,
- Ontologies and Languages for Process Modeling,
- Foundations of Reasoning about Services and/or Processes,
- Composition of Semantic Web Services,
- Innovative network applications, systems and services

Event Chairs

Furtak, Janusz, Military University of Technology, Poland

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Legierski, Jarosław, Orange Labs Poland, Poland

Luckner, Marcin, Warsaw University of Technology, Poland

Szmit, Maciej, Orange Labs Poland, Poland

Program Committee

Afonso, Joao, Foundation for National Scientific Computing, Portugal

Baghdadi, Youcef, Sultan Qaboos University, Oman

Benslimane, Sidi Mohammed, University of Sidi Bel-Abbès, Algeria

Chainbi, Walid, ENISO, Tunisia

Chojnacki, Andrzej, Military University of Technology, Poland

Ciocoiu, Catalin, Orange Labs Products & Services, France

Cocucci, Osvaldo, Orange Labs Products & Services, France

Dabrowski, Andrzej, Warsaw University of Technology, Poland

Davies, John, Glyndwr University, United Kingdom

Fernández, Alberto, Universidad Rey Juan Carlos, Spain

Frankowski, Jacek, Orange Labs, Poland

Fuchs, Lothar, Institute for technical and scientific hydrology, Germany

Furtak, Janusz, Military University of Technology, Poland

Gaaloul, Walid, Institut Mines Télécom, France

García-Domínguez, Antonio, University of Cádiz, Spain

García-Osorio, César, University of Burgos, Spain

Gibert, Philippe, Orange Labs Products and Services, France

Grabowski, Sebastian, Research and Development Centre Orange Labs Poland, Poland

Kaczmarski, Krzysztof, Warsaw University of Technology, Poland

Kapczynski, Adrian, Silesian University of Technology, Poland

Katakis, Ioannis, National and Kapodistrian University of Athens, Greece

Kiedrowicz, Maciej, Military University of Technology, Poland

Korbel, Piotr, Lodz University of Technology, Poland

Kowalczyk, Emil, Orange Labs, Poland

Kowalski, Andrzej, Orange Labs, Poland

López Nores, Martín, University of Vigo, Spain

Maamar, Zakaria, Zayed University, United Arab Emirates

Macukow, Bohdan, Warsaw University of Technology, Poland

Misztal, Michal, Military University of Technology, Poland

Nowicki, Tadeusz, Military University of Technology, Poland

Rahayu, Wenny, La Trobe University, Australia

Richomme, Morgan, Orange Labs, France

Soler, José, Technical University of Denmark, Denmark

Taniar, David, Monash University, Australia

Wary, Jean-Philippe, Orange Labs, France

Wrona, Konrad, NATO Consultation, Netherlands

Zaskórski, Piotr, Military University of Technology, Poland

Zieliński, Zbigniew, Military University of Technology

Żorski, Witold, Military University of Technology, Poland

DOI: 10.15439/2014F473
ACSIS, Vol. 3

SDN Architecture Impact on Network Security

K. Cabaj
Warsaw University of Technology
Nowowiejska 15/19
00-665 Warsaw, Poland, Email: kcabaj@ii.pw.edu.pl

J. Wytrębowicz
Warsaw University of Technology
Nowowiejska 15/19
00-665 Warsaw, Poland, Email: j.wytrebowicz@ii.pw.edu.pl

S. Kukliński
Warsaw University of Technology
Nowowiejska 15/19 00-665 Warsaw, Poland, Email: kuklinski@tele.pw.edu.pl

P. Radziszewski
Warsaw University of Technology
Nowowiejska 15/19
00-665 Warsaw, Poland, Email: pmr@ii.pw.edu.pl

K. Truong Dinh
Warsaw University of Technology
Nowowiejska 15/19 00-665 Warsaw, Poland, Email: k.truongdinh@stud.elka.pw.edu.pl

Abstract—The Software Defined Networking (SDN) paradigm introduces separation of data and control planes for flow-switched networks and enables different approaches to network security than those existing in present IP networks. The centralized control plane, i.e. the SDN controller, can host new security services that profit from the global view of the network and from direct control of switches. Some security services can be deployed as external applications that communicate with the controller. Due to the fact that all unknown traffic must be transmitted for investigation to the controller, maliciously crafted traffic can lead to Denial Of Service (DoS) attack on it. In this paper we analyse features of SDN in the context of security application. Additionally we point out some aspects of SDN networks that, if changed, could improve SDN network security capabilities. Moreover, the last section of the paper presents a detailed description of security application that detects a broad kind of malicious activity using key features of SDN architecture.

I. Introduction

In this paper we analyse the features of SDN that can be used for improving network security. We do not analyse security of SDN per se, however some mechanisms, that directly protect users, improve the security of the SDN network too. Additional information concerning threats, and ideas how SDN network should be secured, can be found in the Kreutz et al paper [1]. Even though the SDN concept is novel, some articles concerning detection of various kinds of known attacks are already published. Data from an SDN controller allow detection of network scans [2], [3], DoS and DDoS attacks [2], [4], and detection of infected Zombie machines that are part of a botnet [3]. Additionally, SDN networks can be easily reconfigured to pass traffic for inspection by various legacy (not SDN capable) security devices, and next automatically react on an attack detected by one of those devices. An SDN network has the ability to easily add new network functionalities. The functionalities added as specialized applications (atop or inside the SDN controller) have access to each flow forwarded by the network. Moreover, an application co-working with the SDN controller can easily add rules to SDN switches, completely changing flow switching or even changing the content of forwarded packets. The security functions are not packet-based but flow-based, what makes protection more efficient. The centralization of control plane operations gives the ability to correlate events from different network nodes, what enables a new approach to network security. In more detail we will discuss all these issues in the third chapter.

The paper is organized as follows. In the next section key features of SDN networks are presented. The third section presents the impacts of the SDN paradigm on security mechanisms implementation. Section IV describes the concept of an application that utilizes evolved SDN networks in order to achieve more efficient attack detection.

II. KEY FEATURES OF SDN

The most expected and promising aspects of SDN networks are associated with:

- centralization of some network operations that enables to base control mechanisms on global network view e.g. traffic engineering,
- easy and standard way in which applications may interact with the network via so called "North-Bound API",
- easy customization of networks.

Open Networking Foundation (ONF) and International Telecommunication Union (ITU) have been recently working on the standardisation of SDN networks. A high level view of the SDN architecture together with the key principles of SDN networks have been presented by ONF [5]. The SDN controller acts as a network "brain" (see Fig. 1), directly communicates with network applications via North-Bound Interface (Control – Application Plane Interface) to provide network state information from data plane, and to translate requirements and high-level policies from applications to low-level commands via South-Bound Interface (Control-Data Plane Interface). The most popular protocol used today for communication between the SDN controller and network data plane is OpenFlow [6].

□ II. This work has been partially conducted as part of the CoSDN (Cognitive Software Defined Networks) project, which is funded by FNR Luxembourg and NCBiR Poland.

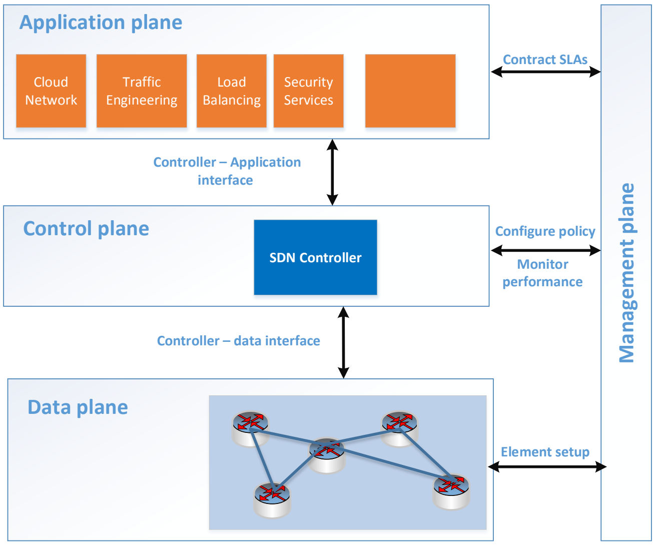


Fig. 1 SDN architecture overview

Fig. 1 describes the general SDN architecture according to network planes and interactions between them. The SDN networks are divided into Data, Control, Application and Management planes.

- Data plane consists of network forwarding elements i.e. switches, which main task is to forward incoming flows to their destinations, making use of routes defined in flow tables.
- Application plane is composed of network service applications, business services, security services, and others, which communicate with the network infrastructure through the SDN controller. They can benefit from abstracted global view of the network according to their own purposes.
- SDN controller is the central point of the network; it gives decisions to the data plane how to forward or modify flows. The controller is also responsible for the transformation of applications' commands to the lower-level communication protocol used by the data plane devices.
- Management plane (according to ONF) is responsible for tasks that are better handled outside the control, application and data planes. It should be isolated and hidden from users. Management entity handles tasks such as setting up the network or configuration of network parameters. It should not be programmable from outside, in order to prevent any kinds of network attacks and to protect the entire network.

North-Bound Interface (NBI) is the interface between applications and the controller. It provides access to network resources from the application level. Although NBI is still not defined, it may provide authorisation and authentication for applications. A role-based authorisation approach has been proposed by Porras et al. [7].

Conflicting rules from applications appear in the controller when some applications require different network behaviour. The conflicts are hard to handle due to the complexity of network control tasks, and an orchestrator is needed. The management in SDN can be implemented in the controller, in the management plane or as a separate application.

The controller is directly connected to network forwarding elements via South-Bound Interface (SBI). OF technology seems to be dominant today, it has been already deployed in many SDN networks [8].

III. IMPACT OF SDN ON NETWORK SECURITY

The SDN concept moves traditional networking from hardware to software with the benefit of automating and simplifying network operations and administration and improving the network performance. As a new technology, SDN is subject to vulnerabilities. In-line with powerful capabilities, which are introduced by SDN networks, various drawbacks of the approach also exist.

Taking into account implementation of security in SDN networks, the following things have to be considered:

- Interactions with network nodes (switches) are performed via the OpenFlow protocol, which has some limitations.
- Global network view: monitoring information is available at the controller; it has the ability to directly manipulate each flow, including possibility to kill it at the source.
- No middleboxes, including NAT (Network Address Translation) or firewalls, are defined in the architecture.

A. OpenFlow limitations in the context of security

Currently, OpenFlow is the most popular protocol used between the SDN controller and network forwarding elements. Although other SDN control interfaces protocol and languages (like FML, Procera, Frenetic [11]) had been designed, the OpenFlow protocol gained the dominant position, and it evolves enabling processing of more and more protocol headers. Despite its popularity, the protocol has some drawbacks. The first big limitation is associated with strict definition of fields that are used by forwarding rules and can be altered by the rules. For example, fields used by IPv6 protocol were not introduced until OpenFlow version 1.3. If hardware or software use older OpenFlow version, the IPv6 traffic cannot be forwarded. The possibility of altering different protocol fields, in an OpenFlow capable switch, enables more complicated functions than forwarding, e.g. NAT or firewall.

It is impossible to implement actions that use other than defined by OpenFlow packet's fields. A solution that resolves this OpenFlow limitation has been recently proposed, it is Protocol Oblivious Forwarding (POF) [12]. A field in POF is defined as sequence of bits starting from a given offset and having a certain length. This provides a flexibility that is the main advantage of the POF paradigm. In contrast, the OpenFlow approach defines a limited number of fields that can be used during matching and altering phase. The POF flexibility allows for rapid development and implementation of new, specialized network protocols, without changes in the switch hardware and without changes in communication with the controller. Moreover, applications that implement new SDN services using POF can make decisions using any part of the ingress packet. This gives a huge flexibility and unimaginable nowadays functionality of future SDN enabled networks. Especially for security related applications, this flexibility can be beneficial, allowing implementation of a “Deep packet inspection” [12], for example finding of known exploits in analysed traffic. Further adoption of POF in network devices hardware structures can lead to performance boost, not achieved in software solutions.

B. Centralized network operations and security

From many years IP network monitoring systems were developed to gather and aggregate data from all network nodes by the use of various protocols, e.g. SNMP [9] or NetFlow [6]. However, access to data related to flow traffic, forwarded in the network, is not easy and not efficient. For example, multiple queries have to be used to gather current state of the network, which could utilize large percentage of management bandwidth and monitored device CPU. In SDN such information is easily accessible at the SDN controller, while in traditional IP networks it has to be sampled on packet basis. From the security point of view, the global awareness concerning all devices in the network is beneficial. An analysis, concerning the whole traffic observed in the network, could lead to detection of distributed attacks, what is impossible on a single network device. Examples of such threats are: a stealth scanning concerning whole network, a set of infected machines, and the advanced persistent threats (APT) [10]. The second big advantage of the global view is associated with the SDN controller's ability to manipulate each flow forwarded through the network. Any reaction to a detected threat can be immediate. From years such reactions were implemented by sending specially crafted reset packets, real-time firewall rules, or by placing security devices in the inline mode, where whole traffic is passed through it. All of those solutions are inefficient and even could lead to degradation of network performance.

The centralized architecture has also some drawbacks. The most evident is associated with performance, when vast amounts of network flows must be analysed in one place. Additionally such architecture introduces a single point of failure. This can lead to congestion of the SDN controller, when many flows are generated in short time, for example as an effect of infection or aggressive scanning attempt. Additionally, as it was described in [1], malicious users can deliberately generate fake traffic to disturb an SDN network. These observations reveal question if all security decisions should be performed directly on the SDN controller. In fact, any successful attack on a centralized controller (may it be a DoS or it's compromise) can result in severe network degradation. Logical distribution of physical controllers might alleviate this danger to some extent, but a meticulous protection of control resources is critical. The protection should cover all aspects – not only technical, but also “social”. In legacy networks this kind of danger is not always critical, and impact of a single security breach can be contained. A carefully thought out network design (e.g. routing policies, OSPF areas, individual link protection) is the solution for security enforcement of today IP networks.

C. Lack of middleboxes in SDN

In currently operating networks many functionalities are implemented in the form of additional devices, so called middle boxes, e.g. NAT devices or firewalls. As was presented in the previous section, OpenFlow limitations prevent implementation of some functions, for example deep packet inspection. Moreover, it is not optimal, from performance point of view, to process all decisions concerning every flow by the SDN controller. A decentralization of some SDN functions, even though it breaks the SDN paradigm, can lead to more efficient and scalable networks. The decentralized functions can be performed locally on SDN switches. This solution demands supplying the SDN switch with an execution environment on which local applications can run. Implementation of security functions in this place has many advantages. As was proven in [14], this can improve detection rate in comparison to traffic observed in aggregated links owned by ISP. Sample description of such solution (PDEE – Programmable Distributed Execution Environment) can be found at [15].

On the other hand, lack of middleboxes in the architecture definition can imply deficiencies in security, however it facilitates end-to-end connectivity, which is needed for some network applications.

A big problem in legacy networks is how to apply and tune traffic engineering rules for tunnelled or encrypted data streams. Without auxiliary mechanisms different tunnelled flows are processed in a unified manner. Of course end devices can alter traffic policies or divide data stream into multiple tunnels in spite of changes in the network core, but only if they are aware of that fact. The SDN approach makes such operations more natural.

A specific kind of middlebox is Intrusion Detection Systems (IDS). Simple IDSs analyze signatures or anomalies, more advanced ones utilize data exploration algorithms. In the next chapter we describe an example of the IDS application, which is suitable for the specifics of SDNs.

IV. SDN SECURITY APPLICATION EXAMPLE: DISTRIBUTED FREQUENT SETS ANALYSER

As was presented in the previous sections, SDN networks enable execution of specialized software that could add new services to the existing networks. Such software could alter simple switches into powerful middle boxes or specialized security devices. What should be emphasized, such change does not need any exchange of deployed hardware and is associated only with addition of new software components. Moreover, the whole network view possessed by the controller allows implementation of advanced methods that could utilize such knowledge. In this section an idea of Distributed Frequent Sets Analyzer (DFSA) systems is presented. The DFSA system takes advantage from experiments with anomaly detection using data mining, and from the features of SDN network. In effect DFSA system could effectively detect broad range of modern network threats.

The idea of data mining algorithms usage for security is based on works [16] [17]. Unfortunately, the integration of three most important elements of such system, i.e. data gathering, data analysis and implementation of actions, is not a seamless process in existing IP networks. For this purpose various mechanisms, techniques and software modules have to be used. Transfers of information between the mentioned elements have impact on the overall security system performance. In contrast, the implementation of such functionality as a security application for SDN network should seamlessly integrate all processes needed for data acquisition, data analysis and reaction to a detected threat.

This section describes a sample security application, which utilizes the well-known features of SDN network and some emerging enhancements that in our opinion can improve the overall network security. Presented concept can be used for detection of various evil or at least anomalous activities performed by network terminals

A. The concept of frequent set analysis

It has been proven that many modern threats, when activated, produce similar patterns in observed traffic. For example network scanning, denial of service attacks (both using one machine or distributed system), botnet activity, sending spam and many more [16]. Discovery of such repeated activity can be a sign of an attack. The data mining techniques could be successfully applied to discover such patterns. Their most important advantage is associated with fact that discovered results are understandable by humans, and can be easily and automatically converted into a response to the detected threat. One of such methods that can detect so called frequent sets is described in [18]. In this method the analysed data is treated as a collection of sets, where each set represents one flow. Each set consists of individual items, which are associated with used protocol, addresses, ports, number of transmitted packets, and overall data size. The number of all sets in the analysed collection that contains this given subset is called support. A frequent set is a subset, whose support is equal or greater than minimalSupport – a parameter defined by the user. Table 1 presents a sample data set with the flows observed by an SDN controller.

TABLE I.
SAMPLE DATA SET USED IN THE EXAMPLE

	Prot.	Src IP	Src Port	Dst IP	Dst Port
1	TCP	10.1.X.X	54333	192.168.Y.Y	80
2	TCP	10.1.X.X	54333	192.168.Y.Y	80
3	TCP	10.1.X.X	54333	192.168.Y.Y	80
4	TCP	172.16.Z.Z	42356	192.168.Y.Y	80
5	TCP	172.16.Z.Z	42456	192.168.Y.Y	8080
6	TCP	172.16.Z.Z	44895	192.168.Y.Y	1080

An operator, who observes the traffic in the network, using his knowledge and experience, sets the minimalSupport parameter. Assumed that in this example we set minimalSuport to 3, various frequent sets can be detected, for example <tcp, *, *, *, * >, <tcp, *, *, *, 80>, <tcp, *, *, 192.168.Y.Y, 80>, <tcp, 10.1.X.X, 54333, 192.168.Y.Y, 80> or <tcp, 172.16.Z.Z, *, 192.168.Y.Y, *>. Asterisks represent items that do not appear in the detected frequent sets. Above frequent sets support initial item sets from table 1 in ranges 1-6, 1-4, 1-4, 1-3 and 4-6 respectively. The most comprehensible patterns are the last two, which are called maximal, due to the fact, that there are no other detected frequent sets in this data set that are supersets of them. Both of these two maximal frequent sets have the support value equal to 3. For further analysis only maximal frequent sets are considered.

Depending on items contained in a frequent set and its support a decision about corresponding flows character can be taken. For example, when detected frequent set has very high support, items are: TCP protocol, TCP port number 25 and only single source address of a desktop machine, it can be assumed with high probability that this machine is infected and actively sends spam. In the other case, when in a long period of observation, frequent set with moderate support value and items associated with TCP protocol, source port and source IP address are detected, presumption that someone used nmap, one of the most known network scanners, can be taken. Detection of such frequent set is caused by this particular scanner implementation, which during scanning uses only one source port. Such analysis can be performed on each network access node and can detect attacker or victim machine directly connected to this particular node. Moreover, the traffic patterns can be observed by a special node that has global view of the network (NetFlow collector, etc.). Such analysis can be performed in a network, which does not support any local detection (the case of existing IP routers). This approach detects scanning activity or other activities that do not appear frequently at a single device, but appear in the whole network.

B. Implementation of DFSA in SDN

The above described method can be efficiently and easily implemented in SDN. Additionally, the ability to reconfigure flow tables in SDN switches can be used for implementation of automatic reaction to detected threats, for example dropping of offending traffic or its degradation. We proposed a hybrid approach, which is based on both local and global analysis of traffic patterns. The Local Frequent Sets Analyser (LFSA) is placed in each SDN switch, and it detects threats that generate vast amounts of traffic; for example infected machines, which send spam or perform DoS attacks. These malicious activities can be efficiently, easily, and without delay detected locally. In effect, the time from detection of an attack to the reaction to it is as short as possible. Moreover, as described in [14], the detection performed at the access node has better accuracy than this performed at aggregation links, for example at ISP management data centre. Additionally, performing all actions locally on one switch reduces the traffic, which is exchanged with the centralized traffic collector and analyser and scales better. However, not all types of attacks can be detected locally. For example stealth scanning of the whole network can be undetected on a single switch, but it can be observed having the global network view. Due to this fact the Global Frequent Sets Analyser (GFSA) is used in our approach.

GFSA is a single module that is placed in the SDN controller or in an SDN application connected to the controller via the NBI interface. LFSA should be implemented at every SDN switch. According to the SDN architecture it is now impossible. However, there exist some enhancements to the concept of SDN, which allow hybrid implementations, for example PDEE [15]. In Fig. 2 the architecture of proposed Distributed Frequent Set Analyser (DFSA) system is presented.

According to the PDEE concept the modules of the DFSA system can be implemented together with the Management applications (M), running at the PDEE execution environments. The LFSA module is placed on each SDN switch and GFSA on the SDN controller. Additionally, DFSA adds to the SDN manager a specialized console module that can be used for the modules configuration (e.g. setting of minimalSupport value used by data mining algorithms). Moreover, the console can be used for reviewing the DFSA logs, which contain information concerning all detected anomalies and performed actions.

For each flow forwarded by the SDN switch, the related set is created and passed to the LFSA. What should be emphasized, cost of data pre-processing is negligible, due to the fact that all needed information is prior collected by the SDN switch. In contrast to mentioned earlier solutions, this computation part is very CPU intensive, as all packets must be directly examined using promiscuous mode or be examined in system firewall and later the firewall processing results are logged and parsed. The process of frequent set discovery is executed at preprogrammed intervals.

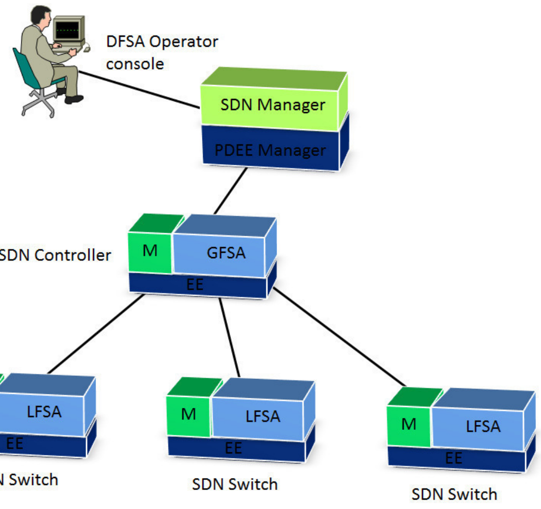


Fig. 2 Architecture of the DFSA system, using the PDEE environment.

Discovered patterns are analysed for symptoms of probable anomalous activity. As it was described in the previous paragraph, the support value of a discovered frequent set and items that the set contains can be used for this purpose. When probability of the malicious activity is very high, the corresponding flows can be stopped by using elements of the discovered frequent set (protocol, source IP, destination port, etc.). For this purpose LFSA inserts additional rules into flow tables of the SDN switch, to drop packets related to malicious flows. In case when malicious activity is not evident, the flow can be degraded and thus slow down attack, but not completely remove it – to preserve the communication under the question. This kind of functionality is nowadays used for flow control (e.g. Random Early Detection), but can be used as a security function as well. When the malicious activity is detected locally, an appropriate action is locally executed. This action not only stops the attack but also protects the SDN controller from DoS attacks directed to it. Additionally, as it has been proven in [13], these attacks can be more accurately detected in the access switch, where the offending machine is connected, rather than in aggregation switch.

At the global level (i.e. at the SDN controller) only the analysis of the aggregated data is performed. The LFSA module, executed in the SDN switch, sends to GFSA implemented in the SDN controller each set that does not appear in discovered frequent sets. This kind of filtering is based on the assumption, that all the activities that generate high volume data traffic are detected at the switch level. There is no need to detect them once again at the central level in the GFSA module. In effect, the GFSA module performs frequent set discovery using aggregated data from all switches. Moreover, data associated with high volume attacks are filtered, and even there is no need to transmit it to the SDN controller. This approach can minimize traffic overhead and CPU cycles at the SDN controller.

Global analysis performed in the GFSA module can be beneficial for detecting massive scanning activity and some stealth scanning techniques. Due to low volume of data associated with those kinds of attacks, observed in a single switch, they cannot be locally detected. However, when data sent from all LFSA modules are received by GFSA and aggregated, detected frequent sets lead to discovery of these threats. Additionally, due to initial filtering that leads to smaller volume of data, aggregation of sets before detection of frequent sets can be performed over a longer time range. After aggregation of frequent sets a discovery is performed. Analysis of discovered patterns is similar to that at the local level. The only change is associated with the manner in which reaction is performed. In this case the SDN controller contacts with all involved SDN switches and installs appropriate rules in their forwarding tables.

V. Conclusion

The SDN paradigm on one hand simplifies the implementation of some security mechanisms, mostly due to centralization of control operations, on the other hand limits distributed approaches. The proposed DFSA system, which uses features of SDN network, can be used for efficient and reliable detection of various network attacks that are observed nowadays in IP networks. The hybrid architecture, which extends the SDN paradigm, allows fast detection of attacks that generate huge amount of traffic, directly at the SDN switch using LFSA modules. Moreover, the GFSA module executed at the SDN controller can be used for detection of attacks that concerns the whole network. Additionally, using SDN network ability to change flow tables, automatic reaction can be implemented as soon as a threat is detected. Implementation of the concept requires a modified SDN, as defined in [15].

Acknowledgment

References

[1] D. Kreutz, F. Ramos, and P. Verissimo, “Towards secure and dependable software-defined networks,” Proceedings of the second ACM SIGCOMM workshop on “Hot topics in software defined networking,” pp. 55-60, 2013, http://dx.doi.org/10.1145/2491185.2491199.

[2] S. A. Mehdi, J. Khalid, and S. A. Khayam, "Revisiting traffic anomaly detection using software defined networking," Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, 2011, http://dx.doi.org/10.1007/978-3-642-23644-0_9.

[3] S. Shin, et al. "Fresco: Modular composable security services for software-defined networks," Internet Society NDSS, 2013.

[4] R. Braga, M. Edjard, and P. Alexandre, "Lightweight DDoS flooding attack detection using NOX/OpenFlow," Local Computer Networks (LCN), 2010 IEEE 35th Conference on. IEEE, 2010, http://dx.doi.org/10.1109/LCN.2010.5735752.

[5] Open Netwok Foundation, “SDN Architecture Overview,” version 1.0, 2013.

[6] N. McKeown, T. Anderson, H. Balakrshman, G. Parulkar, L. Peterson, J. Rexford, S. Shenker and J. Tuner, “OpenFlow: enabling innovation in campus networks,” Sigcomm Comput. Commun., vol. 38, no. 2, pp. 69-74, 2008, http://dx.doi.org/10.1145/1355734.1355746.

[7] Porras, Philip, et al. "A security enforcement kernel for OpenFlow networks," Proceedings of the first workshop on “Hot topics in software defined networks,” ACM, 2012, http://dx.doi.org/10.1145/2342441.2342466.

[8] S. Jain, et al. "B4: Experience with a globally-deployed software defined WAN," Proceedings of the ACM SIGCOMM 2013 conference, http://dx.doi.org/10.1145/2486001.2486019.

[9] J. Case, "A Simple Network Management Protocol (SNMP)," IETF RFC1157, 1990.

[10] C. Tankard, “Advanced Persistent Threats and how to monitor and deter them,” Network security, 2011, http://dx.doi.org/10.1016/s1353-4858(11)70086-1.

[11] A. Doria, J. Salim, R. Haas, H. Khosravi, W. Wang, L. Dong, R. Gopal and J. Halpern, “Forwarding and Control Element Separation (ForCES) Forwarding Element Model,” IETF, 2010.

[12] H. Song, “Protocol oblivious forwarding: Unleash the power of SDN through a future proof forwarding plan,” Sigcomm HotSDN workshop, 2013, http://dx.doi.org/10.1145/2491185.2491190.

[13] A. Nakao, “Deeply programmable network: Emerging technologies for network virtualization and Software Defined Networks,” ITU-T Kaleidoscope, Kyoto, 2013.

[14] S. A. Mehdi, J. Khalid, S. A. Khayam, “Revisiting Traffic Anomaly Detection using Software Defined Networking,” Recent Advances in Intrusion Detection Lecture Notes in Computer Science Volume 6961, 2011, pp. 161-180, http://dx.doi.org/10.1007/978-3-642-23644-0_9.

[15] S. Kukliński, “Programmable Management Framework for Evolved SDN,” IEEE/IFIP Network Operations and Management Symposium, Poland, 2014.

[16] K. Cabaj, K. Szczypiorski, S. Becker, “Towards Self-defending Mechanisms Using Data Mining in the EFIPSANS Framework,” Advances in Multimedia and Network Information System Technologies, Advances in Intelligent and Soft Computing, nr 80, 2010, Springer, pp. 143-151, http://dx.doi.org/10.1007/978-3-642-14989-4_14.

[17] K. Cabaj, Z. Kotulski, P. Szałachowski, et al., “Implementation and testing of Level 2 security architecture for the IIP System,” Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne, SIGMA NOT, vol. LXXXV, nr 8-9/2012, 2012, pp. 1426-1435.

[18] R. Agrawal, T. Imielinski, A Swami, “Mining Association Rules Between Sets of Items in Large Databases,” Proceedings of ACM SIGMOD Int. Conf. Management of Data, 1993, http://dx.doi.org/10.1145/170036.170072.

Position papers of the 2014 Federated Conference on Computer Science and Information Systems pp. 149–156

DOI: 10.15439/2014F41
ACSIS, Vol. 3

The Monte Carlo analysis of the media access time distribution in 802.11n MAC layer

Iwona Dolińska
Akademia Biznesu i Finansów
in Warsaw
ul. Stokłosy 3, 02-787 Warsaw, Poland
Email: i.dolinska@vistula.edu.pl

Antoni Masiukiewicz
Akademia Biznesu i Finansów
in Warsaw
ul. Stokłosy 3, 02-787 Warsaw, Poland
Email: a.masiukiewicz@vistula.edu.pl

Grzegorz Rządkowski
Akademia Biznesu i Finansów
in Warsaw
ul. Stokłosy 3, 02-787 Warsaw, Poland
Email: grzerzad@gmail.com

***Abstract*—Subsequent wireless network standard releases meet better and better the continuously increasing customer requirements in the areas of throughput, coverage and transmission QoS. The network throughput is the key parameter from the user point of view, because the lot of multimedia files are transmitted in such networks. One of the main sources of problems, reducing the transmission quality, is the DCF itself, the basic MAC access method in WLAN. This paper presents the method of analysis of the access to channel in time domain. This method, based on Monte Carlo simulations, allows simulating and presenting the network activity in time domain. The performed simulations show, which components of the MAC scheme absorb most of the time and cause loss of throughput.**

I. Introduction

THE POPULARITY of wireless computer networks has been increasing over many last years. WLANs (Wireless LANs) are built in accoradance with the IEEE 802.11 standard (Wi-Fi). Initially, WLAN networks were treated only as an auxiliary technology augmenting traditional fixed networks. However, with the standard evolution, network throughput and coverage has been still increasing and WLAN networks has become the real competitor to the fixed Ethernet networks. Wireless network standards meet better and better the growing customer requirements for transmission quality [13], [12].

Wireless medium allows for more flexibility of available network locations, but, on the other hand, such networks have also some disadvantages. WLANs are susceptible to disadvantageous interferences from other wireless equipment [8]. Protection against unauthorized access to the network and transmitted data is much more difficult in wireless networks [5]. These networks are based on the shared access to media, what entails many complications in the MAC (media access layer) and the PHY (physical layer) layers. Additionally, in the latest IEEE 802.11 standard version, the MAC layer must deal with the routing tasks in the mesh network (802.11s). Concluding, the wireless standard implementation has much more tasks to perform than fixed network standard implementation.

Wireless networks are very often analyzed in the domain of throughput or in the domain of coverage. Both these domains are important from the point of view of transmission quality providing. But, in the authors opinion, the third domain in which network should be analyzed is the time domain. In the domain of the time we could investigate, if the maximum throughput or the maximum coverage is available during all the transmission time or only during some percentage of time.

One of the key problems, diminishing the transmission quality in the time domain, is the DCF (Distributed Coordination Function) scheme, which is the basic method of MAC access [2], [12]. This scheme was introduced in the first version of the standard and is maintained in the subsequent versions to preserve compatibility [9]. Communication between stations is organized as follows: data is sent in packets, which are always separated with the obligatory time intervals, i.e. interframe spaces (IFS) and backoff (T_{BO}) time. DCF scheme uses DIFS (Distributed Inter Frame Space), SIFS (Short Inter Frame Space) and EIFS (Extended Inter Frame Space) time separators, which values are defined for every standard amendment. Operation of the wireless network in the time domain is shown in Fig. 1. Wireless network operation in the time domain does not depend on the number of stations in the network. Fields marked gray represent data or information packets transmission time, fields marked white represent the dead time T_{DEAD} (silent time), it means the time of obligatory spaces. In the other words, gray color means the channel is active (occupied) and the white color means the channel is not active (not occupied). No data is sent in the dead time. In the authors' opinion, in this time the network bandwidth is wasted. Considering the time domain, the channel throughput is used during a part of time only.

The competition for media access impedes the achievement of an adequate level of QoS (Quality of Service) in WLANs [6], [7], [12] and affects the practical throughput [1], [11], [14]. The MAC scheme is based on the assumption, that before starting trasmission each station must wait until the channel is free. Than it must wait again, when the channel is free during IFS and T_{BO} time. So this communication method reduces capacity in two ways. Firstly, the need to transmit additional headers and preambles in PLCP (Physical Layer Convergence Procedure) sublayer, acknowledgments and control data reduces the bandwidth available for data. Secondly, IFS and T_{BO} time segments separating different packets reduce the bandwidth available for data.

Expected throughput E_{Th}, taking into account most of these parameters, is described as follows [4]:

$$E_{Th} = \frac{K \cdot L_{data} \cdot PRR}{T_{DIFS} + T_{BO}(PRR) + T_{Kdata} + T_{SIFS} + T_{ACK}}. \quad (1)$$

This formula includes three groups of factors. Some factors such as T_{DIFS}, T_{SIFS}, T_{BO} represents the time (wasted or dead time) when all stations are just waiting. T_{ACK} is the time for sending confirmation (acknowledgment frame). The second group of factors concerning the data structure are K, which is the number of aggregated frames and L_{data}, which is the payload carried per frame. The last group concerns selected parameters of the transmission e.g. *PRR* - the packet reception rate. The formula corresponds rather to the average value because it takes into account the average T_{BO}. One cannot calculate real transmission time, because T_{BO} value changes in every communication transaction. Formula (1) doesn't discuss the collisions, *RTS/CTS* (Request to Send/Clear to Send) messages time and the *SIFS* different share in successful transmissions versus collisions.

The *IFS* time intervals, separating every frame transmission, do not reduce the theoretical maximum throughput, but they reduce total throughput per time unit in the time domain (see Fig. 1). During some periods of time the radio channel is not working. Some changes are introduced to improve the dead time problem. In 802.11ac (the new release of 802.11 standard) maximum A-MPDU (aggregate MAC protocol data unit) length is increased from 65 535 octets to 1 048 575 octets. Also, the *RIFS* (reduced interframe space) time interval is not used any more. It is very difficult to estimate the loss of throughput resulting from both mechanisms mentioned above. Some authors suggest that the transmission of whole control data can reduce the throughput available for data by even about 50% [10]. The evaluation of a total throughput decrease is difficult, because it depends on the upper layer transmission type, e.g. UDP or TCP [3].

The authors propose the analysis method of access to the media distribution in time domain based on the Monte Carlo simulations, which allows to simulate the real-time network activity. The description of proposed method and analysis of simulation results are presented in this article. Simulations show, which elements of MAC scheme and in what proportion cause the throughput decrease. The rest of the article is organized as follows. The MAC layer activity in the time domain is described in the second section. All components of data transmission in the time domain are defined there. The third section describes the Monte Carlo method and the way of utilizing this method in WLAN network simulation. The authors define the analysis assumptions in the fourth section. The analysis limitations are described in details in this section. The section five give the simulation results description and in the section sixth this results are concluded.

II. MAC LAYER ACTIVITY IN TIME DOMAIN

As it was mentioned above, DCF is the basic MAC access method in WLAN networks. One communication session involves transmission of one data frame (or control frame) and one *ACK* frame. *ACK* is the obligatory positive acknowledge frame sent by the receiver. In the simulations were used three types of control frames of lenghts as follows: ACK 16 bytes, RTS 20 bytes and CTS 14 bytes.

Before sending a frame, station has to listen to the channel, checking whether the medium is free during *DIFS* time. If yes, backoff algorithm is started [9]. This algorithm differentiates the frame sending start time for many competing stations. The backoff algorithm relies on a draw. Every station has to draw the random value from the (0, CW) interval, where CW means contention window. The TBO (Backoff time) is than calculated as multiplication of this random value and the slot time. After this preparation phase every station has to wait for calculated TBO value. When the two or more stations draw the same shortest value, collision occurs. In this case CW value is doubled (exponentially enlarged) and backoff algorithm starts again. This steps are repeated until successful transmission occurs. The CW may take value between CWmin equal to 15 and CWmax equal to 1023 (dispersion is very big). So, the time needed for one data frame transmission can be described as follows:

$$T_{D1} = \sum_{i=1}^{6} T_i = T_{DIFS} + T_{BO} + T_{PH} + T_{DATA} + T_{SIFS} + T_{ACK}. \quad (2)$$

T_{D1} is the sum of the transmission times of data frame (involves transmission of obligatory PHY header T_{PH} and data T_{DATA}) and *ACK* frame T_{ACK} and all additional time intervals, when the station is waiting all obligatory time spaces (T_{DIFS} and T_{BO} time before sending data frame, T_{SIFS} before *ACK* frame).

In the case of collision, this time is wasted. No data was send, when collision occurred, so the collision time T_{C1} in this communication scheme is equal the data sending time:

$$T_{C1} = T_{D1} = \sum_{i=1}^{6} T_i = T_{DIFS} + T_{BO} + T_{PH} + T_{DATA} + T_{SIFS} + T_{ACK}. \quad (3)$$

An extension of the basic communication schema is a method of the two short control frames exchange at the beginning of communication session [9]. This frames are *RTS* and *CTS*. The *RTS* frame is sent, when station wins the rivalry and it wants to reserve radio channel for the time of sending *RTS/CTS* frames (T_{RTS} and T_{CTS}), data frame, *ACK* frame and all obligatory time spaces. In this case the time needed for one communication session is longer, than in the basic scheme, and can be defined as follows:

$$T_{D2} = \sum_{i=1}^{10} T_i = \\ T_{DIFS} + T_{BO} + T_{RTS} + T_{SIFS} + T_{CTS} + \\ + T_{SIFS} + T_{PH} + T_{DATA} + T_{SIFS} + T_{ACK}. \quad (4)$$

But in the case of collision, the time wasted is much shorter than previously, comparing with (3):

$$T_{C2} = \sum_{i=1}^{5} T_i = \\ T_{DIFS} + T_{BO} + T_{RTS} + T_{SIFS} + T_{CTS}. \quad (5)$$

In the performed simulations this second type of MAC schema has been used (with *RTS/CTS*). All participants (AP and STAs) of wireless communication have to obey the same rules, i.e. they must win the competition to start data transmission. Both AP and station (STA) can be a sender or a receiver. Sample communication session presenting message exchange with the external stations (i.e. station from external networks) is shown in Fig. 2. This packet exchange includes all obligatory time spaces: *DIFS*, *SIFS* and backoff time. In this example we assume, that the first competition was won by STA1 and the second competition was won by STA2. If neither the STA1 nor STA2 wins the competition, data is buffered until success. In the DCF MAC method the T_{BO} parameter has random values (within defined limits) so the authors decided to use the Monte Carlo method in the analysis.

III. The model of 802.11n MAC layer incorporating the Monte Carlo method

Some real world systems contain a lot of elements connected in complex ways. Some of these elements are difficult to define. In such situations, it is sometimes impossible or difficult to define the problem so that it can be solved analytically. We can then attempt to carry out a simulation by using the Monte Carlo method. Monte Carlo methods are a broad class of computational algorithms. They rely on repeated random sampling to obtain numerical results. The word "simulation" means experimenting on the actual system model. Simulation can be made not only, when the problem is difficult to solve analytically. Simulation can be also useful when it is impossible to experiment on a real system. The Monte Carlo method can be used in our situation of the wireless network to model phenomena with significant uncertainty in inputs. It will be mainly used to generate samples from different probability distributions, because we assume that some of our inputs are random variables. Running simulations many times over allows to calculate results similar to those obtained from real experiment.

The basic idea is to construct a mathematical simulation model of the system and introduce into it properly selected data. Then one checks the data of the output of the model and compares them with the available data of the actual system. If the Monte Carlo model meets the expectations and the results appear to be reasonable, then it can be run a sufficient number of times in order to estimate the value of the output to the required level of accuracy. In this way, this method can be used to solve some problems of optimization connected with the proposed model. The very important variable in DCF method is the value which is drawn by each station for T_{BO} computing. Let us recall that, at first each station draws randomly this value from (0, 15) range. Then the station, which drew the smallest number, gains the access to the channel. The DCF scheme could produce the conflicts, when two or more stations have drawn the same smallest number. In this case the draw is repeated, but with a range that is twice longer i.e. (0, 31). In the case of the next conflict, the range will be increased twice again and so on, until reaching the maximum value of 1023. The station can decrease the range to the initial value of (0,15) only after winnig the competition and succesful sending a frame. Because there are a lot of possible solutions it is difficult to find a proper analytical formula.

IV. Assumption for analysis

As it was mentioned above the analysis and simulations were conducted for DCF with *RST/CTS* communication schema (see Fig. 2) with assumption of ideal channel conditions. For the purpose of analysis the wireless network includes AP and stations: one or ten. The main assumption was to simulate the wireless communication as precisely as possible, but with some additional restrictions to make this simulation manageable and workable. Only data frames were taken into account in simulations, not management frames. Every station has always data ready to send. The DCF schema is slightly simplified. It means that retransmission issues, Packet Reception Rate and EIFS are not taken into account. Each transmission is independent session, for wich the new T_{BO} value is chosen. We simulate successful transmissions and collisions, depending on the drawn TBO value for every station. In the case of collision CW value is exponentially enlarged. In the case of successful transmission CW value is reset to minimum. In both these cases TBO is drawn again before next transmission.

Table I presents the most important parameters of radio channel and frames and defined time intervals, used in simulations. The authors assumed that the 2.4 GHz band is used with 20 MHz radio channel width. The data and ACK frame rate is 26 Mbit/s which is the middle rate value for the SISO (single input single output) type of channel. The RTS and CTS frames rate is 6.5 Mbit/s [9]. One communication cycle has contained 200 sessions. The next assumption is that in whole communication cycle the frame length is the same.

In the following formulas the value n describes the number of communication sessions, n_1 - the number of successful sessions, and n_2 - the number of collisions, so $n_1 + n_2 = n$. Transmission time of every DCF schema elements has been calculated, based on parameters showed Table I and control frame lengths. Frame transmission time can be calculated as follows:

$$T_{transmission} = T_{PH} + 8 \cdot N/C, \quad (6)$$

TABLE II
TRANSMISSION TIMES FOR DIFFERENT FRAME LENGTHS AND SOME STANDARD THROUGHPUT VALUES

Throughput [Mbit/s]	**Frame 1540 [bit]**	**Frame 1540 [bit] + preamble**	**Frame 2346 [bit]**	**Frame 2346 [bit] + preamble**
	Transmission time [μsec]			
6,5	1895,38	1927,38	2887,38	2919,38
13	947,69	979,69	1443,69	1475,69
19,5	631,79	663,79	962,46	994,46
26	473,85	505,85	721,85	753,85
39	315,90	347,90	481,23	513,23
52	236,92	268,92	360,92	392,92
58,5	210,60	242,60	320,82	352,82
65	189,54	221,54	288,74	320,74
300	41,07	81,07	62,56	102,56

TABLE III
THE CHOSEN CONTROL FRAMES TRANSMISSION

Throughput [Mbit/s]	**CTS 14 [bit] + preamble**	**RTS 20 [bit] + preamble**	**ACK 20 [bit] + preamble**
	Transmission time [μsec]		
6,5	49,23	56,62	49,238
26	36,31	38,15	36,31

TABLE I
THE CHOSEN PARAMETERS OF DCF AND RADIO CHANNEL

Name	**Value**
Band	2.4 GHz
Radio channel width	20 MHz
RTS, CTS frame rate	6.5 Mbit/s
Data and ACK frame rate	26 Mbit/s
CW range	15-1023
Slot time	20 μsec
SIFS	10 μsec
DIFS	50 μsec
PLCP Preamble	32 μsec
Ethernet frame length	1540 bytes
Maximum frame length (without aggregation)	2346 bytes

where N is data amount in bytes and C is the transmission rate in Mbit/sec. Transmission time of different frame lengths is calculated for a spectrum of possible throughputs. The results are presented in Table II and in Table III. Two different packet lengths are used in the simulation. The length of the typical Ethernet packet is 1540 bytes, while the length of the maximal possible non aggregated packet in the WLAN networks is 2346 bytes. Finally in performed simulations following frame transmission times are used for frame with preamble: for 1540 bit time is equal to 505,85 μsec, for frame 2346 bit time is equal to 721,85 μsec, for CTS 14 bit time is equal to 49,23 μsec, for RTS 20 bit time is equal to 56,62 μsec and for ACK 20 bit time is equal to 36,37 μsec.

For the purpose of analysis some parameters have been defined to describe phenomena on the timeline. The basic information is the dead time, i.e. time, when all users are waiting, although the radio channel is idle. In the case of communication sessions, as showing formula (4), the dead time is the sum of T_{DIFS}, 3 T_{SIFS} and T_{BO} (see Fig. 2). In the case of collisions the dead time is the sum of T_{DIFS}, T_{SIFS} and T_{BO}, so it is a little shorter, like showing formula (5). The dead time can be described, as follows:

$$\begin{aligned} T_{DEAD}(n) = \sum_{i=1}^{n_1} T_{DEADD2i} + \sum_{i=1}^{n_2} T_{DEADC2i} = \\ n_1(T_{DIFS} + 3 \cdot T_{SIFS}) + \sum_{i=1}^{n_1} T_{D2BOi} \\ + n_2(T_{DIFS} + T_{SIFS}) + \sum_{i=1}^{n_2} T_{C2BOi}. \qquad (7) \end{aligned}$$

Second parameter defines the percentage of dead time in one communication cycle and could be described as follows:

$$T\%_{DEAD}(n) = \frac{T_{DEAD}(n)}{\sum_{i=1}^{n_1} T_{D2i} + \sum_{i=1}^{n_2} T_{C2i}} \qquad (8)$$

when T_{D2i} means the time of one successful session and T_{C2i} means the time of one collision session. Next three parameters define the percentage of T_{DIFS}, T_{SIFS} and T_{BO} in one communication cycle:

$$T\%_{DIFS}(n) = \frac{n \cdot T_{DIFS}}{\sum_{i=1}^{n_1} T_{D2i} + \sum_{i=1}^{n_2} T_{C2i}} \qquad (9)$$

$$T\%_{SIFS}(n) = \frac{n \cdot 3 \cdot T_{SIFS} + n_2 \cdot T_{SIFS}}{\sum_{i=1}^{n_1} T_{D2i} + \sum_{i=1}^{n_2} T_{C2i}} \qquad (10)$$

TABLE IV
VARIANTS OF SIMULATIONS

Variant	Number of stations	Length of data packet [bytes]
1	1	1540
2	1	2346
3	10	1540
4	10	2346

TABLE V
DEVIATION OF T_{DEAD} AND T_{BO} VALUES

Time [μsec]	T_{DEAD}	T_{BO}
Average values	98-193	20-250
Maximal values	320-900	100-400

$$T\%_{BO}(n) = \frac{\sum_{i=1}^{n1} T_{D2BOi} + \sum_{i=1}^{n_2} T_{C2BOi}}{\sum_{i=1}^{n_1} T_{D2i} + \sum_{i=1}^{n_2} T_{C2i}} \qquad (11)$$

Additionally, for T_{DEAD} and T_{BO} the average values and the maximum values for every simulation cycle and distribution as a function of the number of communication sessions have been calculated.

V. SIMULATION RESULTS

The simulations were carried out using spreadsheet and the Monte Carlo statistical method. Four simulation variants were used and each variant was characterized by two parameters: the number of stations and the length of the data packet. Variant parameters are presented in the Table IV. The number of stations equal to 1 means that we analyze the network consisting of one AP and one station.

The maximal and average values of the T_{DEAD} and its components (T_{SIFS}, T_{DIFS} and T_{BO}) both in microseconds and as their ratio to the total transmission time ($T_{d2} + T_{c2}$) were calculated. The sum of T_{SIFS} and T_{DIFS} is constant and could be represent by two values. The first value equals 80 microsec (one $DIFS$ and three $SIFS$) and it occurs, when the transmission is completed. The second value of 60 microsec (one $DIFS$ and one $SIFS$) is the result of the collision. Dispersion of T_{DEAD} is mainly the result of the dispersion of T_{BO}. The maximal and average deviation of the T_{DEAD} and T_{BO} values is presented in the Table IV. We can observe, that dispersion of values is very big, what results from CWmax dispersion.

The ratio of time distribution of average and maximal values of T_{DEAD} and its components to the summary transmission time is presented in Fig. 3. The relative ratios (in %) of different components are not constant due to the influence of two factors. Collisions diminish the active time, because there is no data transmission during collision. However, the influence of the T_{BO} value on the T_{DEAD} value is more significant and exists in both situations: when the transmission is successful and when there is a collision. The maximal value of T_{DEAD} (e.g. 900 microsec) could exceed the time necessary for the packet transmission for both packet sizes used in analysis (1540 and 2346 bytes), when the pure data transmission time is respectively 505 and 753 microsec.

T_{BO} is the dominant component of T_{DEAD}, when the network consist of one AP and one station. The situation changes, when the network consist of one AP and 10 stations. T_{DIFS} is the dominant component for average values while T_{BO} is dominant for maximal values. The distribution of relative (in %) ratio of T_{DEAD} and its components to the summary time is presented in Table VI. The presented results indicate that there is:

- Significant difference between maximal and average value especially for T_{BO},
- Quite stable and recurrent ratio (in %) of T_{DEAD} in the summary time for all analyzed variants,
- Increase of average T_{DIFS} influence for AP + 10 stations variant.

Significant deviation between average and maximal value of different dead time components shows, that there is a large difference of single transmission session parameters. The large spread of QoS of the transmission occurs and there is no simple method to predict high value delay resulting from high T_{BO} value.

The distribution of the collision number as a function of the variant of the analysis is shown in Fig. 4. There is a significant increase of the collision number, when the network consist of one AP and 10 stations. The collision probability (ratio of collision in one simulation to 200 i.e. the number of sessions in one simulation) is about 6-7% for the network configuration consisting of one AP + one station. This means that we have approximately 12-14 collisions per 200 sessions. The collision probability increases for the network consisting of one AP + 10 stations and reaches the value of 27% what means 54 collisions per 200 sessions. The increase of collision number for the network consisting of one AP and 10 stations doesn't produce any significant increase neither of average nor of maximal dead time values. The distribution of the dead time value for 200 sessions within one simulation is presented in Fig. 5. We could conclude, that average and maximal value are recurrent, while the actual transmission conditions could have a large dynamism. This is caused by the DCF scheme itself.

VI. CONCLUSIONS

Typical analysis of the DCF scheme in literature concerns rather the possible achievable throughput, which is the most important parameter from the point of view of the best transmission QoS. There are however some root phenomena in the time domain.

Simulation results show that :

a/ T_{DEAD} could reach very high value for single transmission (even 900 microsec, the actual values for present simulation have a significant deviations), while the average and maximal values for repeated simulations are quite stable,

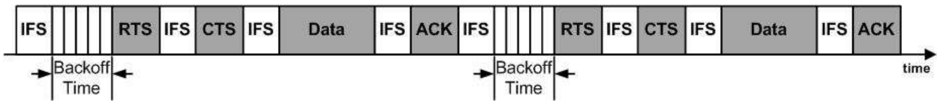


Fig. 1. The activity (gray) and lack of activity (white) in the WLAN radio channel. Source: own preparation.

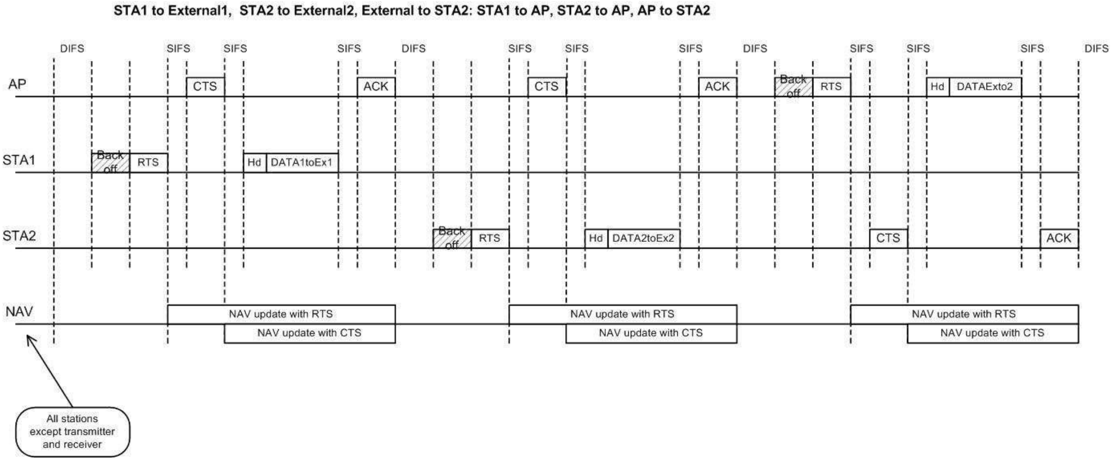


Fig. 2. The example of communication with external station (not belonging to this network). Source: own preparation.

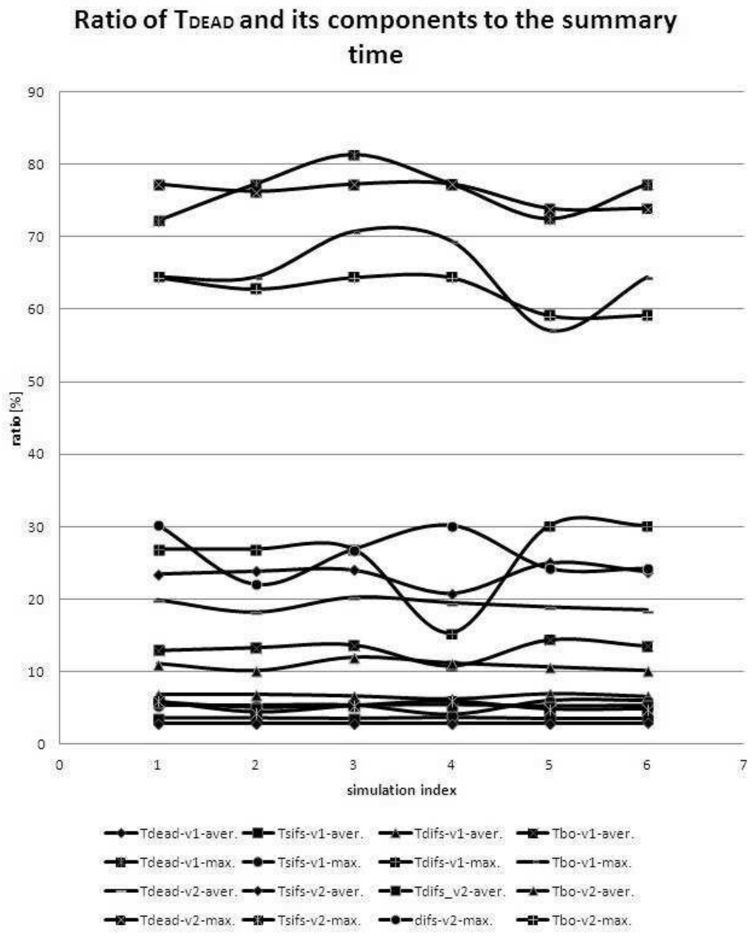


Fig. 3. Ratio of T_{DEAD} and its components to the summary time. Source: own preparation.

TABLE VI
DISTRIBUTION OF T_{BO} AND ITS COMPONENTS IN THE SUMMARY TIME

Analyze variant	Type of value	$T\%_{DEAD}$	$T\%_{BO}$	$T\%_{DIFS}$	$T\%_{SIFS}$
1	Average	20.8-25.0	10.8-14.4	6.3-6.9	3.6
1	Maximal	72.3-81.3	57.0-70.0	15.0-30.0	4.0-6.0
2	Average	18.0-20.0	10.0-12.0	5.3-5.8	2.8
2	Maximal	73.0-77.0	59.0-64.5	24.0-30.0	4.4-6.0
3	Average	21.8-23.7	5.4-7.0	11.9-12.7	4.4
3	Maximal	60.0-71.0	37.0-54.0	30.1	6.0
4	Average	18.5-19.4	4.4-5.2	10.5-11.3	3.6
4	Maximal	60.0-75.1	37.6-60.1	30.1	6.0

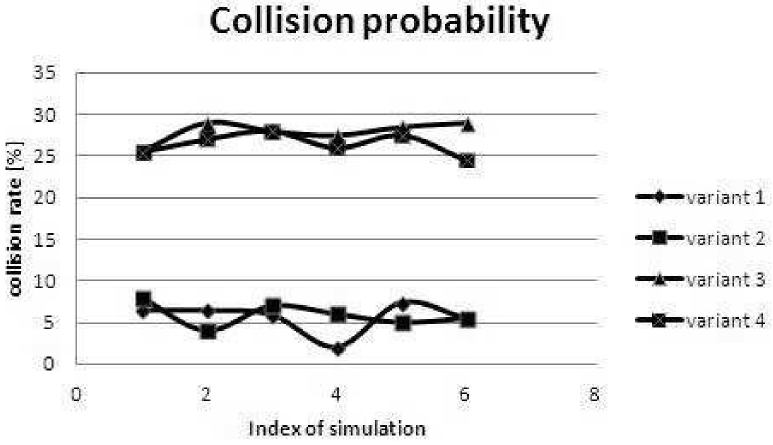


Fig. 4. The collision number distribution (for 4 analyze variants). Source: own preparation.

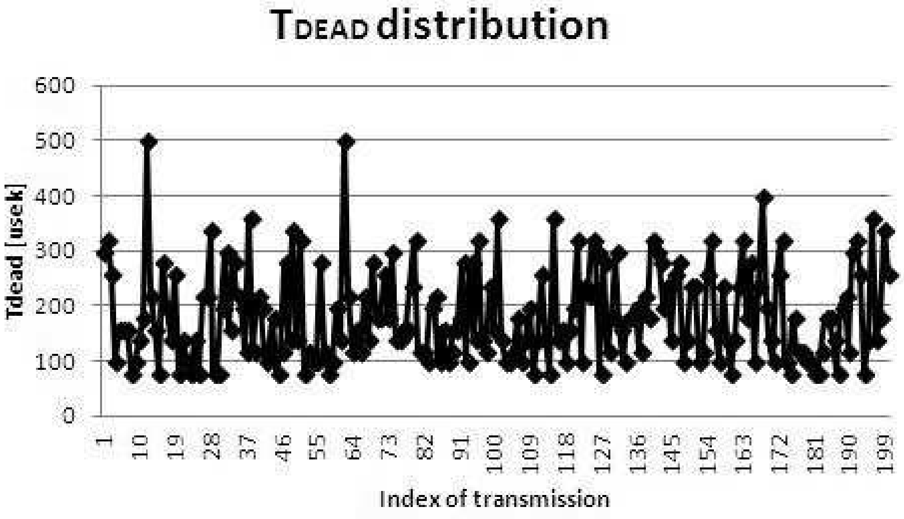


Fig. 5. Distribution of T_{DEAD} for 200 sessions within one simulation (variant 1). Source: own preparation.

b/ the most important component of the dead time is T_{BO}, however for the variant with one AP and 10 stations average values for T_{DIFS} are higher than for T_{BO}.

The final conclusions are as follows:

- the presently used DCF scheme produces quite high differences between QoS of single transmission conditions,
- the average dead time value could be estimated as about 20%.

The further work concerning the reduction of the dead time (especially T_{BO}) could be interesting, while up till now some works concern rather the prioritization issue.

References

[1] M. R. Akhavan 2004. *Study the performance limits of IEEE 802.11 WLANs,* Master Thesis, Lulea University of Technology, Sweden.

[2] V. Bharghavan, A. Demers, S. Shenker and L. Zhang, 1994. "MACAW: A media Access protocol for wireless LAN," *Proceedings of the conference on Communications architectures, protocols and applications,* SIGCOMM, New York, USA, pp. 212-225, http://dx.doi.org/10.1145/190314.190334.

[3] R. Bruno, M. Conti and E. Gregori, 2008. "Throughput Analysis and Measurements in IEEE 802.11 WLANs with TCP and UDP Traffic Flows," *IEEE Transactions On Mobile Computing,* vol. 7, no. 3, pp. 1-16, http://dx.doi.org/10.1109/TMC.2007.70718.

[4] L. Deek, E. Garcia-Villegas, E. Belding, S.-J. Lee and K. Almeroth, 2011. "The Impact of Channel Bonding on 802.11n Network Management," *ACM CoNEXT 2011,* December 6-9. Tokyo, Japan, pp. 1-7, http://dx.doi.org/10.1145/2079296.2079307.

[5] Dolińska, I., 2011. "Wybrane aspekty zapewniania bezpieczeństwa sieci bezprzewodowych IEEE 802.11," *Ekonomiczno-Informatyczny Kwartalnik Teoretyczny,* no. 28, pp. 100-121.

[6] I. Dolińska and A. Masiukiewicz, 2012. "Czynniki ograniczające przepustowość w sieciach standardu 802.11," *Elektronika,* no. 12, pp. 85-88, ISSN 0033-2089.

[7] I. Dolińska, and A. Masiukiewicz, 2012. "Quality of service providing in WLAN networks - possibilities, challenges and perspectives, " in Jałowiecki, P. and P. Łukasiewicz and A. Orłowski, *Information Systems in Management,* Wydawnictwo SGGW, Warszawa, pp. 5-16.

[8] I. Dolińska, A. Masiukiewicz and G. Rządkowski, 2013. "The mathematical model for interference simulation and optimization in 802.11n networks," *Proceedings of the The Concurrency, Specification, and Programming Workshop,* CS&P 2013, Warsaw, Poland, pp. 99-110.

[9] IEEE 802.11-2012, 2012. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE Working Group for WLAN Standards, http://www.ieee802.org.

[10] P. Gajewski and S. Wszelak, 2008. *Technologie bezprzewodowe sieci teleinformatycznych,* WKŁ, Warsaw, Poland.

[11] K.Huang, 2010. *On Wireless Local Area Networks,* Hamilton Institute National University, Ireland Maynooth.

[12] M. H. Manshaei and J.-P. Hubaux, 2010. *Performance Analysis of the IEEE 802.11 Distributed Coordination Function: Bianchi Model,* Mobile Networks: http://mobnet.epfl.ch.

[13] Ni Qiang, L. Romdhani and T. Turletti, 2004. "A Survey of QoS Enhancements for IEEE 802.11 Wireless LAN," *Journal of Wireless Communications and mobile computing,* Wiley, Volume 4 Isssue 5, http://dx.doi.org/10.1002/wcm.v4:5.

[14] Chih-Yu Wang and Hung Yu Wei, 2008. *IEEE 802.11n MAC Enhancement and Performance Evaluation,* LLC 2008, http://dx.doi.org/10.1007/s11036-008-0129-2.

DOI: 10.15439/2014F479
ACSIS, Vol. 3

Non-Destructive Testing in Complexes Cabling Networks using Time Domain Reflectometry and Particle Swarm Optimization

Hamza Boudjefdjouf
Université de Constantine 1
laboratoire d'Electrotechnique de Constantine 25000
Algerie
hamza.boudjefdjouf@lec-umc.org

Rabia Mehasni, Houssem Bouchkara
Université de Constantine 1
laboratoire d'Electrotechnique de Constantine 25000, Algerie
mehasni@yahoo.fr
bouchekara.houssem@gmail.com

Antonio orlandi, Francesco De paulis
UAq EMC Laboratory, Dept of Industrial and Information Engineering and Economics, via G. Gronchi, 18,
67100 - L' AQUILA – ITALY
antonio.orlandi@ing.univaq.it
francesco.depaulis@univaq.it

***Abstract*—Non-Destructive Testing (NDT) is an important area of research, dealing with diagnostic and monitoring the health of the electrical transmission networks and find automatically the failures. One of the recently developped (NDT) techniques is Time Domain Reflectometry methods, they are quite efficient for detecting important damages (hard faults), such as short or open-circuits.**

Interpreting the results obtained with reflectometry instrument for a wiring network requires great expertise, as the reflectometry response can be very complex. Morever, the reflectometry response it self is not self-sufficient to identify and localate the defects in cabling networks. There is the need to solve efficiently the inverse problem which consists of deducing some knowledge about the defects from the response at the input point of the network.

In this paper, TDR and PSO algorithm have been combined and applied to produce a new sufficiently optimized method that permit the extaraction of damages informations from the time domain reflectograms. Finite Difference Time Domain (FDTD) method has been used to produce a training data set with the known of damages. The results obtained from the TDR-PSO algorithm confirmed the theoretical predictions, and gave us exact informations about the complexe structure's health.

I. Introduction

TDR is a measurement technique used to determine the characteristics of electrical lines by observing reflected waveforms [1]. The impedance of the discontinuity can be determined from the amplitude of the reflected signal. The distance to the reflecting impedance can also be determined from the time that a pulse takes to return.

Due to the complex characteristics of the cabling networks which mean that the TDR responses are not self-explanatory some papers use different techniques to read and explain those responses [2] [3] [4].

In this paper a forward model is developed to generate TDR responses and the particle swarm optimization (PSO) technique is used to solve the inverse problem in order to detect and localize faults.

This paper is organized as follows. The proposed TDR-PSO approach is presented in section 2. In section 3, the results are exposed and discussed. Finally, conclusions are drawn in section 4.

II. The Proposed TDR-PSO Approach

The Finite Difference Time Domain (FDTD) method is employed as the forward model to generate the TDR responses based on an input Gaussian pulse. The construction of the forward response is based on the knowledge of the healthy network, such as the network topology thus the number and location of the junctions, the characteristic impedance of the lines, and the length of each line section etc... Moreover the assumption that all the final network branches are terminated on an open load. A basic assumption is that the response simulated by the FDTD algorithm is consistent to the measured response for the healthy network case. This assumption must be verified anytime a new network needs to be analyzed then; the PSO is applied to the inverse process after the generated TDR response using the forward model is compared with the measured one. If a fault is present along the network then the measured TDR response would not match to the simulated response of the healthy network, thus the PSO algorithm operates to modify the topology of a faulty network. This process is repeated until the termination criterion (convergence) is achieved.

A. *The forward model*

The scalar transmission-line equations for two-conductor lines are:

$$\frac{\partial V(z,t)}{\partial z} = -R.I(z,t) - L.\frac{\partial I(z,t)}{\partial t} \quad (1)$$

$$\frac{\partial I(z,t)}{\partial z} = -GV(z,t) - C.\frac{\partial V(z,t)}{\partial t} \quad (2)$$

Where V and I are n x 1 vectors of the line voltages and line currents, respectively. The position along the line is denoted as z and time is denoted as t. The R (resistance), L (inductance), C (capacitance) and G (conductance) are the per-unit-length parameters. The values of these parameters are computed analytically.

$$Z_c = \sqrt{\frac{L}{C}} \qquad (3)$$

As an alternative method of obtaining this general solution we have used the FDTD method which converts the differential equations into recursive finite difference equations [5].
Replacing the derivatives with centered differences in (1) and (2), we obtain the following recurrence equations:

$$\left[\frac{\mathrm{L}}{\Delta \mathrm{t}}+\frac{\mathrm{R}}{2}\right] \mathrm{I}_{\mathrm{k}}^{\mathrm{n}+1}=\left[\frac{\mathrm{L}}{\Delta \mathrm{t}}-\frac{\mathrm{R}}{2}\right] \mathrm{I}_{\mathrm{k}}^{\mathrm{n}}-\frac{\mathrm{L}}{\Delta \mathrm{z}}\left[\mathrm{V}_{\mathrm{k}+1}^{\mathrm{n}}-\mathrm{V}_{\mathrm{k}}^{\mathrm{n}}\right] \qquad (4)$$

$$\left[\frac{\mathrm{C}}{\Delta \mathrm{t}}+\frac{\mathrm{G}}{2}\right] \mathrm{V}_{\mathrm{k}}^{\mathrm{n}+1}=\left[\frac{\mathrm{C}}{\Delta \mathrm{t}}-\frac{\mathrm{G}}{2}\right] \mathrm{V}_{\mathrm{k}}^{\mathrm{n}}-\frac{1}{\Delta \mathrm{z}}\left(\mathrm{I}_{\mathrm{k}}^{\mathrm{n}-1}-\mathrm{I}_{\mathrm{k}-1}^{\mathrm{n}-1}\right) \qquad (5)$$

B. Characteristic impedance and measurement of the coaxial cable:

In this work, the RG58 CU coaxial cable shown in Figure 1 has been used. For this coaxial cable the capacitance per unit length, C, and inductance per unit length, L, are functions of cable geometry [6].
The per-unit length parameters of this coaxial cable are :

$$C = \frac{2\pi\varepsilon}{ln\left(\frac{R_i}{A}\right)} = 100 \times 10^{-12}\ F/m \qquad (6)$$

$$L = \frac{\frac{\mu}{2\pi}}{ln\left(\frac{R_i}{A}\right)} = 250 \times 10^{-9}\ H/m \qquad (7)$$

$$G = Cw.tg(\delta) = 2w \times 10^{-13}\ S/m \qquad (8)$$

$$r = (1/\pi A^2\,\sigma) + (1/\sigma\pi\,(R_0 + R_i)(R_0 - R_i)) \qquad (9)$$
$$= 0.02\ \Omega/\mathrm{m}$$

Where A, R_i and R_0 are the radii of the inner outer and external conductors; μ and ε are the magnetic permeability and dielectric permittivity, respectively, of the material between the conductors, $tg(\delta)$ is the dissipation factor.

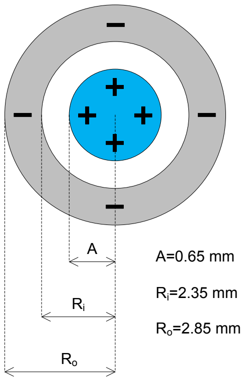


Figure 1: RG 58 CU coaxial cable.

C. Time Domain Reflectometry Measurements

The classical way to do TDR measurements is to inject a signal into the inner conductor of the coaxial cable, which propagates along the cable and when meeting a discontinuity of impedance, a part of its energy is reflected back to the injection point where it is then measured. This reflectogram is used to detect, localize and characterize defects.
A vector network analyzer (VNA Anritsu 10 MHz-2 GHz) has been used to inject the signal into the inner conductor of the RG 58 coaxial cable and to measure the response. The impulse response is deduced from measurement of S_{11} parameter in the frequency domain from 10 MHz to 2 GHz and by using IFFT (Inverse Fast Fourier Transform) to move from frequency to the time domain. A raised cosine pulse, with a rising time of 4 ns and amplitude of 1V has been used.

D. Solving the Inverse Problem Using the Particle Swarm Optimization

D.1 overview

PSO [7] is an evolutionary algorithm for the solution of optimization problems. It belongs to the field of Swarm Intelligence and Collective Intelligence and is a sub-field of Computational Intelligence. It was developed by Eberhart and Kennedy [7] and inspired by social behavior of bird flocking or fish schooling. Several modifications in the PSO algorithm had been done by various researchers [8]. PSO is simple in concept, as it has a few parameters only to be adjusted. It has found applications in various areas like constrained optimization problems, min-max problems, multi-objective optimization problems and many more [8].
The PSO [9] method is regarded as a population-based method, where the population is referred to as a swarm. The swarm consists of n individuals called particles, each of which represents a candidate solution [10]. Each particle i in the swarm holds the following information: (i) it occupies the position $x_{i,}$ (ii) it moves with a velocity v_i, (iii) the best position, the one associated with the best fitness value the particle has achieved so far $pbest_{i,}$ and (iv) the global best

position, the one associated with the best fitness value found among all of the particles gbest.

The fitness of a particle is determined from its position. The fitness is defined in such a way that a particle closer to the solution has higher fitness value than a particle that is far away. In each iteration, velocities and positions of all particles are updated to persuade them to achieve better fitness. The process of updating is repeated iteratively either until a particle reaches the global solution within permissible tolerance limits, or until a sufficiently large number of iterations is reached. Magnitude and direction of movement of a particle is influenced by its previous velocity, its experience and the knowledge it acquires from the swarm through social interaction.

D.2 Velocity and position of the particles

In every iteration, each particle adjusts its own trajectory in the space in order to move towards its best position and the global best according to the following equations:

$$v_{ij}^{t+1} = wv_{ij}^{t} + c_1 rand_{1j}^{t}\left(pbest_{ij}^{t} - x_{ij}^{t}\right) + c_2 rand_{2j}^{t}\left(gbest_{j}^{t} - x_{ij}^{t}\right) \tag{6}$$

$$x_{ij}^{t+1} = x_{ij}^{t} + v_{ij}^{t+1} \tag{7}$$

for j∈1..d where d is the number of dimensions, i∈1..n where n is the number of particles, t is the iteration number, w is the inertia weight, $rand_1$ and rand_2 are two random numbers uniformly distributed in the range [0,1], and c_1 and c_2 the acceleration factors. c_1 is the cognitive acceleration constant. This component propels the particle towards the position where it had the highest fitness. c_2 is the social acceleration constant. This component steers the particle towards the particle that currently has the highest fitness.

The velocity of a particle is bounded between properly chosen limits $v_{min}<v_{id}<v_{max}$ (in most cases $v_{min}=-v_{max}$). Similarly, the position of a particle is restricted between properly chosen constants $x_{min}<x_{id}<x_{max}$.

Afterwards, each particle updates its personal best using the equation (assuming a minimization problem):

$$pbest_i^{t+1} = \begin{cases} pbest_i^t & if\ f(pbest_i^t) \le f(x_i^{t+1}) \\ x_i^{t+1} & if\ f(pbest_i^t) > f(x_i^{t+1}) \end{cases} \tag{8}$$

Finally, the global best of the swarm is updated using the equation (assuming a minimization problem):

$$gbest^{t+1} = \arg\min f(pbest_i^{t+1}) \tag{9}$$

Where$f$, is a function that evaluates the fitness value for a given position.

The general Particle swarm optimization method is described in algorithm 1.

Algorithm 1 Particle Swarm Optimization

```
{Initialization}
for i = 1 to n do
    initialization of position and velocity of each particle
    $p_i=x_i$
    $p_{besti}=\infty$
end for
while termination condition not true do
    for i = 1 to n do
        {Update personal best positions}
        if $f(x_i)< p_{besti}$ then
            $p_i=x_i$
            $p_{besti}= f(x_i)$
        end if
        {Update global best particle}
        if $f(p_{besti})< g_{besti}$ then
            $g_{besti} = p_{besti}$
        end if
    end for
    {Update velocities and positions}
    for i = 1 to n do
        $v_{ij}^{t+1} = wv_{ij}^{t} + c_1 rand_{1j}^{t}(pbest_{ij}^{t} - x_{ij}^{t}) + c_2 rand_{2j}^{t}(gbest_{j}^{t} - x_{ij}^{t})$
        $x_{ij}^{t+1} = x_{ij}^{t} + v_{ij}^{t+1}$
    end for
end while
(11)
```

III. APPLICATIONS AND RESULTS

In order to illustrate the proposed methodology two complexes networks configurations are investigated which are: the YY and the YYY-shaped networks. For each configuration three cases are studied: the healthy network, a network with one fault and a network affected with more than one fault.

Our objective, knowing the topology of the network, is to detect, localize and characterize these faults through finding the length (L_i) and the resistance (R_i) of each branch. The location of each assumed fault is characterized by two parameters: the index of its branch and its distance from the input while the characterization can be either a short circuit or an open circuit.

It worth mentioning that all the TDR responses represented in this article are plotted as a function of distance from the test (or origin) point in order to facilitate relating the main peaks with the configuration of the network.

A. The YY-shaped network

Figure 2 shows the experimental network composed of five branches L_1=1m, L_2=4m, L_3=1m, L_4=0.5m and L_5=1.5m respectively.

Figure 2: The YY-shaped network.

Before starting the fault study of the second configuration, the healthy network is investigated. The TDR responses obtained using measurements and simulations are sketched in figure 3. The analysis of the measured and the simulated TDR presented in figure 3 shows that the positions and amplitudes of the main peaks correspond to the topology of the tested network. Moreover, this figure shows that there is a good matching between the measured and simulated results. The small differences are due to the fact that the skin effect has been neglected.

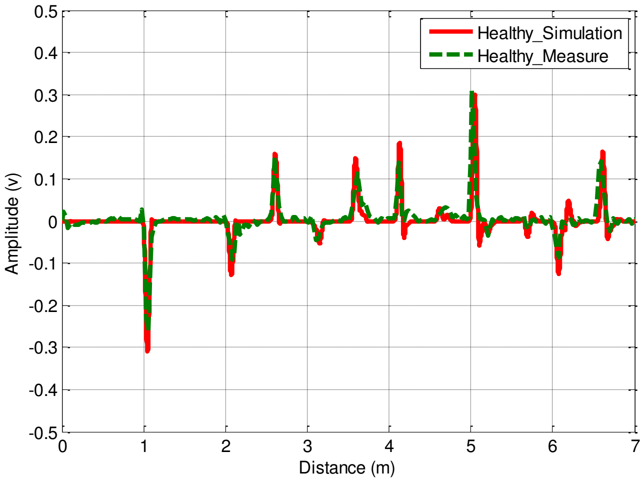


Figure 3: Comparison between the healthy measured and simulated TDR responses of the YY-shaped network.

The faulty study:

In the faulty study two cases are investigated (CASE 1 and CASE 2). The design variables are L_1, L_2, L_3, L_4, L_5, R_1, R_2, R_3, R_4 and R_5.

CASE 1: in this first case the test network is affected by one hard fault (short circuit) in L_4 at 2.4m from the origin point as illustrated in Figure 4.

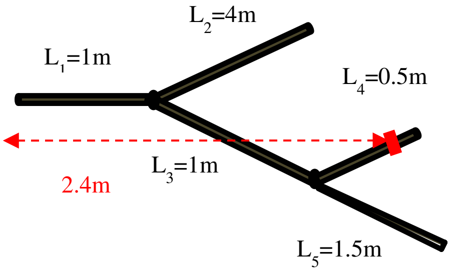


Figure 4: Schematic representation of CASE 1.

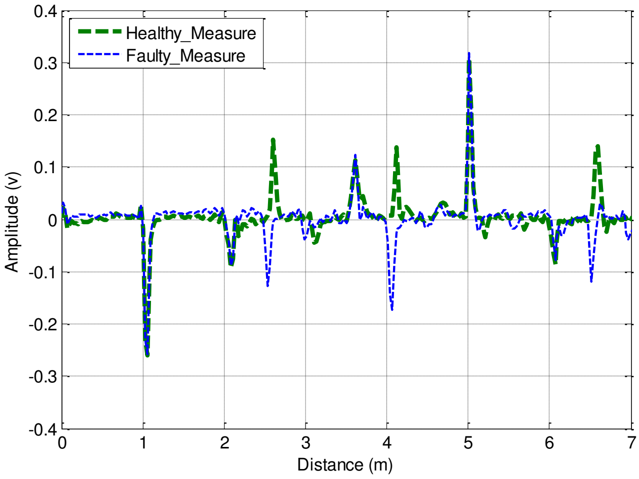


Figure 5: Comparison between the healthy measured and the faulty measured TDR responses for CASE 1.

Figure 5 illustrates clearly the difference between the healthy and the faulty measured responses of the network considered in CASE 1.

After running the TDR-PSO process (50 iterations), the results obtained are: L_1=1m, L_2=4.02m, L_3=1.02, L_4=0.48, L_5=1.53, R_1=1, R_2=1, R_3=1 R_4=0 and R_5=1. The comparison of the obtained results with the known healthy network leads to draw the conclusions summarized in Table 1.

Table I
Fault study for CASE 1.

Detection	The tested network is not healthy. It is affected by one fault.
Cauterization	The nature of the fault is a short circuit one.
Location	The fault is located in branch L_4 at 2.48 m from the origin.

The relative error in locating the fault is 3.33%. Figure shows that the TDR responses of the actual- and reconstructed-faulty networks (using the developed approach) match.

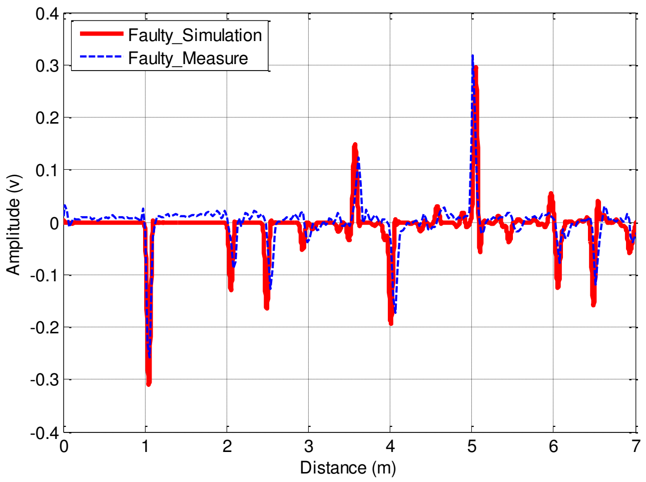


Figure 6: Comparison between the TDR responses measured and reconstructed by the TDR-PSO approach for CASE 1.

CASE 2: This case is illustrated in Figure 7. Here, the test network is assumed to be affected by two hard faults in two different branches: the first fault is an open circuit in L_2 at 3m and the second fault is a short circuit in L_4 at 2.4m.

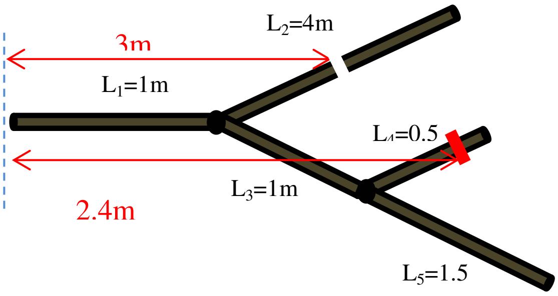


Figure 7: Schematic representation of CASE 2.

After running the TDR-PSO process, the results obtained are: L_1=1m, L_2=2.03m, L_3=1.02, L_4=0.48, L_5=1.53, R_1=1, R_2=1, R_3=1, R_4=0 and R_5=1. The comparison of the obtained results with the known healthy network leads to draw the conclusions summarized in Table 2.

Table II
Faulty study for CASE 2.

Detection	The tested network is not healthy. It is affected by two faults.
Cauterization	Fault one is an open circuit while fault two is a short circuit.
Location	The faults are located in L2 at 3.03m and in L4 at 2.48m, respectively.

The relative errors in locating the first and second faults are: 1% and 3.33%, respectively. Furthermore, Figure 8 shows the excellent matching between the TDR responses of both: the actual- and reconstructed-faulty networks using the proposed approach.

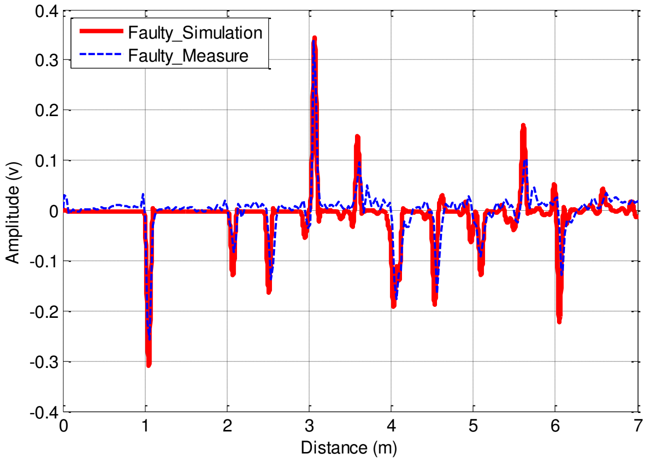


Figure 8: Comparison between the TDR responses measured

and reconstructed by the TDR-PSO approach for CASE 2.

B. The YYY-shaped network

Figure 9 shows the Schematic representation of this network composed of seven branches L_1=1m, L_2=4m, L_3=1m, L_4=0.5m, L_5=1.5m, L_6=0.5m and L_7=2.5m respectively.

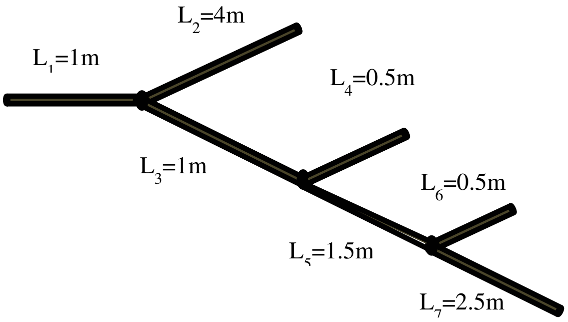


Figure 9: Schematic representation of the healthy YYY-shaped network

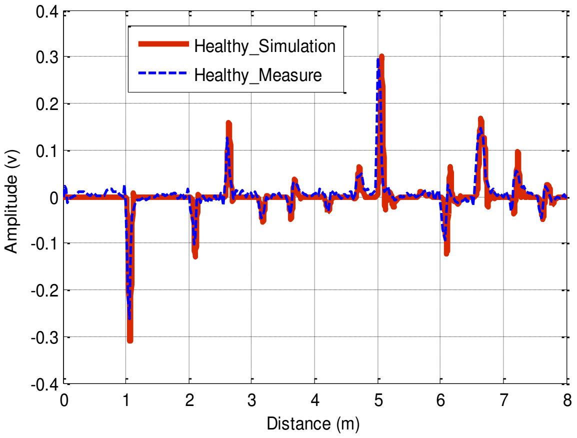


Figure 10: Comparison between the healthy measured and simulated TDR responses of the YYY-shaped network.

This figure shows that there is a good matching between the measured and simulated results.

The fault study:

In the fault study two cases are investigated (CASE 1 and CASE 2). The design variables are L_1, L_2, L_3, L_4, L_5, L_6, L_7, R_1, R_2, R_3, R_4, R_5, R_6, and R_7.

CASE 1: in this first case the test network is affected by two hard faults, short-circuit in L_2 at 3m and an open-circuit in L_4 at 2.4m from the origin point as illustrated in figure 11.

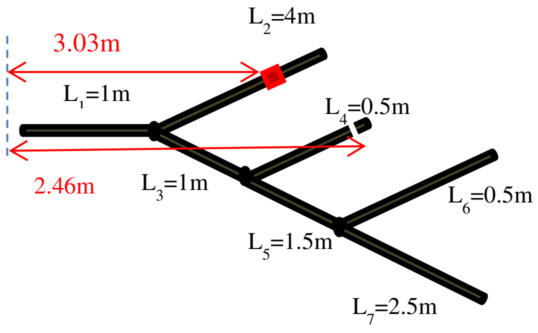


Figure 11: Schematic representation of the case 1

After running the TDR-PSO process (100 iterations), the results obtained are: L_1=1m, L_2=2.03m, L_3=1.02, L_4=0.46, L_5=1.53, L_6=0.53 L_7=2.53, R_1=1, R_2=0, R_3=1, R_4=1 and R_5= R_6= R_7=1. The comparison of the obtained results with the known healthy network leads to draw the conclusions summarized in Table III.

TABLE III

Faulty study for CASE 1.

Detection	The tested network is affected by two hard faults.
Cauterization	Fault one is a short circuit while fault two is an open circuit.
Location	The faults are located in L2 at 3.03m and in L4 at 2.46m, respectively.

Figure 12 shows the excellent matching between the TDR responses of both: the actual- and reconstructed-faulty networks using the proposed approach.

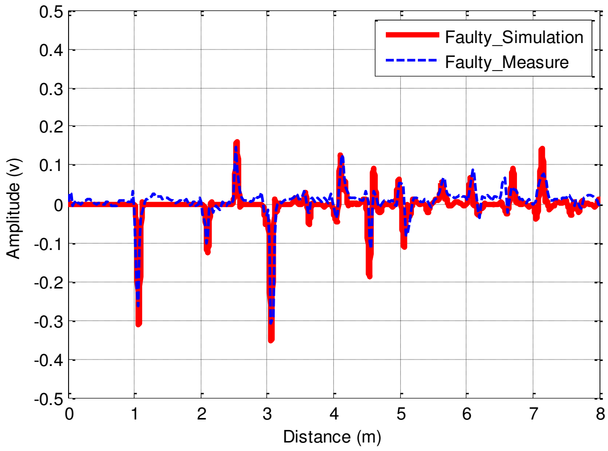


Figure 12: Comparison between the TDR responses measured and reconstructed by the TDR-PSO approach for CASE 1.

CASE 2: in this case the test network is affected by three hard faults; short-circuit in L_2 at 1.4m, another in L_4 at 2.4m and an open-circuit in L_7 at 5.5m from the origin point as illustrated in figure 13.

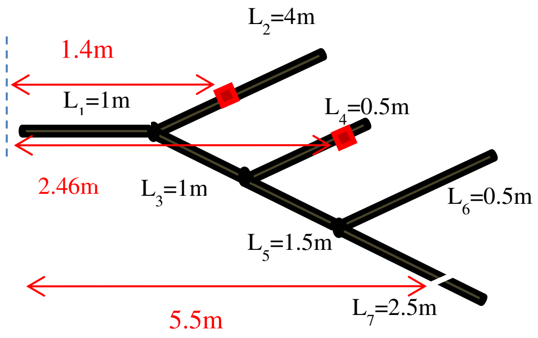


Figure 13: Schematic representation of the case 2.

After running the TDR-PSO process (100 iterations), the results obtained are: L1=1m, L2=0.43m, L3=1.02m, L4=0.46m, L5=1.53m, L6=0.5m, L7=2m, R1=1, R2=0, R3=1, R4=0 and R5=R6=R7=1. The comparison of the obtained results with the known healthy network leads to draw the conclusions summarized in Table IV.

TABLE IV

Faulty study for CASE 2.

Detection	The tested network is affected by three hard faults.
Cauterization	Fault one and two are short circuits while fault three is an open circuit.
Location	The faults are located in L2 at 1.4m and in L4 at 2.46 m, and in L7 at 5.5m respectively.

Figure 14 shows the excellent matching between the TDR responses of both: the actual- and reconstructed-faulty networks using the proposed approach.

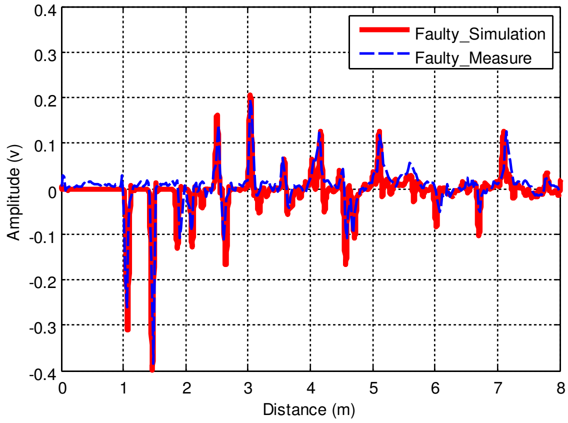


Figure 14: Comparison between the TDR responses measured and reconstructed by the TDR-PSO approach for CASE 2.

IV. CONCLUSION

In this article, a novel methodology based on TDR and PSO for the Non-Destructive Diagnosis of cabling networks is proposed, the physical informations about the network are deduced by processing the reflectometry response and solving the inverse problem by the PSO. The proposed approach has been successfully tested on several cases and for different configurations such as the YY and the YYY-shaped networks. The comparisons of the proposed TDR-PSO approach results with measurements reveal that this approach has a high potential and it is very effective for cabling network diagnosis.

V. REFERENCES

[1]. Furse, Cynthia, Lo, Chet, “Reflectometry for Locating Wiring Faults,” IEEE Trans. Electromagnetic Compatibility. February, 2005.

[2]. M. K. Smail, T. Hacib, L. Pichon, and F. Loete, “Detection and location of defects in wiring networks using time-domain reflectometry and neural networks,” IEEE Trans. Magn, vol. 47, no. 5, may 2011.

[3]. Layane Abboud, Andrea Cozza, and Lionel Pichon, “A Matched-Pulse Approach for Soft-Fault Detection in Complex Wire Networks,” IEEE Trans. Instrumentation and Measurement, vol. 61, no. 6, June 2012.

[4]. H. R. E. H. Bouchekara, M. K. Smail, and G. Dahman“Diagnosis of Multi-Fault Wiring NetworkUsing Time-Domain Reflectometry and Electromagnetism-Like Mechanism,” Electromagnetics, 33:2, 131-143.

[5]. C R Paul, Analysis of Multiconductor transmission lines, New York. Wiley 1994.Halliday, D. and Resnick, R., 1962, Physics, Part II, 2nd ed., John Wiley & Sons, New York.

[6]. Halliday, D. and Resnick, R., 1962, Physics, Part II, 2nd ed., John Wiley & Sons, New York.

[7]. Kennedy, J. Eberhart, R. C. Particle swarm optimization. Proceedings IEEE international conference on neural networks, vol. 7 (1995), pp. 1942– 1948.

[8]. R. Rajendra, D. K. Pratihar, Particle Swarm Optimization Algorithm vs Genetic Algorithm to Develop Integrated Scheme for Obtaining Optimal Mechanical Structure and Adaptive Controller of a Robot Open Access, Intelligent Control and Automation,vol.2 (2011).

[9]. M. El-Abd, Cooperative models of particle swarm optimizers. University of Waterloo, Canada, ProQuest Dissertations and Theses, 2008.

[10]. Robinson, J. and Y. Rahmat-Samii, Particle swarm optimization in electromagnetics, IEEE Transactions on Antennas and Propagation, vol. 52 (2004), pp. 397-407.

DOI: 10.15439/2014F459
ACSIS, Vol. 3

Dynamic Autonomic Network Management: Evaluating the Architectural Challenges of Autonomic Management for Mobile Ubiquitous Access

[1]Clifford C.L Sibanda, Olabisi E. Falowo

Department of Electrical Engineering,
University of Cape Town,
Private Bag X3, Rondebosch 7701 :
[1]clifford@crg.ee.uct.ac.za

***Abstract*—As technology rapidly improves there is more mobile and portable devices available on the market, making the prospects of ubiquitous access to Information Communications Technology (ICT) services a bigger better reality every day. The major hurdle which is the ICT skills shortage can be solved by using autonomic management of the devices on the network and end user equipment. Network and application service providers competing to retain the customer base in order to maintain a guaranteed and healthy income, need to improve network management and stick to service level agreements. This can easily be achieved through enabling network components to automatically configure and optimize their settings, operations and performance. Autonomic network and device management has great advantages including, reduction of human error, reduction on the dependency of the scarce and expensive human skill and much faster introduction of applications, new services and technology, saving the critical and scarce time. However, due to architectural differences major problems arise when a mobile node traverses heterogeneous networks and systems that employ different management paradigms different aspects for similar processes such as Call Admission Control (CAC) mechanisms, Quality of Service (QoS) issues and Security.**

***Index Terms*—Self-configuration, Self-optimization, Ubiquitous access, Mobility, Heterogeneous Networks**

I. Introduction

THE LAST few years have seen rapid developments in technology on both the network side and the devices side. However the speed of the proliferation of high-tech devices to the ordinary user has not and could not be matched by the human Information and Communications Technology (ICT) skills available in the world [1]. Thus configuration management is often left to the poor user who parts with hard earned cash only to enjoy the access to service, data, information, entertainment and the entire digital tech world can offer. The level of skills possessed by the average ordinary user is far less than enough to adequately and optimally use the ICT resources available

The proliferation of the complex, altogether different yet complimentary heterogeneous networks introduces yet another angle that leads to user confusion and inefficient use of resources available. Ubiquitous service access by mobile users across heterogeneous systems, without the bother of changing settings or devices is desirable.

Human interface in network management is hampered by several aspects including but not limited to the already mentioned worldwide shortage of the trained ICT personnel. Human operations are prone to errors and sometimes poor judgment leading to unavailability of ICT resources to impatient end users.

Also, as service providers compete to retain the customer base in order to maintain a healthy income, the need to increase the up time of the network for users to easily access ICT resources is high. On the user equipment side the need to lessen the burden on the user is also desirable, through some means of automation that would make access to service easy and fast.

The need to avail all the new services as soon as possible is extremely unavoidable. One such route to be used to achieve this end is to allow the network and devices to automatically configure, heal, protect and optimize their performance. Thus the introduction of intelligence in the end systems as has been done ` to the core network systems is very vital to ensure the take of autonomic management of the systems.

Some problems arise when a mobile node traverses heterogeneous networks and systems that employ different management paradigms as shown in Figure 1, with different aspects for similar processes such as Call Admission Control (CAC) mechanisms, Quality of Service (QoS) issues and Security [7] [9].

The remainder of this paper is organized as follows: In Section II, we discuss the concepts of Autonomic Computing/Management, in Section III, we discuss the challenges that we identify as related to the Architecture of networks that need addressing if autonomic management is to take off. Section IV contains our suggested research areas for solutions to the challenges and Section V discusses related work.

II. Human administered Network management & Autonomic Computing

Network Management administered by human beings can be viewed as a full time occupation that involves the deployment, maintenance, optimization and upgrade of network components. Deployment would normally involve installa-

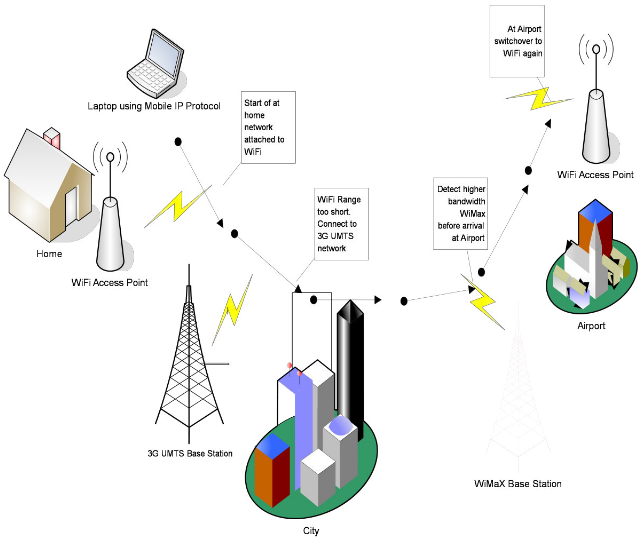

At Airport switchover to WiFi again
Laptop using Mobile IP Protocol
Start of at home network attached to WiFi
WiFi Access Point
Home
WiFi Access Point
WiFi Range too short. Connect to 3G UMTS network
Detect higher bandwidth WiMax before arrival at Airport
Airport
3G UMTS Base Station
WiMaX Base Station
City

Fig 1. Mobile Heterogeneous Network Access

tion, loading of software to interface between the machine and the human, configuration and commissioning the network. Hundred percent uptime of the network is impossible especially if it is in use, hence maintenance is needed. Maintenance could be carried out pro-actively, i.e. to prevent faults from occurring or reactively, i.e. resolving faults that have occurred.

Optimization is generally the changes effected to maximize the gain from the use of the network, as network variables, environment and performance changes on the fly, the need to optimize the network usage is best dealt with using autonomic means. Whereas upgrades effect changes to improve aspects of the network e.g. introducing new drivers or system software that allows for increasing the link speed from 384kb/s to 1Mb/s, these upgrades when automated and occurring in the background allow users to carry on using the network with the prospect of better network experience once the upgrade is complete.

The human dependant network management requires highly skilled personnel to carry out the deployment, maintenance, optimization and upgrades of the network. Accepted, there is a shortage of these skilled ICT personnel to carry out these tasks, hence the need to automate most of the work. Moreover due to the shortage, the highly skilled ICT personnel are in great demand; hence the market offers good remuneration, making them highly mobile. The mobility of the highly skilled ICT personnel at time causes problems as they move with critical network information that they would have gained over the time of their administration.

Also because human beings cannot be at work 24 hours a day, some faults have to wait for a specific person to be on duty for it to be resolved. Human Intervention network management is dependent on the use of tools (shown in Figure 2) that an engineer/network administrator or technician will occasionally consult to make decisions on the status of the network. The frequency of the consultation may also be factor in the status of the system at any given time.

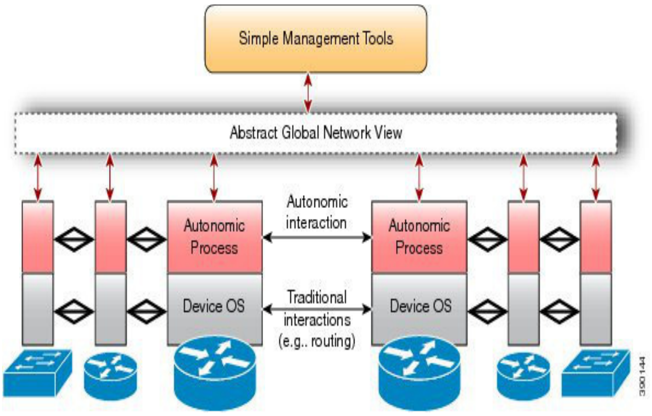


Fig 2. Use of Simple Network Tools with Autonomic [19]

Common tools used by network administration personnel range from protocols such as the Simple Network Management Protocol (SNMP) to Application software on open resources hardware or firmware to those on proprietary hardware and firmware such Cisco equipment using Cisco proprietary hardware, firmware and operating systems. The tools can also include the methodology of network administration, for which several academia papers are available [15], [16], [17] [19].

Network administration management can be classified in different ways such as: the passive versus active network management as well as distributed versus centralized network management.

Passive network administration refers to network administration in which network logging and network configurations are carried out such that they do not affect network operations. This could mean the logs analysis is not real time or online. On the other hand, active network administration refers to real time network logging, configurations and adjustments.

Centralized network administration refers to network management that is guided by one entity or same policies sometimes from a single point. On the other hand, distributed network administration refers to several points/centers of network management and policies creation and implementation.

Autonomic network management refers to the ability of the network to manage itself with minimal human intervention. It is a branch of the Autonomic computing paradigm, and it owes its existence to the Integrated Business Machines (IBM), efforts in the 1990s known as Autonomic management [4]. The efforts of the research in this area have not to date yielded much industry usable solutions.

While the computing and network management area has made great strides from the era of extensive and difficult command line interfaces to Graphic user interface and Web based interfaces, it was not until the introduction of policy based network management ideas that the realization of autonomic network management was any closer to reality.

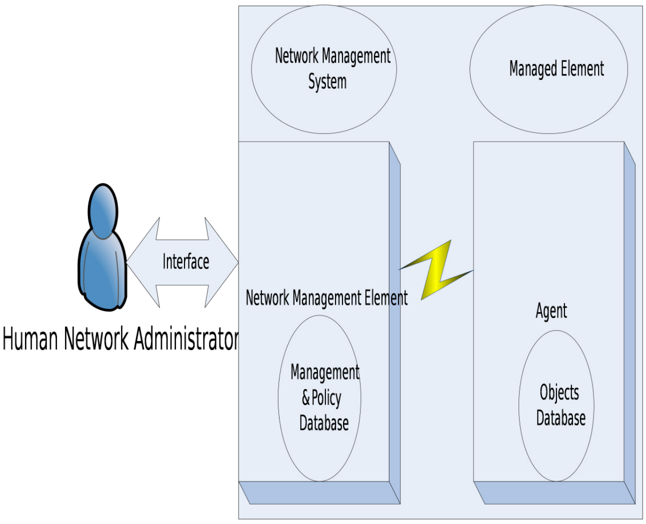

Network Management System
Managed Element
Interface
Human Network Administrator
Network Management Element
Agent
Management & Policy Database
Objects Database

Fig 3. Human Interface Network Management

Autonomic network and device management has several advantages to both the user and the service provider including saving the critical and scarce time, reduction of human error, reduction on the dependency of the scarce and expensive human skill and much faster introduction of applications, new services and technology.

The general principles of autonomic management as envisaged by pioneers of this research paradigm, who included IBM, identified 4 main areas namely self-configuring, self-healing, self-optimizing and self-protecting aspects of automatic management by devices. Several other areas have emerged as research in this arena continues with aspects such as self-aware, self-organization, self-preservation and self-locating, Self-Integration[5].

The four major aspect of Autonomic Network Management Self-CHOP can be as indicated below.

Self-configuration is meant to allow devices to change their configuration as is dictated by the situation and environment.

Self-healing is meant to allow devices to take corrective measures for any systems states that could cause malfunction and disruptions.

Self-optimization is meant to allow devices to take advantage of resources available to the maximum ability

Self-protection is meant to allow devices to enforce appropriate security policies in the event of attacks or perceived intrusions that can cause denial of service or destructive action that can cause loss of service [4].

III. Related Work

Autonomic Network Architecture (ANA) project seeks to develop a network architecture that can self-organize [8]. The research explores ways of organizing and using networks beyond the current Internet technology. The goal is to design and develop a network architecture that is flexible, dynamic, and fully autonomic as a whole. The developed product should be dynamically adaptable according to the working, economical and social needs of the users. One key attribute is that the developed network scales, in a functional, easily extending both horizontally (across systems) as well as vertically (as a solution) [2].

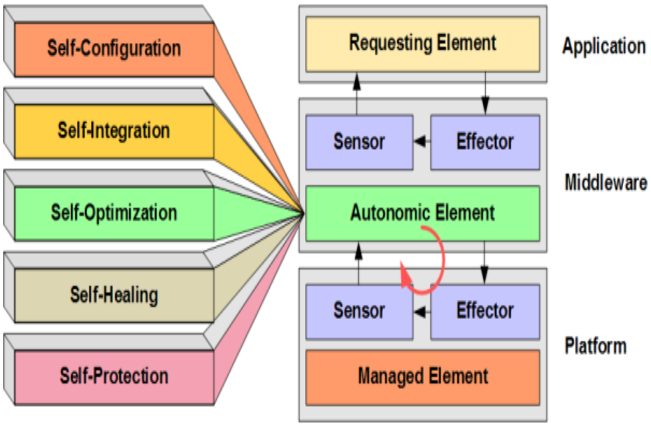


Fig 4. Main Self-Areas of Autonomic Management

Mobility First Future Internet Architecture is part of a bigger project being undertaken in America that seeks to redesign the Internet based on the mobile nodes as opposed to the legacy Internet architecture based on in-situ servers [3]. The projection of this study is that mobile application platforms will replace the fixed application platforms by 2015. With a vision of a future internet that supports mobile devices as ‘first class’ objects, the need to have the mobile nodes operate efficiently, accurately and autonomic is very high.

FOCALE (Foundation Observe Compare Act Learn rEason): This is a distributed architecture that mainly depends on Autonomic Components (AC), where each AC can incorporate the autonomic management functionalities. The main challenge for FOCALE is to accommodate legacy components i.e. already existing network components, and ensure that new Autonomically Enabled Managed Components can also be efficiently integrated and managed. The research also seeks to utilize policy based management of ACs. FOCALE provides a means to reason about the environment and recommend or take appropriate actions, so that the underlying business goals are not violated and, hopefully, optimized [12]. Using sensors to gather information on the environment, FOCALE seeks to implement context-aware policies to change behaviour of Autonomic Components [13] [14]. The context aware policies in our view are relatively closer to achieving SLA honoring. There is need to extend the work in [12] and [13] to ensure enforcement of the SLA within the context of user’s immediate environment. In cases of nonexistent SLA’s or lack of QoS mappings to take care of SLA’s on demand autonomic SLA configurations should be possible.

BIONETS are biologically inspired networks. The human biological system has a stable autonomic system that carries out self-management tasks that ensure balance in the body and thus preserving life [1]. Biological ecosystems also have ways of balancing very complex natural environments. Natural ecosystems tend to balance large populations of diverse organism while efficiently achieving equilibrium through collaboration and competition, yet there is no central controlling entity to organise or manage the equilibrium. The BIONETs project seeks to use the natural systems characteristics to create autonomic networks capable of also managing themselves similarly [6].

IV. Challenges for Autonomic Network Architectures

f devices are enabled to autonomically manage their configuration, state and operations, changes could be effected because of changes in the environment or changes in technology e.g. software updates or version changes. The ultimate goal of the changes is enhanced user experience or more efficient usage of resources or even accommodation of more users for the same service. Changes may include bug fixes or enhanced versions or changes in spectrum used or bandwidth used by devices.

The possibilities of a non recoverable error should total control be left to the devices to change their properties, is also a real danger and thus implementable solutions should allow for recovery and rollbacks. Recovery and roll backs would efficiently be implemented if the devices had enough memory to keep current state before accepting the new state, but the majority of the small devices accessible to users such as cellphones and body area sensors have no memory to hold two different images of software.

In the past few years the speedy convergence of Telecommunications networks and data networks saw an unprecedented upsurge of new applications that readily utilize the advantages offered by the convergence. The major success and effective key driver of convergence, being the phenomenal Internet. Spurred by the all IP networks capable of carrying all kinds of traffic ranging from voice, data to video, the converged network has brought further difficulties in network management.

As such the Internet Engineering task Force (IETF) has been kept busy with modifications of standards, proposals and drafts that help in the management of the Internet. The constant modifications clearly indicate the difficulty in which the current and future applications fit into the originally envisaged architecture of the Internet. There has been a massive increase of the Internet Servers putting the ever questioned lack of central control of the Internet out of the question as it were.

In general the rigid nature of architectural layers of the Internet such as the TCP/IP suite protocols, have literally meant modifications of as many protocols at each small change on the way nodes access the Internet. Several cross layer optimization and workaround solutions have been suggested [11]. However the cross layer solutions have no guarantee of ability to function for the future Internet.

Ad-hoc, mesh and distributed architectures have proved even more popular as the Internet continues to grow, thus the centralized and hierarchical architecture are not the core of the Internet anymore and the future architecture is strongly distributed with high possibilities of the so called core being composed of mobile and ever changing nodes. This idea literally breaks the Internet as we know it and how it is was founded.

The concept of mobile nodes accessing the Internet encouraged a lot of research as the Internet success had for a long time been based on the ability for nodes to route packets via open routes using addressing which had to be static during the initiated connections [10]. The tunneling solutions that emerged attempted to maintain this state for the birth of mobile Internet. To help matters was that the mobile node was always assumed to be communicating with a static server. The future Internet presents even more uncertainty in that the server might be distributed amongst several autonomic nodes which might all be mobile, moreover with no co-ordination whatsoever of the group mobility of the server.

In general the rigid nature of architectural layers of the Internet such as the TCP/IP suite protocols, have literally meant modifications of as many protocols at each small change on the way nodes access the Internet. Several cross layer optimization and workaround solutions have been suggested [11]. However the cross layer solutions have no guarantee of ability to function for the future Internet.

Mobility across heterogeneous networks and systems with different architectures and design also makes autonomic management complex with very little hope of standardization across different proprietary vendor equipment. Call admission control poses a serious problem when it comes to heterogeneous access for mobile nodes.

V. Component Based Solution

This paper envisages the major solutions for the future Internet management will rely on the use of Components and objects as opposed to the hierarchical structures currently used. As such the distributed and flat structure of the network will eventually be realized through non homogeneous mobile nodes serving as both consumers and producers of the content.

Component and object based approaches in the ICT field have shown the advantages of the removal of the single point of failure, component re-use and distributed function value as opposed to centralized and heavily coupled functions. Distributed Components can be combined or re-matched as required. Component and objects upgrades, changes or upgrades can be done without affecting the on-going network operations. [18]

Components are elements that represent independent, interchangeable parts of a system. Components and objects in a system conform to or model one or more interfaces, which allow for the interaction and determine the general behaviour of components.

In general, using components makes a system more flexible, scalable, and reusable. Another advantage of components and objects is that they are replaceable without disrupting the entire systems

For a component to be replaceable, it must meet the following criteria:

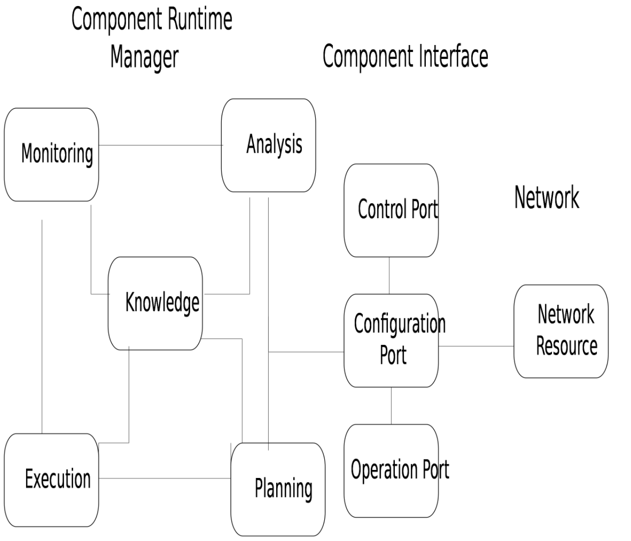


Fig 5. Autonomic Network Management Component visualisation [18]

- The internal structure of the component must be hidden. No dependencies can exist between the contents of the component and other objects.
- Components must provide interfaces so that external objects can interact with them.
- The internal structure of the component must be independent. The internal objects must have no knowledge of the external objects.
- Components must specify their required interfaces so that they have access to external objects.

In models that depict executable systems, components represent the components that are used during the execution of the system. Examples include COM+ objects, JavaBeans™, and Web services.

A component usually is named after the part of the system that it represents.

The boom of small distributed servers with content in the form smartphones, tablets, laptops and personal computers operating as the producer consumer internet client-servers and the current day distributed cloud technology renders little operations for the centralized management and human based management but rather favours the distributed automated policy based management. Business policy is translated into network, applications, storage policies that ensure smooth business operations.

This structure will inform the network management paradigms of the future. The distributed, non homogeneous, non enterprise structure of the network will effectively render the human interface of the network management invalid, but rather a set of policies working on open platforms that regulate access to specific services and groups, however possible resulting in frequent re-configurations by both the network and end user equipment to fit the circumstances.

Hybrid Hierarchal/flat IP Architecture and component based solutions could be the bridge between the current architectures and structures. The hybrid structures will allow for existing systems and structures to co-exist. The emerging systems which will through component based autonomic network management interact with the legacy systems using the hierarchical network management systems.

VI. Future Work

The open research areas that will see the ease of Autonomic Network Architecture design and development easier are not easily quantifiable now. Thus our discussion is not exhaustive but seeks to address the identified problems in this paper.

Desired solutions for Autonomic Network Management should answer the question of assurance of maintained original objective of the network node or the network as a whole as less and less human intervention is effected. This challenge coupled with the security concerns of a network and authentic changes being effected on the network could see the reluctance of industry opting for fully autonomic networks. Version control and verification models for fully autonomic systems are essential.

Control and security in peer environments are as crucial, as the need to fully co-operate in ensuring the flow of information. A balance model of co-operation and trust guidance is important as learning from the environment should not lead to poisoning.

The major solutions lie in the redesign and redefinition of network architecture. As opposed to the layering architecture was pushed by the TCP/IP model, a new component based model will greatly serve the future Internet. Cross layer solutions have attempted to give relief but will not hold up as services and applications continue to evolve in the Future Internet. Components are fully functional units that can interface with any other with a lot of ease. Object oriented components also allow easy re-use and plug-in adaptations.

Mobility as basic feature of the architecture for both the client and server is non-negotiable. Trends have shown a fade in the distinction of client and server, with the emergence of Producer-Consumer models also known as Prosumers. The idea stretches into the paradigm of client nodes being part of the management nodes of the network.

Resource Management and Allocation will take a different approach to please the ever versatile Prosumers market of the future Internet. Dynamic network changes will be the order of the day, but with the question of perceived user satisfaction prioritised. This calls for new call admission and QoS models that follow the changes in the environment

User preferences, profiles and contracts, are now more in the hands of the user than the network administrator and the unpredictable nature of the changes has an impact on the network configurations and operations. As more and more services and applications become available to the digital native users, dynamic autonomic management for the highly mobile, distributed and pervasive nodes, is the only solution.

References

[1] S. Hariri et al, “The autonomic computing paradigm*”, Springer Science 2006.*

[2] H. A. Muller, “Autonomic Computing”, *Technical Note - SEI-2006-TN-006, April 2006, Carnegie Mellon University*

[3] Online - http://mobilityfirst.winlab.rutgers.edu/Vision.html

[4] "An architectural blueprint for Autonomic Computing" , White Paper, IBM, June 2005. *Online http://www.research.ibm.com/autonomic/*

[5] Agoulmine et al, "Challenges of Autonomic Network Management", *Proceedings of IEEE Workshop on Modelling Autonomic Communications Environments 2006*

[6] V. Simon, et al, "Bionets: A new vision of Opportunistic Networks", *Proceedings of WRECOM 2007*

[7] L. Mokhesi, et al, "Context Aware Handoff Decision for Wireless Access Networks using Bayesian Networks", *Proceedings of SAICSIT 2009*.

[8] C. Jelger, et al, "Basic abstractions for an autonomic network architecture", *Proceedings of WoWMoM 2007.*

[9] G. Pujolle, "An autonomic architecture for Network Management and Control" , *New Network Management Trends, UPGRADE Vol. IX, No 6, December 2008*

[10] J. Redi et al, "Mobile IP: A Solution for Transparent, Seamless Mobile Computer Communications", *Upcoming Trends in Mobile Computing and Communications, 1998*

[11] X. Lin, "A Tutorial on Cross-Layer Optimization in Wireless Networks", I*EEE Journal On Selected Areas In Communications, VOL. 24, NO. 8, AUGUST 2006*

[12] J Strassner et al, "FOCALE: A Novel Autonomic Networking Architecture*". Proceedings of Latin-American Autonomic Computing Symposium (LAACS), 2006*

[13] J Strassner et al, "An Autonomic Architecture to Manage Ubiquitous Computing Networks and Applications*". Proceedings IEEE Workshop ICUFN, 2009.*

[14] J. Strassner et al, "A Context-Aware Policy Model to Support Autonomic Networking", *IEEE International Computer Software and Applications Conference 2008.*

[15] Dae-Young Kim, et al, "Ontology-Based Methodology for Managing Heterogeneous Wireless Sensor Networks," *International Journal of Distributed Sensor Networks, vol. 2013.*

[16] Rupali Chopade, et al, Local Area Network Administration Using Mobile, *International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 3, March 2012*.

[17] Thomas A. Limoncelli, et al, The Practice of System and Network Administration, *2nd ed, Person Education, 2007.*

[18] http://acl.ece.arizona.edu/projects/old/Autonomia_ Programmable/index.html

[19] http://www.cisco.com/c/en/us/td/docs/ios-xml/ios/auto_net/configuration/xe-3s/asr903/an-auto-net-xe-3s-asr903-book.html, *March 2014*.

DOI: 10.15439/2014F319
ACSIS, Vol. 3

Send It Safe – A Novel Application for Secure Key Exchange Using Telecommunications Open Middleware APIs

Piotr Wawrzyniak
Orange Polska
Orange Labs
Research and Development Centre
ul. Obrzeżna 7
02-261 Warsaw, Poland
Email: piotr.wawrzyniak@orange.com

Łukasz Wronkowski, Damian Kuniszewski,
Adam Cackowski, Paweł Czapliński,
Karol Szymański
Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
ul. Chopina 12/18
87-100 Toruń, Poland
Email: wronkowski.lukasz@gmail.com

Abstract— Common surveillance of citizens by various intelligence services every year becomes more dangerous threat to our privacy. Recently, number of security enrichment was developed that allows to increase privacy protection during pervasive network use. In this article we present a novel approach that uses telecommunications OpenMiddleware APIs to provide reliable public key exchange protocol.

I. Introduction

NOWADAYS security and privacy issues are getting more and more important for many people using state of the art communication tools like mobile smartphones or internet [1]. The growing need to increase security results in number of applications increasing the privacy and security.

Majority of existing telecommunication security solution generally are intended to be used in IP-based networks, such as internet. This is caused by several factors, among them widespread of internet communication is one of the key driver. However recently growing number of smartphone users results in growth of the importance of mobile security and privacy.

One of the important issues regarding secure communication is the key exchange process when asymmetric ciphers are to be used. Among several available protocols, the scheme based on Diffie-Hellman concept and its derivative Station to Station (STS) protocol [2, 3].

The possibility to use secured communications on mobile devices is often limited by insufficient device capabilities, lack of necessary software libraries or poor internet connectivity.

In this article we present novel application designed for Android-enabled mobile phones which allows to securely exchange cryptography keys with the use of Public Land Mobile Network (PLMN) operator's infrastructure. In fact, our solution makes use of Unstructured Supplementary Service Data (USSD) messages, which provides a finest level of reliability and confidentiality. Number of applications proves that USSD channel can successfully replace IP connectivity in a variety of fields [4-7]. Moreover due to OpenMiddleware Application Programing Interfaces (APIs) network resources, including USSD communication channel can be easily accessed by external services [8-10] with the use of state of the art protocols including RESTful web services.

The remainder of this paper is organized as follows:

- Chapter II provides description of the system architecture, it focuses on the key functions of the entire components,
- In chapter III our proposal of simple, efficient and secure key exchange protocol intended for USSD communication channel is presented,
- Chapter IV describes Human-System Interaction, in particularly focusing on the mobile applications developed for Android-enabled mobile phones.
- Chapter V summarizes the paper and provides possible further extensions.

II. System Architecture

Our system consists of two main parts: application server which act as message proxy between negotiating parties, and mobile application designed for Android-enabled smartphones that allows to exchange encryption keys and facilitate the communication protocol.

Application server act as simple proxy and is the key component of the solution. The importance of that element origins in a USSD communication constraints, which makes it impossible to send USSD datagram directly between two mobile phones (using operator service platforms only). In particular USSD messages can only be send between mobile phone and service platforms and vice versa thus it was necessary to develop proxy service that will enable USSD datagrams exchange between two mobile phones. On the other hand such strong dependency on operator infrastructure makes USSD messages highly reliable communication channel.

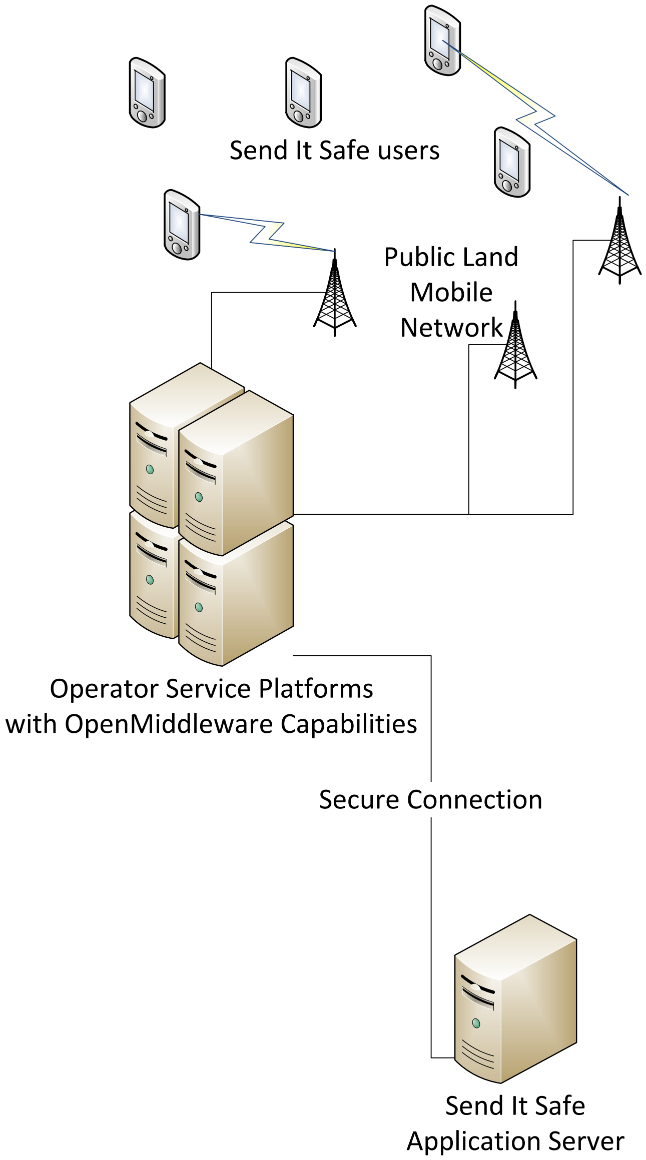

Send It Safe users
Public Land Mobile Network
Operator Service Platforms with OpenMiddleware Capabilities
Secure Connection
Send It Safe Application Server

Fig. 1 The architecture of developed solution

Our proxy server allows to maintain the finest privacy and reliability level. In particular it do not violate the integrity of the payload of exchanged datagrams.

Developed mobile application is designed for Android 4.0 or newer smart phones. It is composed of two cooperating independent components. First one is Android system service that runs in the background. It is started together with the Android OS in independent process. Functionalities of the service includes but are not limited to:

- Capturing SMS messages from the system (before they appear on the screen)
- Capturing received USSD messages
- Categorizing of incoming messages
- Communication with external applications
- Modifying screen dialogs that accompanies USSD messages being sent (percentage progress is displayed instead of standard message)
- Displaying the progress in the system tray while receiving the key
- Supporting retransmission of individual packets when communication error occurs.

Latter part of Android components is Graphical User Interface (GUI) application that is responsible for:

- Concatenating incoming messages
- Encoding and encrypting the messages
- Confirming and checking the identity of the remote party
- • Communication with the Service
- • Dividing outgoing messages to parts that can be sent via USSD protocol
- • Managing key repository (reading stored keys, alternatively it is possible to load the public key of the remote party if we have any in the device repository, verifying whether our keys belong to one pair)
- • Encrypting and sending secured SMS messages and their reception and decryption.

The overall architecture of the proposed solution is provided in Fig. 1.

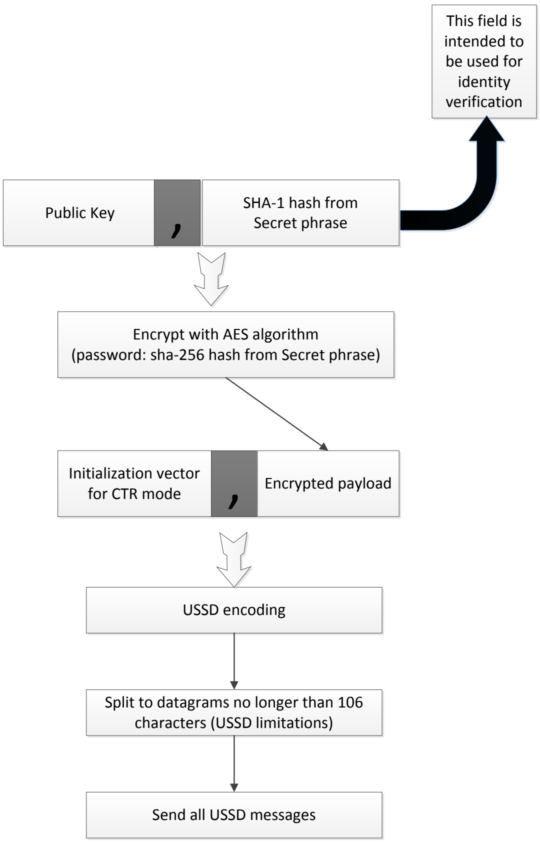

This field is intended to be used for identity verification
Public Key
,
SHA-1 hash from Secret phrase
Encrypt with AES algorithm (password: sha-256 hash from Secret phrase)
Initialization vector for CTR mode
,
Encrypted payload
USSD encoding
Split to datagrams no longer than 106 characters (USSD limitations)
Send all USSD messages

Fig. 2 Implemented key exchange message composing algorithm.

III. KEY EXCHANGE PROTOCOL

In order to use USSD messages for key negotiation, key exchange algorithm was designed and implemented. It uses symmetric key to exchange asymmetric public key between involved parties, which significantly minimizes the number of messages to be exchanged during key negotiation. Minimization of number of datagrams is particularly important factor since it takes about three seconds to exchange single USSD message thus allows to make our system more responsive in human-system interaction experience.

Moreover our algorithm provides efficient mechanism of identity verification which is based on pre-known secret phrase. The secret should be agreed by both negotiating parties prior to entire key exchange. Such approach benefits in strict identity verification capabilities from the one side and allows to keep the number of exchanged datagrams as small as possible on the other. The algorithms for concatenating and initial processing of the key exchange algorithm are presented in Fig. 2 and Fig. 3.

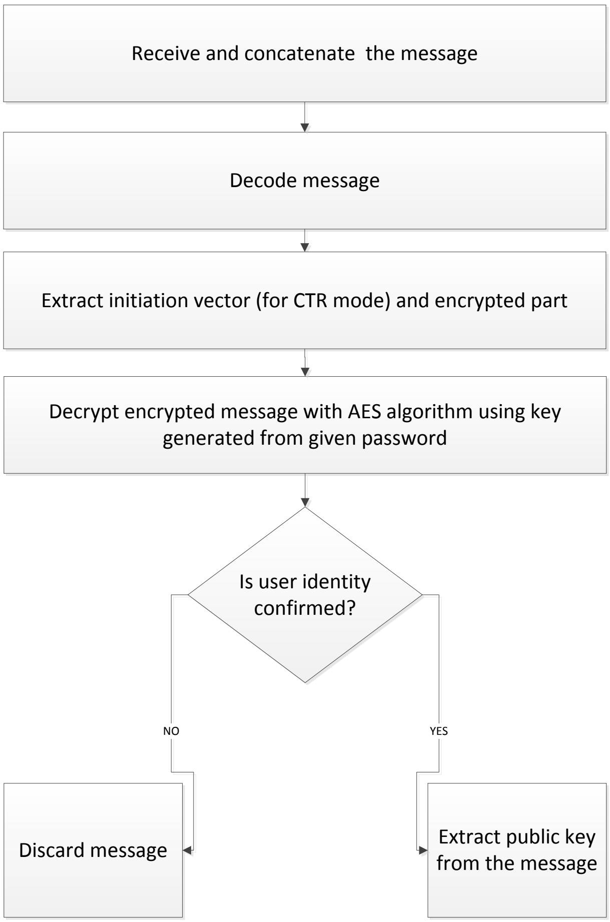

Receive and concatenate the message
Decode message
Extract initiation vector (for CTR mode) and encrypted part
Decrypt encrypted message with AES algorithm using key generated from given password
Is user identity confirmed?
NO
YES
Discard message
Extract public key from the message

Fig. 3 Proposed public key exchange algorithm.

The simplification of the key exchange do not disturb the overall security of the system. Since we use USSD datagrams even this simple approach provides high reliability and security due to fact that entire communication channel is strongly secured.

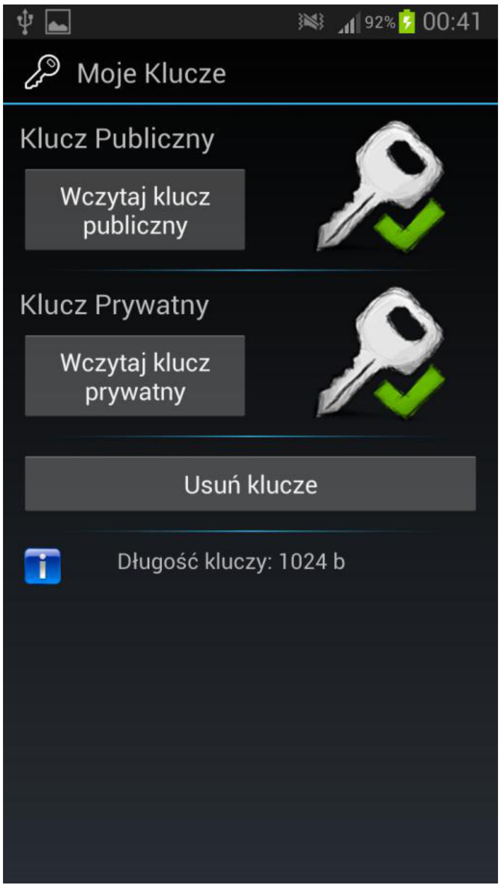


Fig. 4 The “Load Keys” menu (“Wczytaj klucz” means “Load key”). In provided picture both keys were loaded and verified to be the same pair of public and private key.

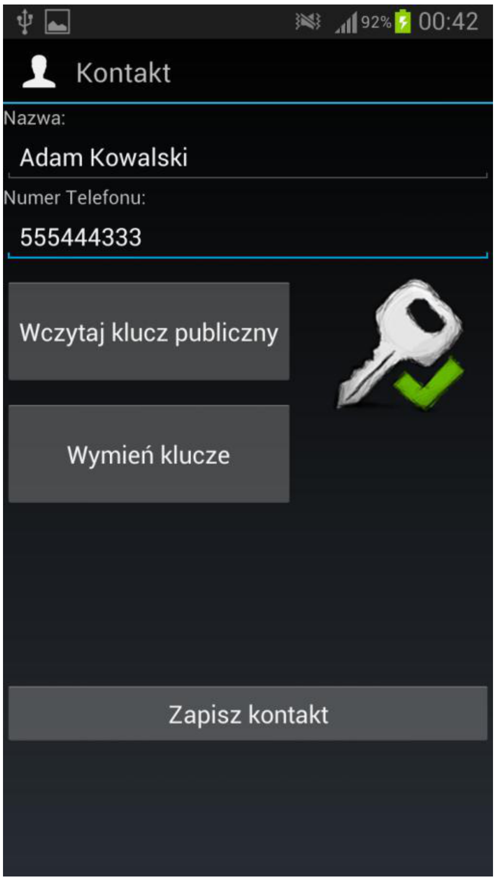
IV. HUMAN-SYSTEM INTERACTION

User might interact with the developed solution with the use of sample application developed as a part of the project. It is primarily intended to be used for key repository management as well as managing and supervising key exchanges, as mentioned in chapter II. This capability is documented in Fig. 4

User interface provides simple visual aids for key management (whether the keys are loaded, validated, etc.) accompanied by the constant monitoring of the key exchange progress. Moreover it can be used to manage contacts (i.e. other people with whom public keys has been already exchanged). The sample menu intended for contact management is provided in Fig. 5.

Moreover exchanged public keys might be exported and easily used by other applications. This feature significantly expand application usability since it can be used for secure and reliable key exchange that are intended to be used by other application or services. This feature makes it also possible to easily incorporate our secure key exchange mechanism into existing services.

Nevertheless the main purpose of the application is to allow secure communication via encrypted SMS or USSD messages with other users. For securing user communication our solution uses RSA algorithm but as it was aforementioned it is possible to store exchanged keys and use them for any purpose.



92% 00:42

Kontakt

Nazwa:

Adam Kowalski

Numer Telefonu:

555444333

Fig. 5 The “Manage Contacts” menu (“Wczytaj klucz publiczny” means “Load public key”, “Wymień klucze” means “Exchange keys”, “Zapisz Kontakt” means “Save contact”).

Upon successful message decryption it remains secured. In particular it stored in dedicated secure storage of the application. Therefore it can be accessed only with the use of our application and is not being displayed in standard messaging application of the phone.

V. Summary

In this article we presented prototype solution designed for secure and reliable key exchange for mobile devices. It makes use of PLMN operator infrastructure which is accessible via OpenMiddleware APIs. Proposed solution is composed by three main parts:

- Android system service, which makes it possible to seamlessly use USSD and SMS communication channels for key exchange protocol,
- Android GUI application designed for key repository management, instant monitoring of the key exchange process and providing tools for communication with the use of secured SMS messages,
- Send It Safe application server, that acts as message proxy. Due to security reasons and in order to provide highest confidentiality level, the proxy do not modify payload of processed datagrams.

Since proposed solution makes extensive use of USSD messages for key exchange, lightweight key exchange protocol has been developed. It allows to minimize the number of exchanged datagrams and provide strong identity verification capabilities.

Moreover exchanged keys can be stored in device memory which makes them accessible to any external application. This approach strongly increase usability of proposed system.

Future system development plan includes implementation of mobile party for other smartphone operating system and implementation of server-side administration panel for performance monitoring.

References

[1] Shklovski, I., Mainwaring, S. D., Skúladóttir, H. H., Borgthorsson, H., & Vej, R. L., “Leakiness and Creepiness in App Space: Perceptions of Privacy and Mobile App Use,” in ACM CHI Conference on Human Factors in Computing Systems, http://dx.doi.org/10.1145/2556288.2557421.

[2] W. Stallings, Cryptography and Network Security: Principles and Practice, Fifth Edition Pearson Education, Inc, 2011

[3] A. J. Menezes, P. C. van Oorschot, S. A. Vanstone, Handbook of Applied Cryptography. CRC Press LLC, 1997.

[4] Bogusz D., Siewruk G., Legierski J., Kunicki J.S.,USSD communication channel as alternative to XML SOAP in mobile Unified Communication applications, Federated Conference on Computer Science and Information Systems. Place: Krakow. September 8-11, 2013

[5] Trusiewicz, P.; Legierski, J., "Parking Reservation - application dedicated for car users based on telecommunications APIs," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.865,869, 8-11 Sept. 2013

[6] Litwiniuk, K.; Czarnecki, T.; Grabowski, S.; Legierski, J., "BusStop — Telco 2.0 application supporting public transport in agglomerations," Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on , vol., no., pp.649,653, 9-12 Sept. 2012

[7] Trusiewicz, P.; Witan, M.; Kuzia, M., "Mobile Payment System - Telco 2.0 application dedicated for payments," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.859,864, 8-11 Sept. 2013

[8] Legierski J., Korbel P.; Telco 2.0 -przykłady praktycznego wykorzystania interfejsów telekomunikacyjnych platform usługowych, KSTIT2011, Przegląd Telekomunikacyjny, 8-9/2011

[9] Wawrzyniak, P.; Korbel, P.; Borowska-Terka, A., "Student information delivery platform using telecommunications open middleware APIs," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.871,874, 8-11 Sept. 2013

[10] Korbel, P.; Wawrzyniak, P.; Grabowski, S.; Krasinska, D., "LocFusion API - Programming interface for accurate multi-source mobile terminal positioning," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.819,823, 8-11 Sept. 2013

3rd International Conference on Wireless Sensor Networks

A FEW years ago, the applications of WSN were rather an interesting example than a powerful technology. Nowadays, this technology attracts still more and more scientific audience. Theoretical works from the past, where WSN principles were investigated, grew into attention-grabbing applications practically integrated by this time in a real life. It could be said, that countless application fields, from military to healthcare, are already covered by WSN. Together with this technology expansion, still new and new tasks and interesting problems are arising. Simultaneously, such application actions stimulate the progress of WSN theory that at the same time unlocks new application possibilities. The typical examples are developments within the "Internet-of-Things" field as well as advancements in eHealth domain with WBAN IEEE 802.15.6 standard progress.

Wireless sensor networks, as the spatially distributed networks consisted of a number of relatively simple, low-cost, low-power components interconnected mutually, provide quite wide application portfolio for different branches of economy. As the main examples could be mentioned military, industry, transport, agriculture and healthcare. However, in the near future, even stronger expansion of WSN application assortment is expected. In order to make this expansion possible, it is necessary to continually work on the solving of typical questions/problems related to the WSN development, e.g. standardization of communication protocols; the lack of energy-efficient power sources; the development of new ultra-low-power microelectronic components; etc.

An integration of WSN within the public data networks as well as within the domains where confidential and private data are processed (e.g. E-Health) brings along problems related to the ethical and legal questions too. Therefore, the terms as social safety or ethical safety should not be neglected.

The problematic of WSN is one of actual activities getting to the fore in the European Research Area since the issue of sensor networks was covered through "IoT" in FP7 program and strong continual extension is planned to be included also in Horizon 2020 program, especially in sections such Smart Transport; Health; Climate Action covered under Societal Challenges Pillar.

It is therefore essential to create an experience-sharing platform for scientific researchers and experts from research institutes, SMEs and companies who work in WSN domain where they can exchange some relevant skills and experiences as well as discuss upcoming trends and new ideas from this field. Moreover, the conference should also serve a function of a kind of networking platform facilitating interconnectivity between participants in case of a future collaboration.

Topics

Original contributions, not currently under review to another journal or conference, are solicited in relevant areas including, but not limited to, the following:

Development of sensor nodes and networks

- Sensor Circuits and Sensor devices – HW
- Applications and Programming of Sensor Network – SW
- Architectures, Protocols and Algorithms of Sensor Network
- Modeling and Simulation of WSN behavior
- Operating systems

Problems dealt in the process of WSN development

- Distributed data processing
- Communication/Standardization of communication protocols
- Time synchronization of sensor network components
- Distribution and auto-localization of sensor network components
- WSN life-time/energy requirements/energy harvesting
- Reliability, Services, QoS and Fault Tolerance in Sensor Networks
- Security and Monitoring of Sensor Networks
- Legal and ethical aspects related to the integration of sensor networks

Applications of WSN

- Military
- Health-care
- Environment monitoring
- Transportation & Infrastructure
- Precision agriculture
- Industry application
- Security systems and Surveillance
- Home automation
- Entertainment – integration of WSN into the social networks
- Other interesting applications

Event Chairs

Hodoň, Michal, University of Žilina, Slovakia
Kapitulík, Ján, University of Žilina, Slovakia
Micek, Juraj, University of Žilina, Slovakia
Ševcik, Peter, University of Žilina, Slovakia

Program Committee

Al-Anbuky, Adnan, Auckland University of Technology, New Zealand

Baranov, Alexander, Russian State University of Aviation Technology, Russia

Dadarlat, Vasile-Teodor, Uniiversita Tehnica Cluj-Napoca, Romania

Diviš, Zdenek, VŠB-TU Ostrava, Czech Republic

Elmahdy, Hesham N., Cairo University, Egypt

Fouchal, Hacene, University of Reims Champagne-Ardenne, France

Giusti, Alessandro, CyRIC - Cyprus Research and Innovation Center

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Gu, Yu, National Institute of Informatics, Japan

Hudik, Martin, University of Zilina

Husár, Peter, Technische Universität Ilmenau, Germany

Jin, Jiong, Swinburne University of Technology, Melbourne

Jurecka, Matus, University of Žilina, Slovakia

Kafetzoglou, Stella, National Technical University of Athens, Greece

Karastoyanov, Dimitar, Bulgarian Academy of Sciences, Bulgaria

Karpiš, Ondrej, University of Žilina, Slovakia

Kochláň, Michal, University of Žilina, Slovakia

Laqua, Daniel, Technische Universität Ilmenau, Germany

Li, Qinxue, Guangdong University of Petrochemical Technology

Monov, Vladimir V., Bulgarian Academy of Sciences, Bulgaria

Ohashi, Masayoshi, Advanced Telecomunications Research Institute International / Fukuoka University, Japan

Púchyová, Jana, University of Žilina, Slovakia

Ramadan, Rabie, Cairo University, Egypt

Scholz, Bernhard, The University of Sydney, Australia

Segal, Michael, Ben-Gurion University of the Negev, Israel

Shaaban, Eman, Ain-Shams university, Egypt

Shu, Lei, Guangdong University of Petrochemical Technology, China

Smirnov, Alexander, Linux-WSN, Linux Based Wireless Sensor Networks, Russia

Staub, Thomas, Data Fusion Research Center (DFRC) AG, Switzerland

Stojmenovic, Ivan, University of Ottawa, Canada

Teslyuk, Vasyl, Lviv Polytechnic National University, Ukraine

Wang, Zhonglei, Karlsruhe Intitute of Technology, Germany

Xiao, Yang, The University of Alabama, United States

Zuo, Liyun, Guangdong University of Petrochemical Technology

DOI: 10.15439/2014F508
ACSIS, Vol. 3

A Design of Application Based Wireless Sensor Node

Sham P Nayse
PHD scholar, Dept Computer Science,
SGBA University Amravati, India.
shamnayse@gmail.com

Mohammad Atique
Associate Prof., Dept Computer Science,
SGBA University Amravati, India.
mohd.atique@gmail.com

Abstract— Proposed wireless sensor node is a highly flexible and programmable system on chip (PSoC) architecture along with integrated RF radio chip, which can be adapted to any IEEE 802.15.4 standard based protocol, working at the band of 2.4GHz. This can be accompanied by the RF antenna matched with the interface circuit for various kinds of sensors and peripherals to form a wireless sensor node. This can be very well used to form a wireless sensor network in different domains. This proposed methodology can be used for specific application for improving the WSN stability and performance. This article will mainly present the design of wireless sensor nodes based on PSoC and CyFi radio chip, including the application specific antenna and hardware design for the system on chip for the sensor node and the simple introduction of the application domain. The proposed nodes are used with a special antenna and application based protocol for communication. This design has a better power handling and performing capacity. Further, we have implemented this for the wireless sensor node as well.

Keywords—PSoC, CyFi, wireless sensor network; cost effective wireless sensor nodes;

I. INTRODUCTION

The System on chip SoC is the arrival of the new technological revolution; the world enters the information and VLSI age. To make use of the information, the first thing that we must do is to obtain accurate and reliable data. We can use sensors to obtain information in the natural parameters with low power and low overhead fields in data communication.

WSN has already penetrated into numerous areas such as industrial production, space development, ocean exploration, environmental protection, resource survey, medical diagnostics, biotechnology, and even conservation, etc. It is no exaggeration to say that, from the deep ocean to the vast space, along with a variety of complex engineering systems and every modern project, is inseparable from a variety of sensors.

The Wireless Sensor Network is composed of a set of sensor nodes. The sensor nodes form the communication network in the of multi-hop and self organization fashion[1]. Using sensor nodes and its network, the information of the monitored objects are distributed to the target area, which can then be collected. This collected information will pass to the upper layer of wireless communication stack in an abstracted way. Therefore, it is very important to design a kind of efficient and practical wireless sensor node [3].

II. THE OVERALL STRUCTURE OF WIRELESS SENSOR NETWORK SYSTEM

The wireless sensor network system architecture is as shown in figure 1.It has three parts; (1) sensor nodes; (2) the sink node/ WSN gateway; (3) monitoring center [3]. Sensor nodes have the capabilities such as data acquisition, signal processing and wireless communication. It is both: the initiator and the transmitter of the information frame. The Self deviser Network which is self-organized and has multi-hop routes sends the collected monitoring data to the WSN gateway. WSN gateway, also known as convergent node, transmits the data collected by the sensor nodes to the monitoring center through the serial communication; The terminal monitoring center mainly carries out tasks such as managing and questing the data from the network, sending the networking request, asking the specified nodes to sample data ,etc. But in the task, the function and the terminal monitoring center is divided into two parts: data acquisition and data management. Figure 1 shows the over all picture of wireless sensor network. These have no of sensors nodes.

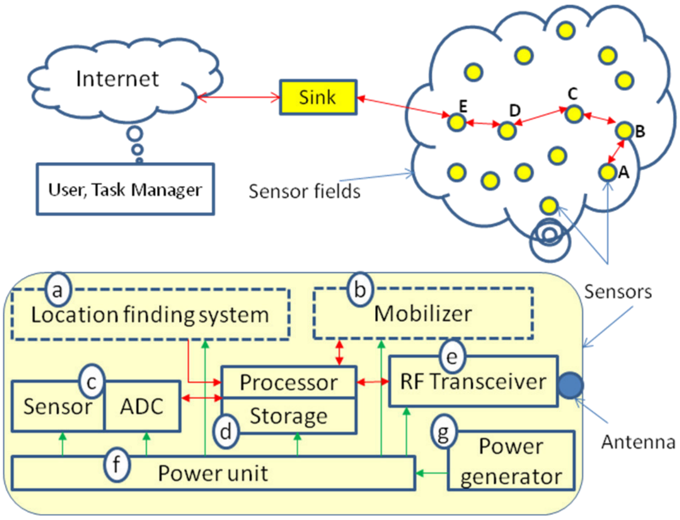


Figure 1: Architecture of Wireless Sensor Network

III. HARDWARE DESIGN

There Wireless sensor nodes (part a to g of Figure 1) are the key part of sensor network. The node is generally made up of data collecting unit (c), data processing (d) and transmitting unit (e). These are three essential parts which are shown in figure 1. The next parts are location finding system (a) and mobilizers (b), which are required for the moving object application. In fact there are two more parts which too play a very important role in wireless sensor node. These are: antenna which is a part of the transmitting unit, and battery (g) which supplies power to all. Data collecting unit is responsible for collecting information within the monitoring area and completing the data conversion from the analog value to its digital value. The wireless sensor nodes achieve the acquisition of different physical quantities by being equipped with different sensor modules. Data processing and transmitting unit consists of two modules: data processing and data transmitting.

The central processing unit (CPU) is responsible for controlling and handling the routing protocol, simultaneous localization and power management of the entire sensor node; The data transmitting module is responsible for conducting wireless communication with other nodes, exchanging controlled messages, sending and receiving the collected data, and so on and so forth. The data transmitting unit mainly consists of corresponding communication protocols (mainly MAC or physical level protocol) and low-power; short range wireless communication module and antenna [4]

Data processing and protocol management units makes use of the program and data memory of PSoC device which integrates with the CPU core with a choice of M8C, 8051 or ARM as shown in Table 1. Data acquisition components like PA (programmable gain amplifier), ADC, multiplexer, programs routine and protocol stack (Tiny OS) can transmit data wirelessly, as further described in the next section. Data processing and transmitting unit is the core of the application specific wireless sensor node, which is responsible for collecting the parameters of various required physical quantities or signals. It requires sampling rate and certain amount of data. The choice of processor is crucial while designing the node.

PSoC family & CyFi is a true and flexible, programmable and reconfigurable system on radio chip (SoC) of CMOS solution [9]. This solution can improve performance and meet the application goal with IEEE 802.15.4 based 2.4GHz ISM band and can meet the request of low cost and low power. It combines with a high performance 2.4GHz DSSS (Direct Sequence Spread Spectrum) RF transceiver CyFi and a compact and efficient PSoC device. The design of the PSoC combines the 8 to 64 Kbyte RAM and powerful library of user modules [10].

Table 1: PSoC Family and features

PSoC 1	PSoC 3	PSoC 4	PSoC 5
Performance optimized 8-bit M8U	Performance optimized single cycle 8-bit 8051 core	High-performance 32-bit ARM Cortex-M0	High-performance 32-bit ARM Cortex-M3
Up to 24 MHz, 4 MIPS Flash 4 KB to 32 KB SRAM 256 bytes -2 KB Operation 1.7 V- 5.25 V	Up to 67 MHz, 33 MIPS Flash 8 KB to 64 KB SRAM 3 KB to 8 KB Operation 0.5 V to 5.5 V	Up to 48 MHz, MIPS Flash 16 KB to 32 KB SRAM 4 KB Operation 1.71 V to 5.5 V	Up to 67 MHz, 84 MIPS Flash 32 KB to 256 KB SRAM 16 KB to 64 KB Operation 2.7 V to 5.5 V
1 Delta-Sigma ADC (6 to 14-bit) 131 ksps @ 8-bit Voltage Precision ±1.53 % Up to two DACs (6 to 8-bit)	1 Delta-Sigma ADC (8 to 20-bit) 192 ksps @ 12-bit Voltage Precision ±0.1% Up to four DACs (8-bit)	1 SAR ADC (12-bit) 1 Msps @ 12-bit Up to 2 DACs (8-bit)	1 Delta-Sigma ADC (8 to 20-bit); 2 SAR ADCs (12-bit) 192 ksps @12-bit;1 Msps @ 12-bit Voltage Precision ±1.0% Up to four DACs (8-bit)
Active: 2 mA, Sleep: 3 µA FS USB 2.0, I²C, SPI, UART	Active: 1.2 mA, Sleep: 1 µA, Hibernate: 200 nA FS USB 2.0, I²C, SPI, UART, CAN, LIN, I²S	Active: 1.6 mA, Sleep: 1.3 µA, Hibernate: 150 nA I²C, SPI, UART	Active: 2 mA, Sleep: 2 µA, Hibernate: 300 nA FS USB 2.0, I²C, SPI, UART, LIN, I²S
Requires ICE Cube and FlexPods	On-chip JTAG, Debug and Trace; SWD, SWV	On-chip JTAG, Debug and Trace; SWD, SWV	On-chip JTAG, Debug and Trace; SWD, SWV
Up to 64 I/O	Up to 72 I/O	Up to 36 I/O	Up to 72 I/O

There are three different architectures for the selection of CPU core. These depend on different members of PSoC families like PSoC-1 with M8C core, PSoC-3 8051 core, PSoC-4 low power Cortex-M0, PSoC-5 Cortex-M3. Each of the members has option of flash memory space such as 32,64 and 128 k Byte and the digital and analog block to optimize the combination of complexity and cost. The device size ranges from 7 × 7mm 48-pin package. It uses the 0.18 & micro CMOS standard technology with embedded flash memory. This can be integrated with different choices of user module which fix either in digital or analog array. The proposed node extends the use of the previous architecture of PSoC-1 to higher families, either to PSoC3 or PSoC-4 chip for improving the processing capability of the sensor node. The separate part CyFi can be the RF front end for any device of the PSoC, memory and microcontroller in a single chip. It uses an 8-bit to 32 bit architecture on selection of PSoC family, with 128-64 KB programmable flash memory and 8-64 KB RAM. It further includes analog-digital arrays for ADC, DAC, Filters, analog amplifiers, timers, PWM, counter, comparator, watchdog Timer, different clock, grounding flexibility Power section, power monitoring, with sleep mode timer, power On Reset, brown out detection, and up to 72 programmable I/O pins. Choice of user module can be added in the hibernate mode for the better performance of power aware sensor node design [2], the detail classification of which is shown in table 1.

This is the latest PSoC-4 device chip using recent and latest CMOS process technology. The current consumption in the receiving and transmitting mode, is lower than 1.2 mA or 25 mA . Sleep mode and the characteristics that can use very short time to complete the mode conversion is especially suitable for the applications which require long battery life.

The main features of the CyFi chip are as follows:

CyFi is a 2.4-GHz direct sequence spread spectrum (DSSS) & GFSK radio transceiver. It operates in the unlicensed worldwide industrial, scientific, and medical (ISM) band (2.400 GHz to 2.483 GHz). Its operating current is 21 mA (transmit at –5 dBm) and sleep current is less than 1 μA. The transmit power is up to +4 dBm and receiver's sensitivity is up to –97 dBm. CyFi can operate in DSSS modulation techniques with data rates up to 250 kbps. With the Gaussian frequency-shift keying (GFSK) modulation it can go up to 1 Mbps data rate. It requires less external component count that is only crystal and antenna marching circuit for impedance matching [10]. CyFi even supports new advanced features like auto transaction sequencer (ATS) - no MCU intervention, framing, length, CRC16, auto acknowledge (ACK), power management unit (PMU) for MCU, fast startup and fast channel changes for hopping. It has separate 16 byte transmitter, and receiver FIFO buffers. It has a dynamic data rate reception, which receives signal strength indication (RSSI). CyFi can communicate with PSoC devices with serial peripheral interface (SPI). It can be controlled while in sleep mode also. It has a microcontroller interface of 4-MHz SPI. It has in built battery voltage monitoring circuitry which can support even coin-cell operated applications because of its operating voltage which ranges from 1.8 V to 3.6 V with temperatures from 0 °C to 70 °C. It is available in very small space saving package 40-pin QFN 6 × 6 mm. The actual node of this configuration is shown in figure 2. It is exactly the size of Nokia rechargeable battery and can fit at the rear side of the node. It can be used as a standard node in the low cost solution for laboratory or field.

This node has a high degree integration however its work architecture is simple (as it is shown in figure 1). If the PSoC device and CyFi chip is coupled with a small number of electronic components and sensor (temperature), it can become a good versatile node as shown in Figure 2 . It will be able to send and receive wireless data. In the figure, the CiFy module leads to two sets of interfaces: one with SPI to PSoC device and the other to antenna. The antenna in the figure is λ / 4 dipole antenna and patch type application specific will be the better option [4], It also can increase communication distance by the way of increasing common used antennas. The λ / 4 dipole patch antenna length can be calculated as:

$$L = 14250 / f$$

The unit of f is MHz, the length unit is cm, therefore the length of 2450MHz antenna is 2.9 cm. The length of the antenna is 2.9cm, as shown in figure 2. If the antenna is shorter, it will affect the communication distance of RF modules. The WSN gateway conducts level translation by the other wire network, this option is also available in PSoC chip, depending on the serial ports and the computer program used to communicate.

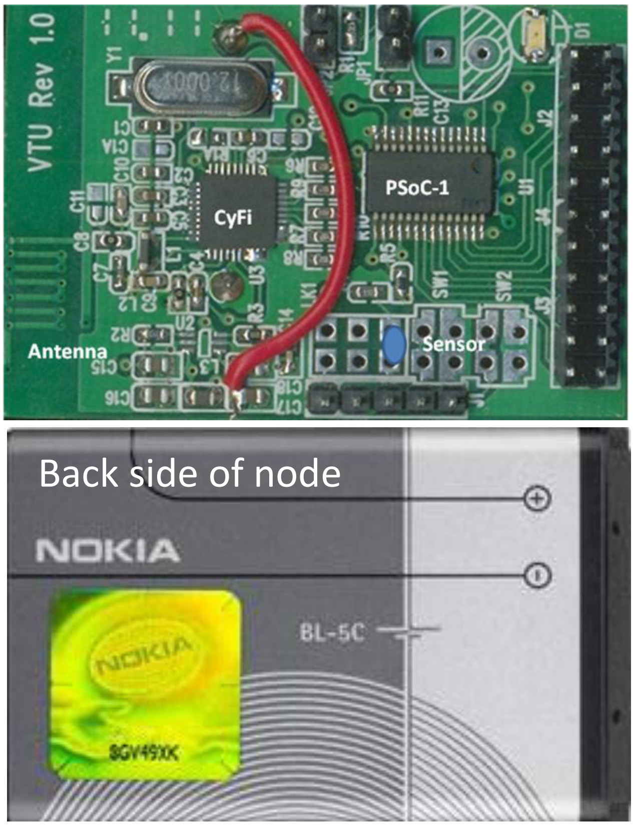


Figure 2: Proposed node

In this way the CyFi module can connect with a number of sensors through PSoC device, which can be selected as per the application requirement like light sensors, temperature sensors, etc. A typical design of the circuit is shown in Figure 4. The program and the user module selection for the application specific wireless sensor node are flexible. It should be noted that all the required components are user selectable and controlled in programs. Even the key parameter like gain of opamp, band pass filter frequency channel freq, channel bandwidth, etc, can be changed dynamically using application specific protocol for communication. The CyFi can support the change in the channel frequency while communication with very narrow band with about 1 MHz. Thus, in the entire ISM band, user can get the felicity hoop channels of an interval of 1 MHz. In other way CyFi can support the 98 channels for communication with 1 MHz bandwidth of each [10].

IV. Software design

Figure 3 shows the software of these types of sensor nodes. These have are four layers. 1. Tiny OS, 2. Application Specific Communication Protocol, 3. Control logic in C for switching the various user modules in PSoC device. 4. Build and configure code as per user routing and selection. Its part one is the compiler environment used by the program, i.e the TinyOS. TinyOS is an open source OS developed by UC Berkeley, which is specifically designed for embedded wireless sensor network. This operating system, based on the architecture of component makes rapid update possible, which in turn reduces the code length [1,3,6]. TinyOS components include network protocols, distributed servers, sensor-driven and data identification tools. It has a good power management due to the event-driven execution model. The model also allows flexibility for scheduling, and can even be applied to several hardware independant platforms.

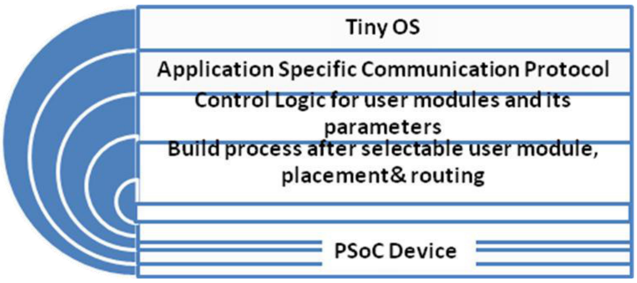

Tiny OS
Application Specific Communication Protocol
Control Logic for user modules and its parameters
Build process after selectable user module, placement& routing
PSoC Device

Figure 3: Layers of software development process

The layered architecture of TinyOS's component-based makes reduction of application possible to be achieved [8]. It only needs to compile the useful components for an application. A specific application usually has a top-level configuration file, completes the connections between component interfaces, and then assembles the component into a specific application. Components contain configuration and module. The achievement of configuration describes the connection of interfaces, commands or events between different components specifically. Top-level is an accessory of configuration file. The achievement of module describes commands and event functions specifically. Interface is a two-way channel between the components. Through stating a set of commands and event function, it does the interfacing for the different functions and the event notification. Interface provider is responsible for improving the interface commands. Interface user is responsible for using interface events. An interface can be provided and called by multiple components.

The part 2 and 3 has to be done in the C programming and it varies as per the application requirement. This software portion of the sensor network nodes is mainly responsible for collecting the physical sensor data and controlling CyFi transceiver modules through instructions received from Tiny OS. This can be done with the help of PSoC creator or designer. The part 4 of the software is the building process in cypress designer or creator, used for generating build code after the selection and placement of user modules in programmable array block of PSoC. As per the requirement of the application, the user can select appropriate user module and can place it at a proper place of diital and analog array. He can then rout the connection between user modules and I/O pins. Even after the code is built, the user programme (C code) can keep the changes in the key parameters dynamically and can improve the performance of the node and network in power prospective, accuracy and so on [2].

The basic idea of the system software programming is that the PSoC devices is initialized, which sets the mux input and gain of programmable amplifier. Sampling clock of the ADC, sets the timer. The then timer does AD conversion on the sensor at regular intervals through the selected ADC input channel , and fills the data into data packets. It then sends the data packets through the CyFi module, and finally falls into sleep. When the set sleep time ends, it repeats the above steps again. When sleeping, if data packets sent by other wireless nodes are received, then it reads the data packets to determine whether they are consistent with the conditions. If so, it transmits the data packets, otherwise discards them.

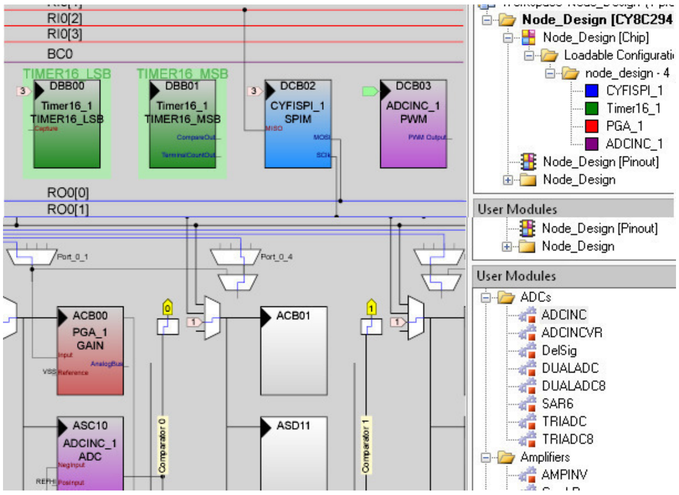


Figure 4: PSoC user module's placement and routing

Figure 4 shows a snap shot of PSoC developer IDE. This is used for PSoC-1 family. Higher family devices, i.e. PSoC 3 to 5 require PSoC Creator IDE. These IDE are different because of theier different architecture of core CPU and related compiler issues. Internal block shown in figure 4 is the mixed array block. It is either digital or analog. These are most flexible, just is like FPGA. These blocks can reconfigure as per the component required (like ADC, DAC, Counter, Timer, PGA etc). This block can be filled by the user module (pre define code for different functions by Cypress library). It has all types of user module right from programmable gain amplifier, instrumentation amplifier, to low pass, band pass filter and so on. These user modules can occupy the appropriate blocks either digital or analog or both. These features are also available in PSoC creator IDE but look wise, it is different. This entire design work bench has the flexibility in routing for input, output and internal connection between blocks. That is the reason, why in this architecture, the pin compatibility between various devices and packages is not an issue. The key pins like power and reset are computable as per the package. All the i/o functions require pinsto be defined by the user.

WSN gateway needs to read commands from the computer through the serial or USB ports and forward the commands. At the same time it even transmits the data packets sent by the wireless sensor nodes to the computer management software through the serial or USB ports, in order to facilitate the data management.

After the CyFi transceiver module receives the instructions from the WSN gateway or other neighboring nodes, the network nodes will be awakened. The processor then determines node number of the instructions. If the object of the instruction is the current node, then the node begins working. Otherwise the network node will be skipped. In addition, the monitoring software also needs to manage the entire network. In general, the monitoring software receives the data WSN gateway through serial ports, completes the processing of data summary and stores the data in the database. The data can then be queried or deleted as and when required, on the basis ofnd querying of each node's working status.

V. Conclusion

Although there are few sensor nodes available as CTOS but none of these have this type of flexibility to improve the performance of wireless sensor node and its network. This node supports the flexibility at almost all layer of technology: from gain of the sensor signal amplifier to the frequency hopping in wireless communication channels. It does this with the help of software control. The few nodes mentioned above are developed and are under testing with application and are just waiting for the final approved result from laboratories. This wireless sensor node and its networks has been developed in a different domain of application. But its development and application in a practical life will be perfect and comprehensive. Therefore, the design of application specific wireless sensor nodes based on the PSoC and CyFi, does good to the further research and development of Ad-Hoc sensor network, and it may as well play a significant role in the field of WSN.

References

[1] Sujian Zhao et al “Software Design of a Simple Wireless Sensor Network System Based on CC2430” Wireless Communications, Networking and Mobile Computing (WiCOM), 2011 7th IEEE International Conference. PP 1-4 2011.

[2] Sham P Nayse, Mohammad Atique, Anita Agrawal et al, “Power Aware Wireless Sensor Node Design Using PSoC” International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 4 PP 1178-1182,April 2013.

[3] El Kateeb, A. et al, “Wireless Sensor Nodes Processor Architecture and Design” Advanced Information Networking and Applications – 2nd IEEE International Conference. PP 892 – 897, 2008

[4] Ching-Wen Chen et al, “Design of a Low-Power and Low-Latency MAC Protocol with Nodes Grouping and transmission Pipelining in Wireless Sensor Networks; PP 143 - 147 2008

[5] Bo Chen ; Tomizuka, M. “Open Architecture Design of Structural Health Monitoring Software in Wireless Sensor Nodes” Mechtronic and Embedded Systems and Applications, International Conference PP 19 - 24 2008.

[6] Feng Tian ; Xiaowu Xu “Design of Wireless Sensor Networks Node in Coalmine” Intelligent Computation Technology and Automation, Second IEEE International Conference, Vol. 4 PP 66 - 69 2009.

[7] Dzung, D. et al, “Design and implementation of a real-time wireless sensor/actuator communication system” Emerging Technologies and Factory Automation, 10th IEEE Conference Vol.2 PP 442-2005

[8] Sasilatha, T. ; Raja, J. “Design and analysis of a 1.2V, 2.4 GHz Low Power ASK Transmitter for Wireless Micro Sensor Nodes" Wireless Communication and Sensor Networks, Third International PP 17 - 20 , 2007

[9] Doboli, A. et al; "Dynamic reconfiguration in a PSoC device",. FPT IEEE International Conference PP 361-363, 2009

[10] Data sheet and technical manual Cypress Inc.

DOI: 10.15439/2014F176
ACSIS, Vol. 3

The concept of authentication in WSNs using TPM

Janusz Furtak
Military University of Technology
ul. Kaliskiego 2,
00-908 Warszawa, Poland
Email: jfurtak@wat.edu.pl

Jan Chudzikiewicz
Military University of Technology
ul. Kaliskiego 2,
00-908 Warszawa, Poland
Email: jchudzikiewicz@wat.edu.pl

Abstract— This document describes how to use the Trusted Platform Module (TPM) to authenticate sensors in wireless sensors network which create a sensors' domain. Model of the wireless sensor network is presented. There are three types of nodes in the domain. The M node is an authentication authority in sensors' domain – it stores credentials of all nodes of domain. The M node is also the recipient of the data emitted by the domain sensors. The S node is the source of sensors data (i.e. air temperature, concentration of sulfur dioxide, wear of ammunition, etc.). The rM node is acting as backup for M node. The concept of main operations available in the sensors' domain related to: managing of sensors in the domain, authentication of sensors and regeneration of the node credentials is presented. The concept is a proprietary solution developed by the authors of the paper.

I. Introduction

The wireless sensor networks (WSNs) consist of large number of ultra-small, low-power and inexpensive wireless sensor nodes with sensing, computing and communication capabilities [1], [2]. The popularity of the WSNs causes that are used in many areas like for example: military, ecological, health-related areas etc. These applications often include the monitoring and processing of sensitive information or location of soldiers on the battlefield. Security is therefore important in WSNs. We need use secure communication mechanisms in WSN to ensure confidentiality, authenticity and integrity of the nodes and data. Security mechanisms deployed in WSNs should involve collaborations among the nodes due to the decentralized nature of the networks and absence of any infrastructure. The situation becomes critical when the nodes are equipped with cryptographic materials such as keys and other important data in the sensor nodes. Moreover, adversaries can introduce fake nodes similar to the nodes available in the network which further leave the sensor nodes as un-trusted entities.

The researchers in WSN security have proposed various security schemes which are optimized for these networks with resource constraints. A number of secure and efficient routing protocols [3], [4], secure data aggregation protocols [5], [6], [7], [8] and additional security mechanisms such as Trusted Platform Module (TPM) [9], [10], [11], [12] etc. has been proposed by several researchers in WSN security. Taking this into consideration WSNs among others could be divided into different security levels [13], [14]:

- **Availability**, which ensures that the desired network services are available even in the presence of denial-of-service attacks;
- **Authorization**, which ensures that only authorized sensors can be involved in providing information to network services;
- **Authentication**, which ensures that the communication from one node to another node is genuine, that is, a malicious node cannot masquerade as a trusted network node;
- **Confidentiality**, which ensures that a given message cannot be understood by anyone other than the desired recipients
- **Integrity**, which ensures that a message sent from one node to another is not modified by malicious intermediate nodes;
- **Nonrepudiation**, which denotes that a node cannot deny sending a message it has previously sent;
- **Freshness**, which implies that the data is recent and ensures that no adversary can replay old messages.

Most of them are very important for military applications. Secrecy is part of its nature; and data (sensed/ aggregation/processing) is required to remain confidential. This is critical to the successful operation of a military application. Enemy tracking and targeting are among the most useful applications of wireless sensor networking in military terms. Considering the above, the secure method of transmitting and storing data in WSNs is proposed in the paper. The Trusted Platform Module (TPM) is the basis of the presented method. A TPM is used for secure storing the necessary data to authenticate the nodes, and generate symmetric keys, and asymmetric keys (private/public).

The second section provides proposed architecture of WSNs, and basic definitions. The basic data stored in every nodes (depending on the role they played in the network e.g. domain master (node M), and slave (node S)), and the basic data structures used in the nodes are defined in the section. In the third section the procedures to ensure proper authentication of sensors in domain and correct data transfer between sensors are described. Finally, a few concluding remarks are presented.

This concept is a proprietary solution developed by the authors of the paper. Some inspirations for the development of the presented concept the authors of the method drew from solutions used in DNS.

II. The model of Wireless sensor network with authentication

In the domain of sensors there are two authorities. The first is the node (Data Collector) which is the recipient of the data emitted by the domain sensors. The node which manages the Root of trust is the second authority. The Root of trust is to be used to authenticate all sensors involved in the exchange of data between elements of the domain of sensors. The second authority is to act as a master of domain and will be called the node M. The presented concept assumes that both the role of the recipient of data from the sensors (i.e. Data Collector) and the role of the master of domain plays the same node.

In the sensors' domain is exactly the one node that acts as the domain master (node M). To this domain belong sensors of type slave (nodes S), which are registered by the node M. Nodes S are the source of data. Node S is initiated and authenticated by node M of domain. Node M stores the root trust of sensors' domain. The sensors' domain structure is shown in **Fig. 1**

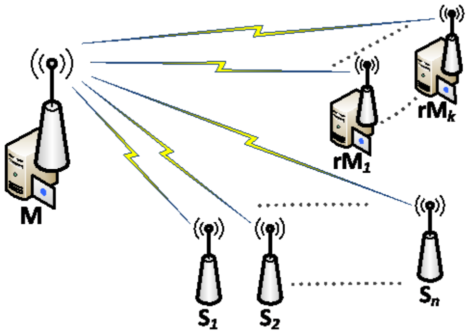


Fig. 1 The structure of sensors' domain

In the domain may be designated nodes acting as backups masters (replicas of master - rM). Such a node may be a S node after the establishment the role rM for him, on condition that its hardware and software resources provide this capability. In the domain may be no node type rM (this is not recommended), but there may be a few such nodes. The task of node rM is to store a copy of the root of trust from the node M of domain. Updating the copy should be done by one method selected from the following [1]:

- after the modification of the root of trust on the node M;
- at fixed intervals of time;
- on demand of node M;
- on demand of node rM.

From the viewpoint of authentication procedures nodes M and rM for nodes S are the same. Node rM can become a new node M of domain after changing its role, due to proven inactivity of old node M. In this case the node, which has so far acted as a node M, becomes a node rM, or node S, or is removed.

When the sensor does not function, is turned off or damaged, it is assumed that this node is in a non-active state, and when the sensor is functioning, then the node is in the active state.

If a node acts as M and remains in a non-active state for longer than a predetermined period of time, the procedure for the designation of the "new" master of the nodes is started. The "new" master is designated from among nodes, which until now were playing the role of rM. When in the domain there are no nodes rM, the new master shall be designated among the S nodes. A node that has lost the role M as a result of prolonged inactivity, but retains the efficiency, after obtaining the active state can act as rM or only S (it might be needed restart the procedure for initiating node).

If a node acts as rM and remains in a non-active state for longer than a predetermined period of time, and there are no others nodes rM in the domain, procedure for designating "new" node rM from nodes S is started. A node that has lost the role rM as a result of prolonged inactivity, but retains the efficiency, after obtaining the active state can act as rM or only S (it might be needed restart the procedure for initiating node).

Sensor, which acts as a node M receives data from S nodes.

Minimum requirements for a sensor type S are as follows:

- sensor must be equipped with a TPM (see the next section);
- sensor must have an interface that allows direct connection to the node M (e.g. via USB) in the registration procedure of the node in the domain;
- the ability to send sensor data (i.e. measurement data) to M node using only wireless connection.

In order to enable automatic authentication procedure of the node and regeneration procedure for S node credentials, S node should be able to receive data transmitted by node M via a wireless connection. Otherwise, the node authentication procedure is not possible and change of credentials of this node will be possible only after the re-registration of the node. Nodes that are designed to play the role of M or rM must be able to bi-directional communication with other nodes, and should also have adequate resources in terms of power, processing capability and storage capacity.

A. Trusted Platform Module

In the presented model for authentication sensors are used mechanisms offered by the Trusted Platform Module (TPM). It is assumed that each element of the domain of sensors is equipped with TPM.

[1] Analysis of the advantages and disadvantages and the choice of how to upgrade the resources of nodes rM are not subject of this study.

TPM is an implementation of a standard developed by the Trusted Computing Group [15]. This module is designed to support the cryptographic procedures and protocols that can be used for securing data [16]. Trusted Platform Module provides the following functions:

- generating an asymmetric key pair,
- secure storage of keys,
- generating an electronic signatures,
- encryption and decryption,
- implementation of an operation defined by the standard PKCS #11.

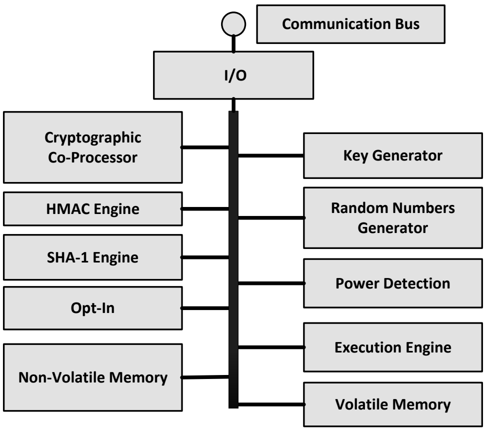


Fig. 2 TPM Component Architecture (based on [15])

The following algorithms are typically implemented in TPM [17]: RSA, SHA-1, HMAC and AES[2]. In addition, each TPM chip stores a unique serial number and its RSA private key that is never available to read. TPM components are shown in **Fig. 2**.

B. Resources of sensors

Each sensor is equipped with a TPM. In the resources of TPM are stored the necessary data to authenticate the node acting as the S in domain. The structure of the data is shown on **Fig. 3**. Sensors, which are to play the role of M or RM must be equipped with additional memory, which is intended to store the description of the domain and descriptions of remaining domain nodes.

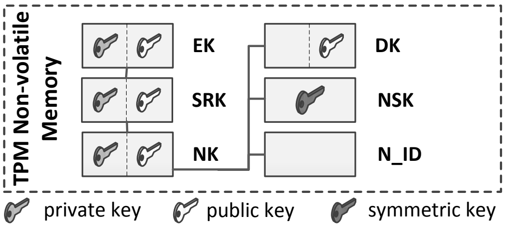


Fig. 3 The data stored on S node

Content of credentials stored in non-volatile memory of the TPM, which are used by a node S (Slave data):

- EK (Endorsment Key) - key pair (private/public) generated in the production phase of the TPM;
- SRK (Storage Root Key) - key pair (private/public) generated during the process of initiating the TPM in the procedure for registering a S node in the domain of sensors;
- NK (Node Key) - key pair (private/public) of node; generated during the procedure for registering a S node in the domain of sensors; acts as the parent for the remaining keys stored in the resources of the node;
- DK (Domain Key) – public part of the key of sensors' domain to which the node belongs; obtained during the procedure for registering the node in the domain;
- NSK (Node Symmetric Key) – symmetric key to encrypt the data sent from this node to M node; obtained during the procedure for registering the node in the domain and renovated in the regeneration procedure of S node credentials;
- N_ID (Node ID) – ID of the sensor (e.g. IPv6 link local address of sensor).

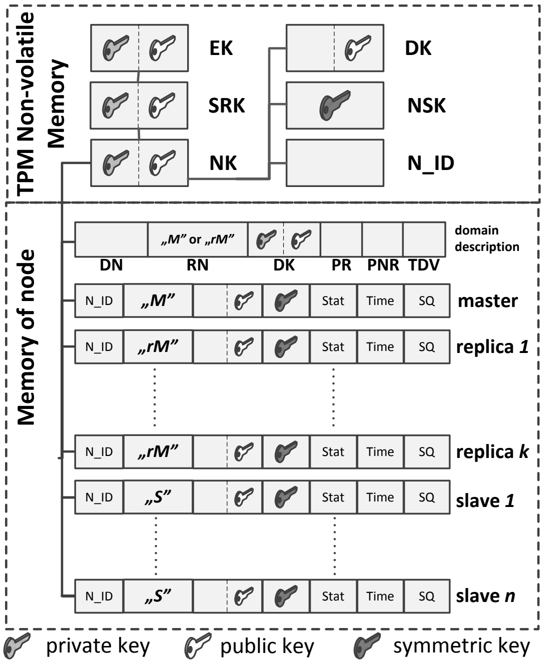


Fig. 4 The data stored on M node or rM node

Credentials stored by the node M (the structure of the data is shown on **Fig. 3Fig. 4**) consist of three resources: Slave Data, Domain description and Description of domain nodes. The content of these resources is as the following:

- **Slave data** (the same as for the S node);
- **Domain description**:
  - DN (Domain Name) – the name of domain;
  - RN (Role of Node) – determine whether below data are the resource of master node or the resource of replica of master; it is synonymous with the role it plays in the domain; may have one of values: M or rM;
  - DK (Domain Key) – key pair (private/public) of sensors' domain; generated during the procedure generated in the process of creating the domain

[2] TPM uses a symmetric algorithm AES to protect the confidentiality of the session in which it participates following the recommendations of the TCG. However, symmetric encryption functions are not normally accessible outside the TPM.

of sensors and establishing the role of the "master" in the domain for the first node;

- PR (Period of Replication) – the time after which the rM node is required to establish communication with the node M and refresh the domain data;
- PNR (Period of Non-success Replication) - the time after which the node rM is obliged to repeat the attempt to establish communication with the M node if the previous attempt refreshing the domain data was not successful;
- TDV (Time of data validity) – after this time and the inability to refresh, the domain data are invalid and node becomes a node S.

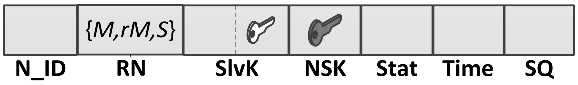


Fig. 5 The data structure describing a node

- **Description of domain nodes**. Description of each node contains the following data (the structure of the data is showed on Fig. 5):
  - N_ID (Node ID) – ID of the sensor;
  - RN (Role of Node) – the role filled by the node in the domain; it can take values from the set {M, rM, S};
  - SlvK - public part of an asymmetric key N_ID node of sensors' domain;
  - NSK - symmetric key to encrypt the data sent from this node to M node; obtained during the procedure for registering the node in the domain and renovated in the procedure for the regeneration of S node credentials;
  - Stat - status of the node; it can take one of the values: non-active(-1), active(0), active non-confirmed (n), where n is the number of consecutive unsuccessful attempts to establish communication with the node
  - Time - moment of the last and the effective transmission [3];
  - SQ - the sequence number of the last sent frame (modified after each message).

Data EK, SRK, NK, DK, NSK and N_ID are stored in non-volatile memory of the TPM, and the remaining data are stored in the sensor resource and secured using the NK key.

III. Operations in the wireless sensor network with authentication

In order to ensure proper authentication of sensors in domain and correct data transfer between sensors, in the domain should be available the following procedures:

1. Procedure for initiating M node.
2. Procedure for registering the S node in the domain of sensors.
3. Procedure for removing rM or S node from the sensors' domain.
4. Authentication procedure of the node.
5. Integration test of nodes in sensors' domain.
6. Procedure for the regeneration of S node credentials.
7. Procedure of sending data from S node to M node.
8. Procedure of reading data on M node which were received from S node.
9. Procedure for giving role rM in the domain for S node.
10. Procedure for updating resources of rM node based on resources of M node.
11. Procedure for changing the node role from rM role to role M;
12. Procedure for determining the "new" node M after the failure of the "old" node M.
13. Integration test of resources of M and rM nodes.

In this study in the following sections are comprehensively described the procedures listed in paragraphs 1-8. Other functions related to the management of the nodes being the "replicas of the master" are a subject of another study.

A. The procedure for initiating M node

This procedure is intended to create the domain of sensors and to initiate the node that will serve as the master of the domain.

Input data:

- N_ID - node identifier;
- DN - sensors' domain name;
- time periods (i.e. PR, PNR and TDV) associated with the operation of nodes rM..

The procedure for initiating M node comprises the following steps:

1. Take ownership of the TPM and SRK key generation.
2. Generate asymmetric key (NK) and symmetric key (NSK) for the node.
3. Put NK, NSK and N_ID into non-volatile memory of the TPM.
4. Generate asymmetric key (DK) for sensors' domain and put the public part of that key in non-volatile memory of the TPM.
5. Prepare of the domain description, which includes the fields DN, RN, DK, PR, PNR, TDV and then wrap this description using the public portion of the NK key. The RN field should have a content of "M".
6. Prepare of the M node description and then wrap this description using the public part of the NK key. The fields of the description should have the following values:
   N_ID = input data N_ID
   RN = „M"
   SlvK = public part of the node NK key
   NSK = the node NSK key

[3] It was assumed that Time field is modified each time the field SQ is modified. In order not to complicate the understanding of the procedures outlined in the following sections, this field has not been included in these procedures.

Stat = 0
Time = current time
SQ = random number from the range <0; 65535>.

7. Save the M node description in M node resources.

B. The procedure for registering the S node in the domain of sensors

In the procedure of registration S node in the domain is assumed that during this procedure S node is connected to the node M via the USB interface[4].

Input data:

- N_ID - node identifier;
- public part of the DK key.

The procedure for registering S node in the domain comprises the following steps:

1. Install S node in USB port of M node.
2. Take ownership of the TPM and SRK key generation.
3. Generate asymmetric key (NK) and symmetric key (NSK) for the node.
4. Put NK, NSK and N_ID into non-volatile memory of the TPM of S node.
5. Obtain the public part of the DK key from non-volatile memory of the TPM of M node and save it into non-volatile memory of the TPM of S node.
8. Prepare of the S node description and then wrap this description using the public portion of the NK key. The fields of the description should have the following values:
   N_ID = input data N_ID
   RN = „S”
   SlvK = public part of the S node NK key which is registered
   NSK = the NSK key of node which is registered
   Stat = 0
   Time = current time
   SQ = random number from the range <0; 65535>.
6. Save the S node description in M node resources.
7. Uninstall the S node from USB port of M node

C. The procedure for removing rM or S node from the sensors' domain

The procedure for removing a node is technically quite simple activity. A bigger problem is the answer to the question: under what conditions make this activity? The problems from this area that should be resolved include the following:

- Is the node is removed after one ineffective node authentication procedure (or maybe after n unsuccessful attempts)?
- After how many unsuccessful attempts to authenticate the node is removed?
- How long the node can remain in a non-active state before it is removed?

The following procedure does not take into account the identified problems. It is assumed that decisions on the above issues have already been taken, and the procedure can be performed only by the node M.

Input data:

- N_ID - identifier of node to remove
- Description of N_ID node recorded in the tree of trust stored on resources of M node.

The procedure for removing rM or S node from the sensors' domain comprises the following steps:

1. Prepare a remove packet (Optional[5]):

remove packet

code	id	empty	name

   where:
   code = 5 for remove packet;
   id – SQ field from description of removed node;
   empty - zeroed field;
   name -identifier of removing node (i.e. M node).
2. Wrap the remove packet with the SlvK key of N_ID node and send the packet from M node to rM or S node.
3. Remove the N_ID node description from resources of M node.

D. The authentication procedure of the node.

The procedure may be initiated by node M (authenticator) to confirm the identity of the node rM or S, and can also be initiated by the node S to confirm the identity of the node M. The procedure is based on PPP Challenge Handshake Authentication Protocol (CHAP)[18].

Input data:

- N_ID - identifier of node to check;
- Description of N_ID node recorded in the tree of trust stored on resources of M node.

The authentication procedure of the node S or rM initiated by node M comprises the following steps:

1. Prepare a challenge packet:

challenge packet

code	id	rand	name

   where:
   code = 1 for challenge packet;
   id – SQ field from description of checked node;
   rand - random number from the range <0; 65535>;
   name -identifier of checking node (i.e. M node).
2. Increment the stat field in description of checked node.
3. Wrap the challenge packet with the SlvK key of checked node (optional)[6].
4. Send the packet from M node to rM or S node.

[4] If it was not possible to use the USB interface, in order to ensure the safety of the registration procedure, is required to develop additional ways of mutual authentication of both parties involved in the registration.

[5] If the removed node is to be informed of the fact of the removal of that node from the sensors' domain, the steps 1 and 2 of the procedure are required.

[6] Given the limitations in terms of energy consumption and shortage of computing power of sensor you can skip steps 3, 6, 10, and 13, but then the packets will be sent in clear text and it will be the security vulnerability of the system.

5. Receive the challenge packet on checked node from M node (unwrap the packet with the private part of NK key, if needed) and prepare a response packet:

   response packet

code	id	hash	name

   where:
   code = 2 for response packet;
   id – **id** field from challenge packet;
   hash - value of hash function (SHA-1) determined for concatenation of the following fields:
   - **id** field from challenge packet;
   - **rand** field from challenge packet;
   - Symmetric Key (NSK) of checked node;

   name -identifier of checked node.
6. Wrap the response packet with the private NK key of node.
7. Send the response packet to M node.
8. Receive the response packet on M node (unwrap the packet with the SlvK key of N_ID node, if needed). Verify data by comparing the value of hash field from response packet and value determined for concatenation of the following fields:
   - **id** field from challenge packet;
   - **rand** field from challenge packet;
   - NSK field from description of checked node.
9. If authentication is successful, the stat field in description of checked node is zeroed, the SQ field in description of checked node is incremented:

   success packet

code	id	ok	name

   where:
   code = 3 for success packet;
   id – **id** field from challenge packet;
   ok - success message for checked node;
   name -identifier of checking node (i.e. M node).
10. Wrap the success packet with the SlvK key of checked node.
11. Send the packet from M node to rM or S node.
12. If authentication fails, a failure packet is send to checked node:

    failure packet

code	id	fail	name

    where:
    code = 4 for failure packet;
    id – **id** field from challenge packet;
    fail - failure message for checked node
    name -identifier of checking node (i.e. M node).
13. Wrap the failure packet with the SlvK key of checked node.
14. Send the packet from M node to rM or S node.
15. In case of receiving the success package on checked node, the Sequential Number (SQ of checked node) of last send packet is incremented. In other case (i.e. receiving the failure packet or not collecting any package) the SQ of checked node is not modified.

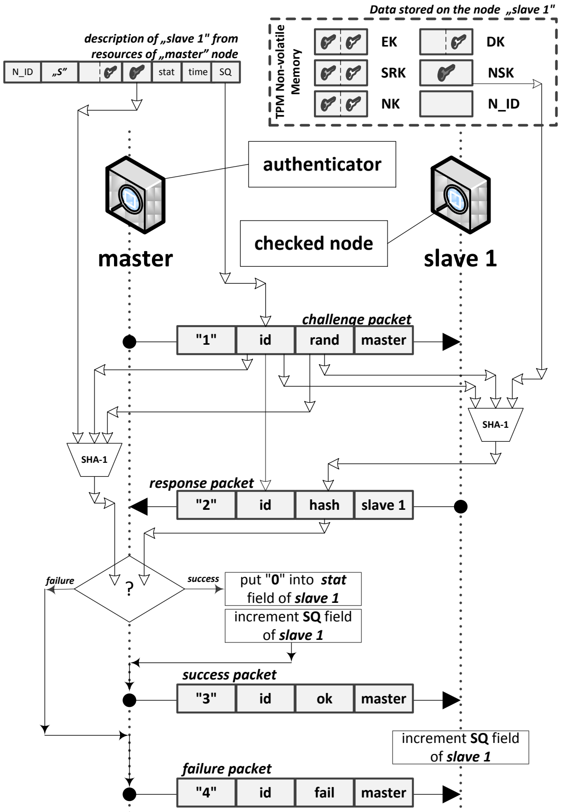


Fig. 6 The authentication procedure of *"slave 1"* node initiated by "master" node (packages are sent in clear text).

Authenticator (i.e. the node which initiates the authentication procedure) controls the frequency and timing challenges. If the above described procedure is initiated by the node S, S node takes over the authenticator role. The authentication procedure of *"slave 1"* node initiated by "master" node is showed on Fig. 6.

E. The integration test of nodes in sensors' domain

The integration test of nodes in sensors' domain is initiated by the node M. This procedure involves running the authentication procedures for all sensors' domain nodes whose descriptions are stored in the resource of node M. The procedure is to be run on demand.

F. Procedure for the regeneration of S node credentials

Procedure for the regeneration of S node credentials is initiated by node M in one of the following cases:

- overflow the sequence number SQ;
- after exceeding a fixed number of packets sent between nodes M and S;
- after a fixed time interval of validity of credentials.

The procedure can also be run on demand and then can be initiated either from the node M and node S.

Input data:

- N_ID - identifier of node to check;
- Description of N_ID node recorded in the tree of trust stored on resources of M node.

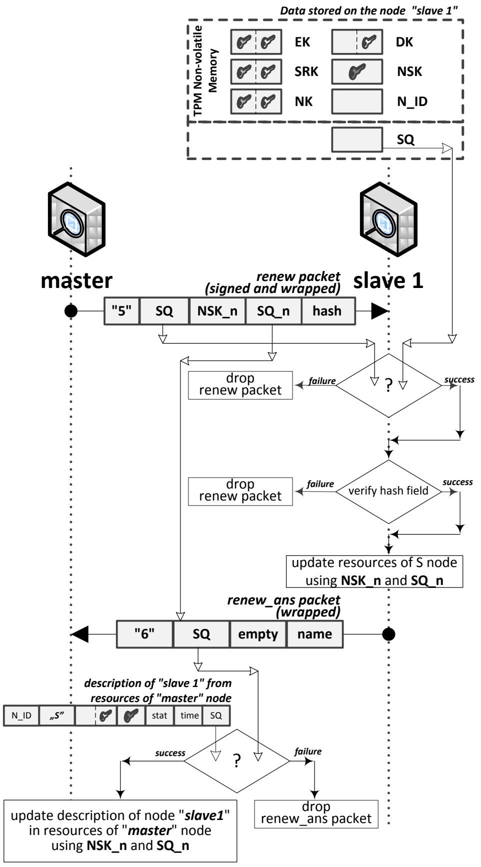

Data stored on the node "slave 1"
TPM Non-volatile Memory
EK
SRK
NK
DK
NSK
N_ID
SQ
master
slave 1
renew packet (signed and wrapped)
"5"
SQ
NSK_n
SQ_n
hash
drop renew packet
failure
?
success
drop renew packet
failure
verify hash field
success
update resources of S node using NSK_n and SQ_n
renew_ans packet (wrapped)
"6"
SQ
empty
name
description of "slave 1" from resources of "master" node
N_ID
„S"
stat
time
SQ
success
?
failure
update description of node "slave1" in resources of "master" node using NSK_n and SQ_n
drop renew_ans packet

Fig. 7 The procedure for the regeneration of *"slave 1"* node credentials initiated by "master" node.

The procedure for regeneration of S node credentials initiated by node M comprises the following steps (Fig. 7):

1. Prepare a renew packet:

renew packet

code	SQ	NSK_n	SQ _n	hash

where:
code = 5 for renew packet;
SQ = current SQ incremented by 1
NSK_n – new symmetric NSK key for S node;
SQ_n – new sequential number for S node;
hash - value of hash function (SHA-1) determined for concatenation of **NSK_n** and **SQ_n** fields.

2. Sign **NSK_n**, **SQ_n**, and **hash** fields using private part of DK key.
3. Wrap renew packet using Slvk of S node.
4. Send the packet to S node.
5. Receive the renew packet on S node and unwrap the packet using private part of NK key of S node.
6. Compare SQ field from renew packet and SQ of S node. If not equal, drop packet.
7. Unsign **NSK_n**, **SQ_n**, and **hash** fields using public part of DK key.
8. Verify data by comparing the value of hash field from renew packet and value determined for concatenation of the **NSK_n** and **SQ_n**, fields.
9. If not a success, drop packet, otherwise update resources of S node using **NSK_n** and **SQ_n** fields and prepare renew_ans packet:

renew_ans packet

code	SQ	empty	name

where:
code = 6 for success packet;
SQ – **SQ_n** field from renew packet;
empty - zeroed field;
name -identifier of checking node (i.e. M node).

10. Wrap renew_ans packet using public part of NK key of S node and send the renew_ans packet to M node.
11. Receive the packet on M node, unwrap it using private part of NK key, and verify SQ field. If success, update the description of S node in resources of M node on basis of NSK_n and SQ_n fields.

G. The procedure of sending data from S node to M node

Input data:

- N_ID – identifier of node;
- SD – sensor's data
- NSK – symmetric key of S node
- DK – public part of domain key.

sensor packet

code	N_ID	SD	SQ	Hash

Fig. 8 The structure of the frame containing the sensor data

The structure of the frame containing the sensor data is showed on Fig. 8. It includes the following fields:

code = 7 for sensor packet
N_ID = input data N_ID
SD = Sensor's Data encrypted with the NSK
SQ = current SQ incremented by 1;
Hash = the value of the hash function determined on the basis of fields N_ID, SD and SQ

The procedure of sending data from S node to M node comprises the following steps:

1. Preparing of the frame containing the sensor data, as shown on Fig. 8.
2. Encrypting of the frame using the public part of DK.
3. Sending the frame to M node;
4. Incrementing SQ field in resources of S node.

H. The procedure of reading data on M node which were received from S node.

Input data:

- Received frame from S node;
- Resources of M node.

The procedure of receiving data on M node from S node comprises the following steps:

1. Receiving of the frame, as shown on Fig. 9.
2. Unwrapping of the frame using the private part of DK.
3. Searching the description of N_ID node in resources of node M. If not a success, the N_ID node is unrecognized.

4. Comparing SQ field from received frame and SQ field from node description. If not equal, the SQ is incorrect.
5. Updating the description of N_ID node:
   stat = 0
   Time = current time
   SQ = **SQ**+1
6. Decrypting of the SD field using the NSK of “slave 1” node.

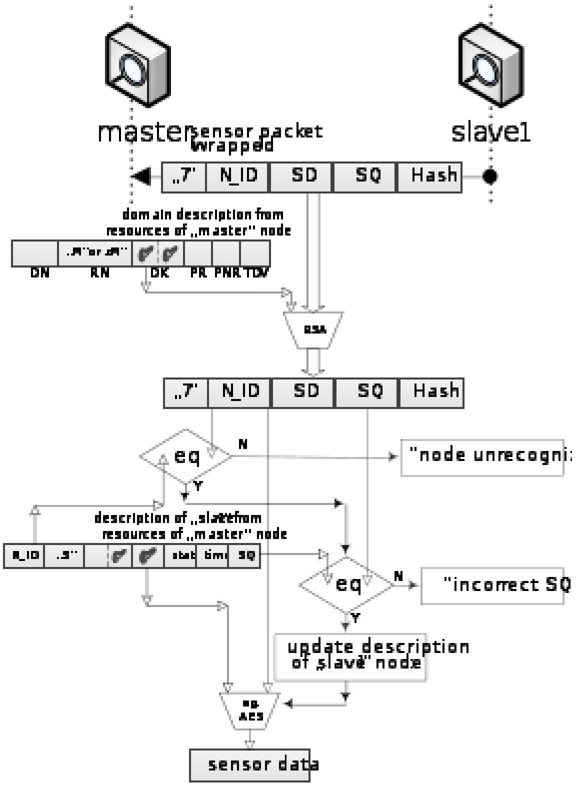


Fig. 9 The procedure of reading data on “master” node which were received from “slave 1” node

IV. Conclusion

This paper presents the model and concept of authentication in sensors' domain. For this purpose, the mechanisms provided by the TPM are used.

Solutions related to authentication elements involved in secure exchange of data require effective exchange of keys between these elements while maintaining the ability to communicate between these elements. The problem of sensors authentication is of great importance for the wireless sensor networks especially that they are mostly employed in critical areas.

On the other hand, the nature of nodes in WSNs gives rise to constraints such as limited energy, processing capability, and storage capacity. The selection of the appropriate cryptographic methods depends on the processing capability of sensors, indicating that there is no universal solution for all networks of sensors.

Taking this into consideration the TPM use is proposed for managing the root of trust and as a tool for securing the data exchange between sensors. Use of TPM will enable credentials processing by the hardware. Most of the operations is done by M node in the domain. The M node receives all the data from sensors and is used to authenticate the rest of nodes. The other nodes do not need large resources. Further work will aim at implementation of the proposed method in built model of WSN and then verifying its properties in a real environment.

References

[1] K. Sohraby, D. Minoli, T. Znati, „Wireless Sensor Networks Technology, Protocols, and Applications”, Wiley, New Jersey 2007, DOI: 10.1002/047011276X.

[2] R. Faludi, “Building Wireless Sensor Networks”, O’Reilly, 2011.

[3] A. Perrig et al., “SPINS: Security Protocols for Sensor Networks”, Wireless Networks, vol. 8, no. 5, Sept. 2002, pp. 521–34, DOI: 10.1023/A:1016598314198.

[4] Boyle D., „Securing Wireless Sensor Networks: Security Architectures”, Journal Of Networks, Vol. 3, No. 1, January 2008, pp.65-77.

[5] A. Al-Dhelaan, “Pairwise Key Establishment Scheme for Hypercube-based Wireless Sensor Networks”, Recent Researches in Computer Science.

[6] Y Mohd Yussoff, H. Hashim, M. Dani Baba, “Identity-based Trusted Authentication in Wireless Sensor Network”, International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.

[7] L. Hu and D. Evans, “Secure Aggregation for Wireless Networks,” Wksp. Security and Assurance in Ad Hoc Networks, 2003.

[8] B. Przydatek, D. Song, and A. Perrig, “SIA: Secure Information Aggregation in Sensor Networks,” SenSys '03: Proc. 1st Int'l. Conf. Embedded Networked Sensor Systems, New York: ACM Press, 2003, pp. 255–65, DOI: 10.1145/958491.958521.

[9] W. Hu, H. Tan, P. Corke, W. Chan Shih, S. Jha, “Toward Trusted Wireless Sensor Networks”, ACM Transactions on Sensor Networks, Vol. 7, No. 1, Article 5, August 2010, DOI: 10.1145/1806895.1806900.

[10] C. Krauß, F. Stumpf, C. Eckert, “Detecting Node Compromise in Hybrid Wireless Sensor Networks Using Attestation Techniques”, Lecture Notes in Computer Science Volume 4572, Springer-Verlag Berlin Heidelberg 2007, pp. 203–217, DOI: 10.1007/978-3-540-73275-4_15.

[11] J. Furtak, T. Pałys, J. Chudzikiewicz, “How to use the TPM in the method of secure data exchange using Flash RAM media”, Proceedings of the Federated Conference on Computer Science and Information Systems, 2013, pp. 831–838.

[12] Hu W., Corke P., Chan Shih W., Overs L., „secFleck: A Public Key Technology Platform for Wireless Sensor Networks”, Wireless Sensor Networks, Lecture Notes in Computer Science Volume 5432, 2009, pp 296-311, DOI: 10.1007/978-3-642-00224-3_19.

[13] Y. Wang, G. Attebury, B. Ramamurthy, “A survey of security issues in Wireless sensor networks”, IEEE Communications Surveys & Tutorials, , Volume 8, No. 2, 2ND Quarter 2006, DOI: 10.1109/COMST.2006.315852.

[14] J. Sen, “A Survey on Wireless Sensor Network Security”, International Journal of Communication Networks and Information Security, Vol. 1, No. 2, August 2009.

[15] *TPM Main Part 1 Design Principles. Specification Version 1.2. Revision 116,* Trusted Computing Group, Incorporated, 2011

[16] *TCG Software Stack (TSS) Specification Version 1.2 Part1: Commands and Structures* (http://www.trustedcomputinggroup.org /files/resource_files/6479CD77-1D09-3519-AD89EAD1BC8C97F0 /TSS_1_2_Errata_A-final.pdf).

[17] S. Kinney, “Trusted platform module basics: using TPM in embedded systems”, Embedded Technology Series,Elsevier Inc., 2006

[18] Simpson, W., "PPP Challenge Handshake Authentication Protocol (CHAP)", RFC 1994, August 1996.

DOI: 10.15439/2014F155
ACSIS, Vol. 3

Toward implementing an efficient gateway for wireless sensor networks

Gundars Miezitis,
Riga Technical University
Faculty of Computer Science and
Information Technology
Meza str.1/321 LV-1007 Riga,
Latvia
E-mail: gundars.miezitis@rtu.lv

Romans Taranovs,
Riga Technical University
Faculty of Computer Science and
Information Technology
Meza str. 1/3-322 LV-1007 Riga,
Latvia
Email: romans.taranovs@rtu.lv

Abstract—**Timely and energy efficient data delivery is important in wireless sensor network applications. To reduce the probability of wireless sensor network disconnection from user a lower energy consumption gateway must be used, but data delivery speed should be maintained as high as possible. So our purpose is to compare and analyze common approach of building gateway for wireless sensor network applications, where high performance ARM boards are used as gateways, with less common approach, where simple eight bit microcontroller boards can work as gateway. We develop two similar test-beds using two available boards – DiGi Wi-9C (high-end) and ATXmega Xplained-A1 (low-end). We test both boards using the same data processing algorithm and by measuring delivery speed and energy consumption we make conclusions. Contrary to our expectations simple eight bit microcontroller showed even better results than we had expected. While this board consumed less energy it guaranteed faster and more stable (little or no delivery speed deviations) data delivery. Thus we concluded that using simpler hardware can not only reduce energy consumption, but ensure high data delivery speed as well.**

Keywords—**sensor network, gateway, data transmission speed, energy efficiency**

I. Introduction

Wireless sensor networks (WSN) enable subtle monitoring of the environment, buildings and human activity. Built from small devices about the size of a matchbox, called sensor nodes, that compute, relay data to each other and sense phenomena. This allows a user to monitor objects of interests for long periods of time and in greater detail. Developing energy efficient protocols and algorithms for WSNs has always been an important research question, but it is essential to improve the WSNs connectivity to other networks for increased usability and WSN system value. By interconnecting WSN with a global network, like the Internet, GSM, accessing WSN data can be transmitted from distance – another building, country or even continent. Thus bringing closer to realization of Internet of Things (IoT) [8].

To enable network interconnection the transition from one network to the other is required. One of possible options in WSNs interconnection is by using gateway (GW) – which basically translates one network data stream to other and vice versa and additionally can perform data aggregation. In our study-case it is between TCP/IP and WSN. GW usage relieves work load from sensor nodes and enables WSN to use specialized protocols, as well WSNs are more scalable and elastic to changes [1]. To our knowledge very few different gateway approaches have been proposed [2,5,7]. They include a) designed application layer GW for interconnecting WSN and TCP/IP networks using PXA270 board [2], b) an ARM based GW for interconnecting WSN and TCP/IP [5] and c) Samsung S3C2440 400MHz CPU [7]; all of which have been based on high-end devices. No justification or guidelines have been provided for the choice of hardware for GW.

The lack of a justification may be due to different application requirements or researchers assumption that GW has unlimited energy resources. But there are applications, mainly outside of cities, where gateways can't be plugged in and batteries must be used for GWs as well. One such application we can mention is equipping tractor with sensors to achieve autonomous work execution and provisioning users with most relevant and fresh data. Thus we are interested in – how the hardware choice for GW impacts performance and how it influences WSN data transmission when communicating with user, and network lifetime as well. Furthermore we wish to provide some easy to apply guidelines that could be used to make the right choice in GW hardware.

Most common approach for building WSN GW is to use ARM processor based boards (or other similar high-end device) running OS, like Linux; this was done in [2,5,7]. From our experience we know that this approach is more profitable – because use of OS can reduce application development expenses and complexity; furthermore, if constant energy source is available high-end board is a logical choice. But if we consider WSN system to be energy constrained, that includes GW as well, choosing or constructing appropriate board must be done carefully. As far as we know we are first to compare hardware that is used in building energy constrained GW for WSN. This may be due to fact that researchers want to check their developed approach overlooking energy efficiency in GW.

As we mentioned common approach used for GWs, is based on high performance ARM (32-bit) processors. But there exists a high-performance low-energy consumption 8-bit microcontroller based boards that could be used for GWs as well. The comparison of two selected board's performance under load conditions, from each end, is main attraction of this research and conclusions are based on obtained results. By comparing both boards we hope to obtain results that would clarify choice of hardware and if it can significantly affect system operation altogether and relive choice among available boards leading to more suited GW for.

The rest of the paper is organized as follows. In II we describe related work. In III we describe constructed test beds and software for GW testing and methods used for evaluation. In IV we show and analyze obtained results. In V we have implemented similar test beds for more detailed comparison. And finally in VI we conclude our research.

II. Related Work

GWs provide simple and easy to implement interconnection among two different networks by using intermediate node between these two networks. In [2] is designed application layer GW for interconnecting WSN and TCP/IP networks using PXA270 board. In [4] is proposed network interconnection address mapping mechanism, simple structure of GW node and mechanism for multiple GW selection. Paper focuses on theoretical network interconnection and GW structure. In [5] is proposed yet another ARM based GW for interconnecting WSN and TCP/IP. In [6] are discussed GW mobility aspects when interconnected with GWs. And in [7] Samsung S3C2440 400MHz CPU is proposed for WSN GW.

Alternatives include using more complicated, but more transparent method – TCP/IP protocol over sensor nodes as presented in [1,4]. In [1] is presented TCP/IP on sensor nodes a.k.a., overlay concept, for WSNs. In [3] SunSPOTS are described that are powerful sensor nodes (based on ARM processor, running at 200 MHz) and runs TCP/IP stack. This approach uses sensor nodes equipped with a TCP/IP protocol stack. By implementing TCP/IP stack on sensor nodes, border between WSN and TCP/IP network becomes blurry and often is referred as "Internet of Things" [8]. Drawbacks include increased energy consumption and communication overheads – because of TCP/IP packet size, due to noisy environment links between nodes are unreliable, but TCP/IP has been created as reliable protocol stack. Among advantages are routing and medium access – they are fully developed and should be easily adapted.

III. Description of test beds

For our research purposes two slightly different test beds were proposed. Common for both test beds are – sensor nodes, gateway (interconnected controller and sink node) and main computer, located in local network. Due to low energy consumption – ultra-low-power MCU, low-power radio module and protocol stack – SimpliciTI – eZ430-RF2500 boards were chosen as sensor nodes [9]. Connection between sensor nodes and gateway are wireless, but connection between gateway and main computer can be both – wireless or wired, as long as it ensures TCP/IP stack. Due to implementation simplicity wired – TCP/IP – connection setting was used.

To enable gateway connection with sensor nodes, controller board must be connected with one sensor node, called sink node, trough available hardware communication interface – this is widely used approach in GW interconnection. For us two possibilities exist – SPI (Serial Peripheral Interface) and USART (Universal Synchronous Asynchronous Receiver Transmitter). SPI is more suited for applications that require faster data delivery, while USART is more common communication interface and will be frequently available, but with more communication overhead. SPI speed is directly affected by boards source clock and generally can be derived as – maximum transfer speed is half of used clock speed (this is true for selected boards, but not for all microcontrollers/processors in general). Due to fact that sensor node maximum clock frequency is 16 MHz, maximum SPI transfer frequency is chosen as 8 MHz. But using higher sensor node clock frequency will lead to faster energy source depletion – developer must choose from these tradeoffs.

Based on common test bed assumptions DiGi Wi-9C board test bed is depicted in Fig. 1. and Xplained board test bed is depicted in Fig. 2. The difference can be seen in Fig. 2. – new board had to be included – ENC28J60-H – because Xplained board doesn't have built-in Ethernet support.

A. Design challenges

After choosing all hardware to test implementation, a few design challenges had to be overcome, as described below.

First was connecting sink node to gateway controller. No free SPI interface was available, because in Texas Instruments board drivers interface was used for other purposes. One possibility is to implement software SPI to connect both boards. Drawback of this is that transfer of data is slower than it is possible with hardware SPI. Second was to introduce SPI imitation mode, where incoming data were modulated from board that imitates sensor node functionality with maximum transfer rates.

The second challenge was interconnection of XMEGA XPLAINED-A1 and Ethernet controller ENC28J60-H. No TCP/IP stack was implemented on ATXmega's, but fortunately for ATmegas there were examples - uIP. So we had to port existing TCP/IP example to work with ATXmega. Although this wasn't done perfectly basic communication between PC and XMEGA XPLAINED-A1 was possible and worked without bugs.

Third was related to SPI driver on DiGI board. Available SPI driver for Linux operated only in master mode so both GW controllers were implemented in master mode and sink node as slaves. This means that sink node has to buffer data that are to be forwarded to GW controller so that they are always ready for sending after master node request.

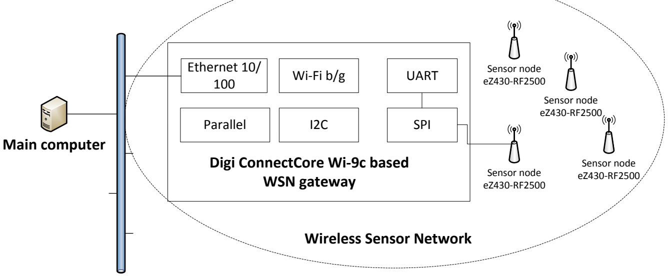


Fig. 1. ConnectCore Wi-9C test bed.

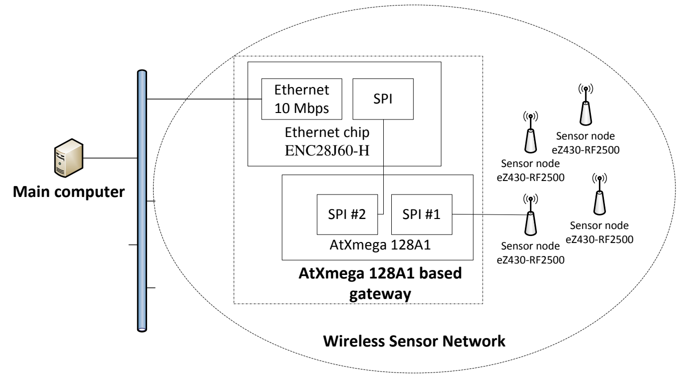


Fig. 2. ATxmega128A1 based gateway integration into WSN an interconnection with main computer.

B. Gateway operation algorithm for testing purposes

Based on proposed test beds, GW operation algorithm was developed that could ensure valid and reliable results. This algorithm is depicted in Fig. 3.

Algorithm executes after following steps:

1) Setting up board – variables, hardware communication interfaces and Ethernet interface.
2) Testing cycle can begin. It is repeated 100 times to ensure large sample count.
   a) Each test cycle consists of reading data from sink node using SPI or USART (depend on executed test). From each sink node 20 bytes are read. Assuming that each sensor node has sent value of 12 bits long (sensor nodes Analog to Digital converter resolution), 10 sensor nodes have sent measured value to sink node.
   b) After reading sensor node data it is inserted in predefined UDP (User Datagram Protocol) packets payload;
   c) And data are sent to user via Ethernet connection.

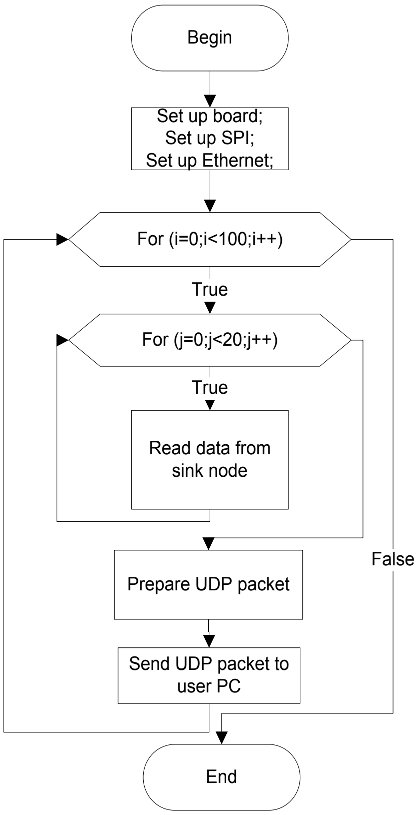


Fig. 3. Gateway test program’s flowchart.

Developed algorithm had to be implemented on both boards. Linux API (Application Programming Interface) functions and C/C++ were used due for DigGi Wi-9C while pure C was used for ATXMeaga board. We came to conclusions that using API makes programming less complicated and application development faster.

C. Tools used for data delivery speed measurement

Network protocol analyzer - Wireshark - was used as a tool to measure data transfer speed on user side. By using this application we can see what data we have received trough Ethernet interface and at what speed. It makes data checking and result processing easier.

But using Wireshark can only show total data transfer speed on user PC. If data transfer speed between sink node and gateway controller must be measured reliably, logic analyzer, that is connected to communication wires and captures transferred data, must be used. Here, data transfer time and amount can be measured, using logic analyzers software, to calculate data transfer rate. We used Intronix LA1034 Logicport logic analyzer.

IV. Results and evaluation

In this section we will discuss main results obtained after testing GW in proposed test beds. In all following figures transfer speed is depicted in **kBps** and time in **ms**.

A. Full communication link speed test

This test evaluates full communication path – from sink node to user PC, and speed performance of both selected gateway boards. Obtained results: Figure 4 a) shows, that by using ATXmega128A1 based GW, data transfer is almost eight times faster than DiGi board in 8MHz imitation mode and more than two times faster in software SPI mode. Even ATXmega128A1 software SPI mode is working faster than DiGi hardware SPI, which implies that ATXmega board can perform data forwarding with less overhead. Furthermore, it was measured that ATXmega128A1 gateway uses around 300 mA while DiGi gateway uses around 800 mA while operating as gateway.

Obtained results were different than what we had anticipated. We had expected that with DiGi board data transfer from sink node to user PC would be faster, but as we can see in Figure 4 a) this assumption was incorrect. To make sure that results are correct, test was performed for four more times and average values were presented. During these tests we noticed that DiGi board values deviates more than ATXmega. But in discussion about deviation we go into more details later. We can see that by choosing simpler hardware board data can be transferred even faster and with less energy consumed.

There are few drawbacks in our tests. First, we are aware that to receive more general results, more than two boards should be compared, but due to resource limitation it is not possible. When paper was prepared Atmel had released even more energy efficient and faster SAM4L boards, which could perform even better in these tests. Second, Linux kernel that was used wasn’t fully real time kernel, unfortunately present Linux kernel did not support real time patch. Thus further researches should be done to confirm our results. Third, sensor nodes with free hardware SPI should be used to avoid imitating this connection or implementing software SPI.

As we mentioned earlier common approach is to choose powerful ARM based boards, but when implementing real application gateways these results should be taken into account, because this could ensure longer network operation and thus smaller network maintenance costs.

B. Sink node to gateway connection

This connection was tested by sending large data amount (4000 bytes) from sink node to gateway controller and average transfer speed calculated and results were depicted in Figure 4 b). DiGi board in hardware/imitation SPI mode present small connection speed improvements compared to software SPI mode – only about two times. While ATXmeag128A1 gateway in software SPI mode already was three times faster than DiGi board and in hardware/imitation SPI mode more than twenty five times faster than software SPI mode.

This might be a little unexpected, but when we observed logic analyzers measured data, the gap in average transfer speeds was clear. ATXmega board had no delays between transmitted sink node bytes while DiGi board had random (at least we didn't notice any regularity) delays between byte transmissions. We want to remind that gateway controller acts as master due to DiGi limitations – it does not support SPI master mode, while ATXmeg board does not have this limitation.

C. Gateway to user PC connection

Connection was tested by sending large amount of UDP packets (100 packets) from gateway controller to user PC and network analyzer Wireshark was used for measuring transfer speed and results are depicted in Figure 4 c). From results it is seen that ATXmega board can transfer data two times faster than DiGi board. When performing this test we noticed that DiGi gateway has great variation in transfer speed, even as much as two times which is depicted in Figure 5 c). From both previous tests we can see that DiGi has more deviations which reduce total delivery speed.

D. Communication speed tests for gateway

These results present same tendency what was seen in full communication cycle, i.e. ATXmega based gateway transfer data faster than DiGi based gateway. This in the end leads to faster total speed.

Due to differences in presented results, we came to conclusion that total transfer time is formed out of three different components, of witch second is indirectly observable:

a) Time to transfer data from sink node to gateway – data transfer using SPI interface;
b) Time to switch between communication streams in gateway – copying data from SPI buffer to Ethernet buffer;
c) Time to transfer data from gateway to user PC – data transfer using Ethernet UDP packets;

Thus introducing additional time that can influence total transfer speed.

E. Switching between communication streams

Based on obtained connection speeds, switching between communication streams was calculated and depicted in Figure 4 d). It was performed equally for all tests, which shows that DiGi board is yet again deviating while aggregating data.

With this our test was concluded and interesting and new results were obtained. Due to limited resources we have, we invite other researchers to perform their own tests and compare results.

Intuitively we believed that DiGi board would perform better due to fact that it runs faster and has greater resources, but as results indicate we were mistaken – low-end board with less resources outperformed high-end ARM board when operating as gateway performing data delivery from WSN to user PC in local network. Furthermore achieves this with less energy consumed. This is particularly important when gateway must work by using battery and possibly if application requires gateway to be mobile. Next step could be examination of mobility impact on gateway and WSN collaboration. Of course there are few factors that should be taken into account. First, delivery speed decreases if more data control – filtering, aggregation must be performed, because data handling time increases. Second, if distance from WSN to user increases, delivery speed most likely will decrease, because route increases and more network devices must be employed to send data.

V. Alternate gateway connection and gateway speed deviation tests

To make results more detailed and comparable to other settings we continued our test with:

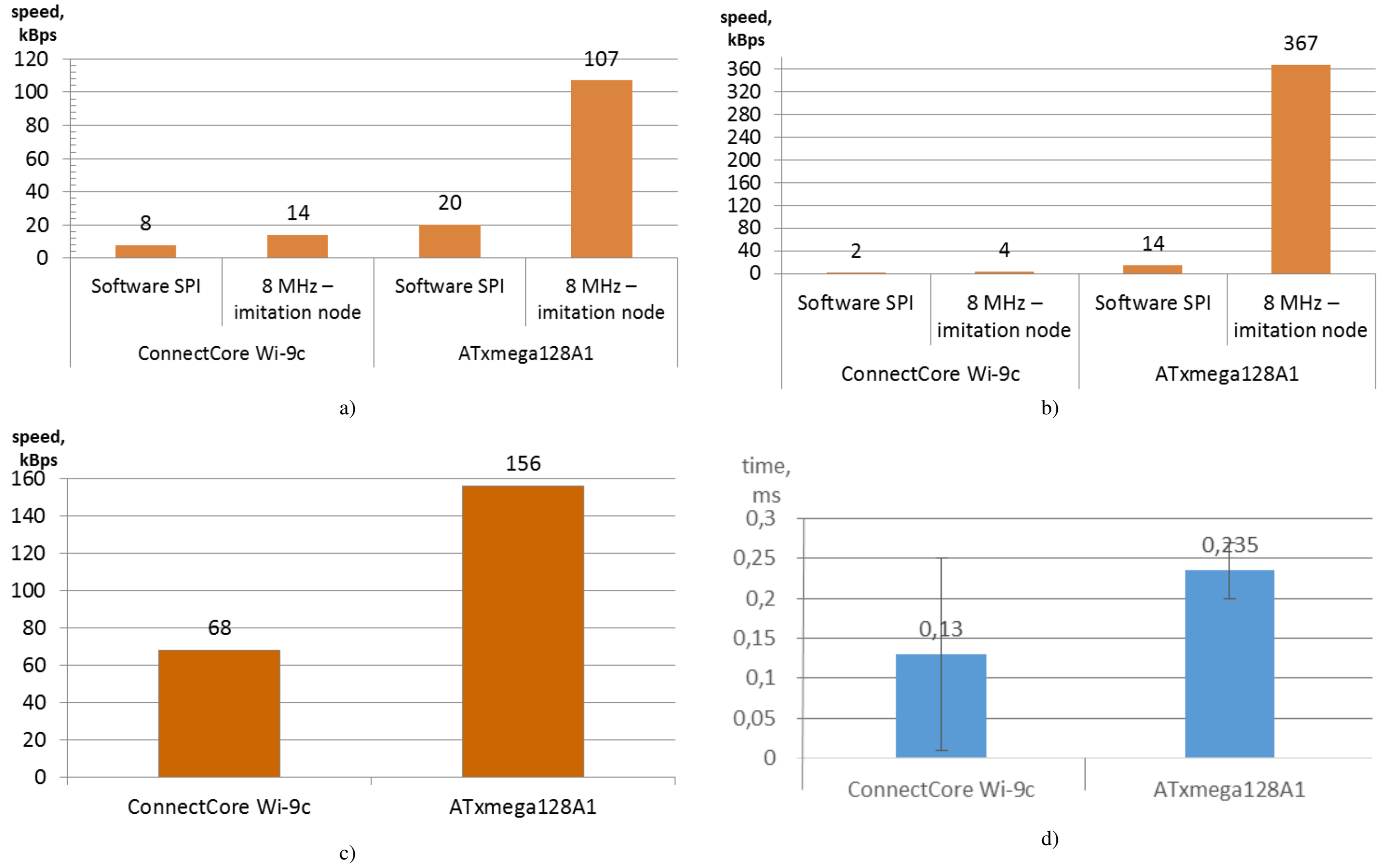
a) Transfer speed variation calculations for both connections;
b) Replacing SPI with UART, which is more common interface;

While transfer speed from sink node to user PC was measured, speed variations were calculated from min and max speed values from these observed results and depicted in Figure 5 a), b) and c). As can be seen in these three figures, ATXmega128A1 gateway variants very little and is even constant in some connections, but DiGi gateway variations highly greater than ATXmega GW. We think this is due to OS which needs to allocate/reallocate different resources which can introduce different and somewhat random delays in performance. This implies that DiGi can't be used for reliable (in sense of guaranteed data delivery time) communication because transfer speed can wary and guaranteed can be only lowest transfer speed. Thus when higher guaranteed data transfer rates should be guaranteed no OS should be used.

Since many sensor nodes use UART interface performance using this interface was measured as well. One drawback of UART in comparison with SPI is that UART frame has only 80% of useful data (8N1) in its frame while SPI has 100% - meaning only data is transferred. The same as in first test full communication link transfer speed between sink node and user PC were performed. Results are depicted in Figure 5 d). If compared with results from Figure 4 b) it is possible to see, that even ATXmega software SPI mode performs faster than UART. Although UART is more freely available, if transfer speed is important criterion it is advised not to use it.

A few lessons that were learnt during implementing presented GW were obtained. First, programming DiGi board was much faster and easier, because more examples are available for Linux, both for Ethernet and SPI. While ATXmega programming took a lot longer and debugging was more complicated. Second, even theoretically less powerful device can outperform high end device if certain conditions are met. Third, choosing gateway components can be difficult, but in the end it can present better performance as was seen in our case.

Our future work includes building a mobile gateway that could be used in wireless sensor network to relay local data to user and this research makes a contribution to the kind of board



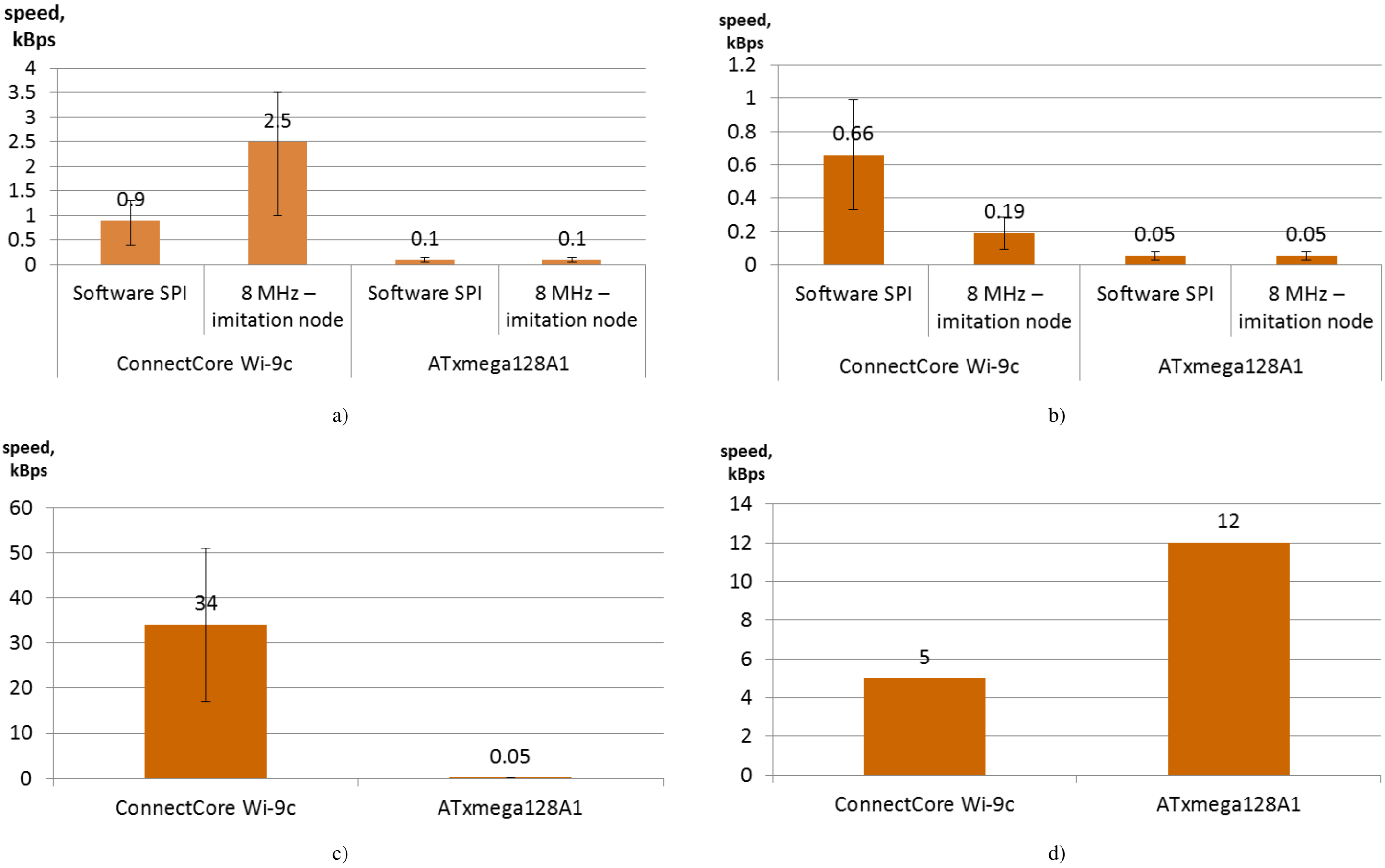
a)

b)

c)

d)

Fig. 4 Average transfer speed: a) from sink node to user PC; b) from gateway to sink node; c) from gateway to user PC; and d) average switching time between communication streams in gateways



a)

b)

c)

d)

Fig. 5 Speed variation in a) sink node to user PC communication; b) sink node to gateway communication; c) gateway to user PC communication; and d) transfer rate from sink node to user PC using UART at 115200 baud rate

we should be using when implementing mobile gateway, i.e. we will consider using 8 bit microcontroller based gateway, since a lower energy consumption will provide longer operation. Most important resource for this type gateway is energy, although by using mobility gateway could be recharged more easily. Amongst new challenges could be adding wireless capabilities to this gateway.

VI. Conclusions

Our paper presents comparison of two different gateways envisioned for use in wireless sensor networks. Main concern that is investigated is choosing suitable hardware for gateway. Common approach, as we have seen in previous researches is to choose ARM based boards that run Linux OS. We suggested using less complex and less powerful board thus reducing energy consumption and furthermore data transfer rate shouldn't decrease significantly. Two available boards were chosen – ATXmega Xplained-A1 and DiGi Wi-9C board – and compared in similar test beds performing identical task – forwarding data from wireless sensor network sink node to user PC.

As obtained results imply using less powerful board without OS, can ensure smaller energy consumption and even increase data delivery speed, thus being more suited for wireless sensor network applications where data delivery speed is important. Furthermore, using the same board results are more stable over time, i.e. delivery speed is the same today as was yesterday. Further advantages of using less complex hardware is that overall costs for wireless sensor network can be reduced. Among disadvantages – programming becomes more complex when no API (which is in OS case) is used; possible that gateway must be designed by developer, because not always all necessary hardware is included in one board.

One more conclusion is from observing variations in results. Especially among sink node and GW when SPI is used. We observed that random delays between transferred data bytes was present. To our believes this is one of main reason why in the end ATXmega board outperformed DiGi board. Greatest drawback of OS based solutions is that to operate OS some user invisible processes are performed and some delays are introduced leading to uncertainty which is undesired in timely applications.

Lastly we want to mention some tradeoffs we encountered and observations we saw during developing our first gateways and give other developers some pointers what type of board would be more suited for certain applications:

1) If implementation should be done in short time preferable are boards with OS, like Linux. Because using API noticeably decreases development time. Furthermore using API provides wider application possibilities faster.
2) If gateway should ensure less energy consumption or little as possible speed deviation, or more control over hardware then board with no OS should be used (as seen in paper even less powerful board can be feasible).
3) If wireless sensor network costs are important then less powerful board can be used.

Acknowledgment

This work has been partly supported by the European Social Fund within the project «Support for the implementation of doctoral studies at Riga Technical University

References

[1] Lei, Shu and Jin, Wang and Hui, Xu and Cho, Jinsung and Lee, Sungyoung. Connecting Sensor Networks with TCP/IP Network. Springer-Verlag. p. 330--334 2006 http://dx.doi.org/10.1007/11610496_44

[2] Song, Ping and Chen, Chang and Li, Kejie and Sui, Li. The Design and Realization of Embedded Gateway Based on WSN. IEEE Computer Society. p. 32--36 2008 http://dx.doi.org/10.1109/CSSE.2008.889

[3] Guinard, Dominique and Trifa, Vlad and Pham, Thomas and Liechti, Olivier. Towards Physical Mashups in the Web of Things. IEEE Press. p. 196--199 2009 http://dl.acm.org/citation.cfm?id=1802340.1802386

[4] Han, Yanyan and Li, Deshi and Chen, Jian and Wang, Tianyu. Research on Wireless Sensor Network and Carrying Network Integration Based on Gateway. IEEE Computer Society. p. 748--751 2011 http://dx.doi.org/10.1109/iThings/CPSCom.2011.18

[5] Ye, Dun-fan and Min, Liang-liang and Wang, Wei. Design and Implementation of Wireless Sensor Network Gateway Based on Environmental Monitoring. IEEE Computer Society. p. 289--292 2009 http://dx.doi.org/10.1109/ESIAT.2009.194

[6] Shakya, Mukesh and Zhang, Jianhua and Zhang, Ping and Lampe, Mattias. Design and Optimization of Wireless Sensor Network with Mobile Gateway. IEEE Computer Society. p. 415--420 2007 http://dx.doi.org/10.1109/AINAW.2007.146

[7] Zhu, Qian and Wang, Ruicong and Chen, Qi and Liu, Yan and Qin, Weijun. IOT Gateway: BridgingWireless Sensor Networks into Internet of Things. IEEE Computer Society. p. 347--352 2010 http://dx.doi.org/10.1109/EUC.2010.58

[8] Atzori, Luigi and Iera, Antonio and Morabito, Giacomo. The Internet of Things: A Survey. Elsevier North-Holland, Inc.. p. 2787--2805 2010 http://dx.doi.org/10.1016/j.comnet.2010.05.010

[9] eZ430-RF2500 Development Tool: User's Guide, Texas Instruments Inc., Dallas, TX, 2009

Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems.

Events that constitute IT4BMS are:

- **ABICT'14** - 5th International Workshop on Advances in Business ICT
- **AITM'14** - 12th Conference on Advanced Information Technologies for Management
- **IT4L'14** - 3rd Workshop on Information Technologies for Logistics
- **KAM'14** - 20th Conference on Knowledge Acquisition and Management
- **SS4SI'14** - 1st International Symposium on Service Systems for Social Innovation

5th International Workshop on Advances in Business ICT

ABICT focuses on Advances in Business ICT approached from a multidisciplinary perspective. It will provide an international forum for scientists/experts from academia and industry to discuss and exchange current results, applications, new ideas of ongoing research and experience on all aspects of Business Intelligence. ABICT will be also an opportunity to demonstrate different ideas and tools for developing and supporting organizational creativity, as well as advances in decision support systems.

We kindly invite contributions originating from any area of computer science, information technology and computational solutions for different applications areas, data integration and organizational implementation of ABICT, as well as practical ABICT solutions.

Topics

Topics include (but are not limited to):

- Advanced technologies of data processing, content processing and information indexing
- Analytics as a service
- Big Data: benefits and challenges
- Business Analytics
- Business applications of social networks
- Business data mining and knowledge discovery
- Business Intelligence
- Business Rules
- Business-oriented time series data mining, analysis, and processing
- Cloud based Business Intelligence
- Creativity Support Tools
- Customer Relationship Management, social Customer Relationship Management
- Data driven marketing
- Data Warehousing
- Decision support
- Information forensics and security, information management, risk assessment and analysis
- ICT technologies in enterprise management
- Knowledge Management (for better Decision Support, Collaboration and Competitiveness)
- Legal text processing
- Semantic Web and Ontologies in Business ICT
- Virtual Enterprise
- Web 2.0 and Web 3.0 in fusing Business Intelligence systems and Decision Support Systems
- Web-Based Data Management Systems

Event Chairs

Mach-Król, Maria, Katowice University of Economics, Poland

Olszak, Celina M., University of Economics in Katowice, Poland

Pełech-Pilichowski, Tomasz, AGH University of Science and Technology, Poland

Program Committee

Abramowicz, Witold, Poznań University of Economics

Badica, Amelia, University of Craiova, Romania

Berio, Giuseppe, Universite de Bretagne Sud, France

Chiu, Dickson K. W., Dickson Computer Systems, Hong Kong S.A.R., China

Christozov, Dimitar, American University in Bulgaria, Bulgaria

Gaweł, Bartłomiej, AGH University of Science and Technology

Kacprzyk, Janusz, Institute of Computer Science, Polish Academy of Sciences, Poland

Khachidze, Manana, Tbilisi State University, Georgia

Konikowska, Beata, Institute of Computer Science, Poland

Koohang, Alex, Macon State Collage, United States

Korwin-Pawlowski, Michael L., Universite du Quebec en Outaouais, Canada

Kulczycki, Piotr, Systems Research Institute, Polish Academy of Sciences, Poland

Ligęza, Antoni, AGH University of Science and Technology, Poland

Loucopoulos, Peri, Harokopio University of Athens, Greece

Maamar, Zakaria, Zayed University, United Arab Emirates

Michalik, Krzysztof

Nycz, Malgorzata, Wroclaw University of Economics, Poland

Ogihara, Mitsunori, University of Miami, United States

Owoc, Mieczyslaw, Wroclaw University of Economics, Poland

Petryshyn, Lubomyr, AGH University of Science and Technology, Poland

Prasad, T. V., Visvodaya Technical Academy, India

Pulvermueller, Elke, University Osnabrueck, Germany

Reimer, Ulrich, University of Applied Sciences St. Gallen, Switzerland

Rossi, Gustavo, National University of La Plata, Argentina

Salem, Abdel-Badeeh M., Ain Shams University, Egypt

Sauer, Jurgen, University of Oldenburg, Germany

Szpyrka, Marcin, AGH University of Science and Technology, Poland

Teufel, Stephanie, University of Fribourg, Switzerland

Whatley, Janice, University of Salford, United Kingdom

Wrycza, Stanisław, University of Gdansk, Poland

Zadrozny, Slawomir, Systems Research Institute, Poland

Zurada, Jozef, University of Louisville, United States

Zurada, Jozef, College of Business University of Louisville, Louisville

Position papers of the 2014 Federated Conference on Computer Science and Information Systems pp. 201–208

DOI: 10.15439/2014F275
ACSIS, Vol. 3

Enhancement of the ValueSec Risk Management Model

Andrzej Bialas
Institute of Innovative Technologies EMAG,
ul. Leopolda 31, 40-189 Katowice, Poland
Email: a.bialas@emag.pl

Abstract—**The paper concerns the ValueSec methodology and tool which support decisions related to the security measures selection in different application contexts. The ValueSec project, financed by the European Commission Seventh Framework Programme (FP7), considers security measures which properly affect risk, are cost effective, bring benefits and are free of different restrictions (political, social, legal, psychological, etc.). These restrictions, called here qualitative factors (criteria), are hard to identify and assess. The ValueSec methodology is based on three pillars: risk assessment, cost-benefits assessment and qualitative criteria assessment. The paper discusses the project results by identifying their positive and negative features and proposing to enhance the ValueSec methodology. The focus is on one of the possible enhancements, i.e. monitoring factors which influence the measure effectiveness during its operation. The proposed concept shows how the shortage of resources needed for the measure implementation and operation impacts the measure efficiency during the operation.**

I. INTRODUCTION

THE paper presents how to enhance the risk management framework elaborated in the ValueSec project, financed by the European Commission Seventh Framework Programme (FP7). The project was performed by 11 partners from Germany, Finland, Norway, Spain, Poland, and Israel, including the Institute of Innovative Technologies EMAG [1]. The results of ValueSec, i.e. the methodology and tools (ValueSec toolset) are dedicated to the security decision makers, policy makers, architects and other stakeholders to support them in strategic decisions concerning the selection of security measures in a certain context of application. Decisions about security measures selections are very complex because each decision requires the trade-off between many factors of diversified nature. Additionally, some of these factors are multi-directional and often opposite to each other.

It was assumed in ValueSec that the selected measures should:

- properly affect the risk,
- be cost-effective,
- take into account non-financial restrictions.

Basically, the main focus area of ValueSec is security. On the other hand, however, the interdisciplinary character of the project lies in economical, political, social, legal, psychological, and other issues (called qualitative factors) which are taken into account here. Their consideration in ValueSec is the basic added value of the project.

The diversified, multidirectional, positive and negative effects form a vector of values related to the security measure. The optimization of this function, from different points of view and decision contexts, is the main objective of the ValueSec project, expressed by its full title "ValueSec – Mastering the Value Function of Security Measures".

Other project aims are:

- to reduce the uncertainty related to the decision context,
- to reduce the fuzziness of the decision process,
- to provide better decisions argumentation for stakeholders, who have diverging priorities, and for citizens, who are usually unable to recognize whether the decisions reflect their interests.

The ValueSec methodology was validated in five application domains [2], called contexts (by running certain scenarios and applying security measures to them, called here use cases):

- public mass event – for the scenario "Valencia's Formula One Race Track" the following are assessed: CCTV, scanners and frequency inhibitors; they are called use cases and are focused on the improved surveillance and detection systems;
- public mass transportation – for the scenario "rolling stock depot security" the following are assessed: the use of a train portal and different access control and face recognition sensors;
- air transportation/airport security – for the scenario "Norwegian airports security" the following are assessed: the implementation of security measures for electronic screening of liquids, aerosols and gels (LAG's) [3],
- communal security planning – for the scenario "Flood protection based on the experience of the German Bundesland Saxony-Anhalt (LSA) during the 2002 and 2013 floods of the Elbe and Mulde

rivers” the following are assessed: the implementation of crisis management software, establishing a standardized secure communication network, and standardization of command & control equipment and management tools & software [4],

- cyber threat – for the scenario “Cyber-security smart grid attack based on the targeted viruses, like Stuxnet” the following are assessed: different security measures applied to different areas/layers like IT infrastructures, IT systems, physical security and procedures.

The paper reviews the ValueSec researches, from ideas to the tool prototype (Section 2). Section 3 discusses how to improve the ValueSec framework, Section 4 and 5 compare the current and the enhanced processes of the security measures selection. Section 5 discusses the implementation of the proposed solutions in the ValueSec toolset. The last section concludes the work and presents some plans for the future in this field.

II. ValueSec Methodology and Toolset

The ValueSec methodology does not define a complete risk management framework [5], but its key parts focused on the multidimensional assessment of the security measure before the decision related to its implementation in the considered context is made [1]. The ValueSec methodology, which supports decision makers, can be applied when the decision should be taken with respect to the secured undertaking, event, object or project. This methodology is not used to manage (to monitor, to maintain) the security. It is used rather for one time ventures than for permanent activities.

The general scheme of the ValueSec decision making framework is shown in Fig. 1. In the considered context and scenario the decision maker prepares a set of security measures to assess in this application. Next he/she analyses protected assets or processes, identifies available resources, budget and social values. Each measure is assessed with respect to the risk affected, cost-benefits brought and non-financial restrictions which affect the measure during the operation.

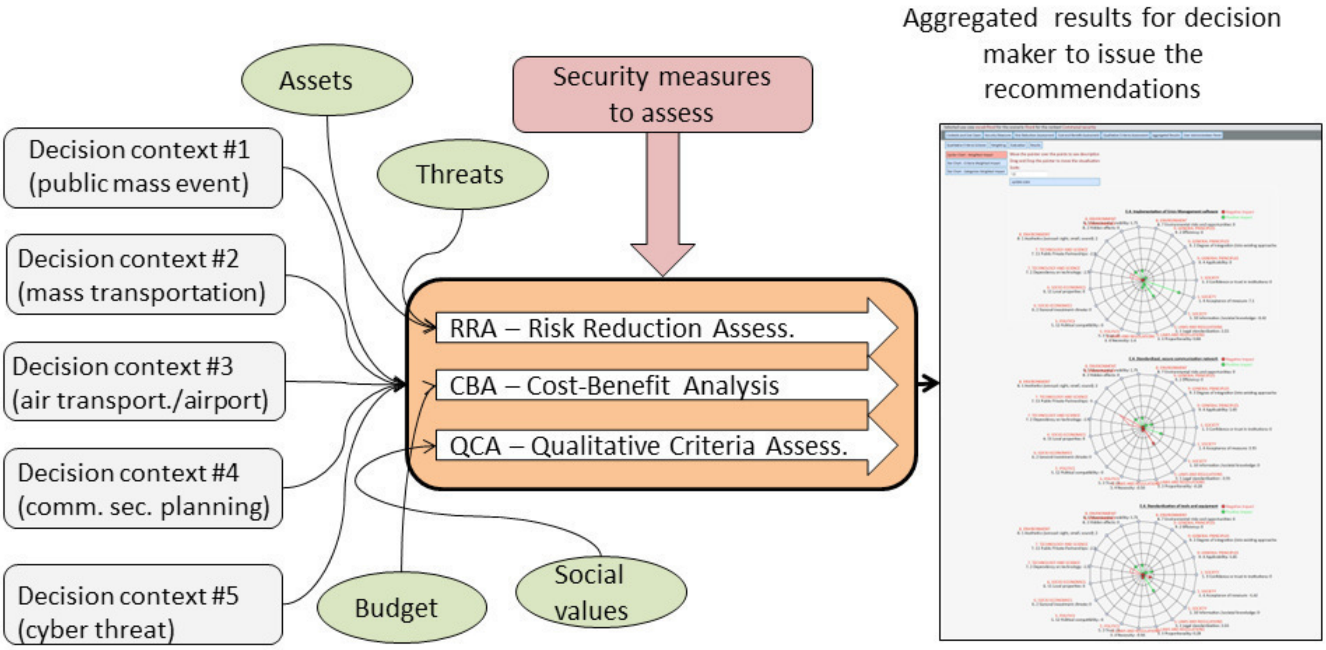

Aggregated results for decision maker to issue the recommendations
Assets
Security measures to assess
Threats
Decision context #1 (public mass event)
Decision context #2 (mass transportation)
Decision context #3 (air transport./airport)
Decision context #4 (comm. sec. planning)
Decision context #5 (cyber threat)
RRA – Risk Reduction Assess.
CBA – Cost-Benefit Analysis
QCA – Qualitative Criteria Assess.
Budget
Social values

Fig. 1 General scheme of the ValueSec decision framework

The results of the security measure assessment facilitate decision making within different threat and risks, financial, political and social aspects.

As a result, different information related to the assessed measures is obtained. This information needs to be analyzed and synthetized to obtain the aggregated results useful for decision makers to elaborate the final recommendation.

The ValueSec framework is implemented as the ValueSec toolset with tree distinguished pillars:

- Risk Reduction Assessment (RRA) pillar [6],
- Cost-Benefit Analysis (CBA) pillar [7],
- Qualitative Criteria Assessment (QCA) pillar [1].

The RRA pillar is based on four RRA components elaborated by the consortium members and assigned for specific contexts:

- Riger (elaborated by the consortium member ATOS) assigned for the public mass event context; it is an asset-oriented risk analyzer;
- RAS (elaborated by the consortium member TUM) assigned for the public mass transportation and air transportation/airport security contexts; it is a process-oriented risk analyzer and a simulation tool;

- OSCAD (elaborated by the consortium member EMAG) dedicated for communal security planning; it is an asset/process-oriented risk analyzer;
- Lancelot (elaborated by the consortium member WCK) used for cyber threat; it is an asset/process-oriented risk analyzer.

During the framework operations, a given RRA component is used twice:

- to assess existing (inherent) risk,
- to assess the risk after the considered measure implementation.

From all preselected variants of security measures adequately affecting risk, those should be selected, which are cost-benefit effective and are free from non-financial restrictions (with the use of the CBA and the QCA pillar respectively).

For the monetary Cost-Benefit Analysis (CBA) three main categories are distinguished:

- investment costs,
- operating costs,
- future benefits.

Each of these main categories is configurable and has its subcategories and sub-subcategories. For example, the category of investment costs has the following subcategories: initial planning, initial procurement process cost, procurement, setup and integration, initial set of spare parts.

The category of operational cost encompasses the subcategories: personnel, basic supplies, customization and adaptation, logistics, quality control, safety and security external services, etc.

Benefit category includes the subcategories: reduction of casualties (saved lives, reduction of injured people), reduction of damages of property, infrastructure, critical infrastructure, and environment, reduction of operational costs or resources, reduction of infrastructure fees, growing business profits, image-related benefits, reduced probability/frequency of threats, etc.

The CBA tool allows to determine the different commonly used key indicators, like: Net Present Value (NPV), Present Value of Benefits/Costs (PVB/PVC), Benefit Cost Ratio (BCR), Internal Rate of Return (IRR), Break even, Pay Back Period years (PBR), etc.

The security measures, properly affecting risk and having acceptable cost-benefit characteristics, are passed for the QCA pillar. This pillar is responsible for the analysis of restrictions with the use of varied factors which are difficult to determine [1]. The following main categories of immaterial parameters of security-related decision making are considered:

- general principles,
- social parameters (social group level),
- individuals (personal level),
- legal regulations,
- social laws and ethics,
- politics,
- socio-economics,
- technology and science,
- living environment and natural environment.

Each category is configurable and has several subcategories. Some of them – relevant for the given analysis – are selected by the QCA tool user. The tool allows to eliminate the "overlapping" or "doublecounts" items, to identify the interdependencies between subcategories, etc. For each subcategory its positive and negative impact is quantitatively assessed with the use of the predefined utility functions.

The ValueSec toolset offers different kinds of diagrams and tabular data reports as the aggregated results of assessment for each of the considered security measures.

III. Range of the possible enhancements

The ValueSec methodology is based on three independent pillars, which can be iteratively used to elaborate the aggregated results dealing with the assessed security measures in the decision context. The RRA and CBA pillars are used by risk managers but the QCA pillar is the innovative added value of the ValueSec project.

The validation shows that the ValueSec methodology, supported by the toolset, can be useful in five previously mentioned contexts and has considerable potential of applications in other domains. The questions are: Does this framework have only positive features? Can it be improved or extended? How can this be done?

During the elaboration and validation of the ValueSec project results some ideas and concepts were identified.

1. The ValueSec methodology and its supporting toolset provide a lot of diversified information for the decision maker (aggregated results) and, in this sense, support the decision making process.

Please note that the decisions themselves are not supported by any specialized methodology but are elaborated heuristically by people. In this field there is potential to extend the ValueSec methodology by applying commonly used methods, e.g. MCDM/A (Multiple-criteria decision making/analysis). The ValueSec output can be adapted and used as input for the chosen methodology applied to automate the decision process. This is performed to facilitate the work of decision makers.

2. The ValueSec methodology and toolset are focused on the security planning and do not tackle the security measure implementation and use.

The selection of security measures which properly affect risk, are cost-benefits effective and free of restrictions related to the qualitative criteria – does not guarantee a full success. This is due to the fact that these measures can be later improperly implemented, monitored, and the resources for their management can be insufficient. There is a danger that all activities performed according to the ValueSec methodology may be thwarted later, during implementation and operation of the measure. For this reason it is proposed to conduct a security measures sensitivity analysis against the factors that may decrease the security measure efficiency during the future operation. Moreover, performance indicators tracking the effectiveness of the applied measures can be useful.

3. The ValueSec methodology analyses the risk before and after the security measure implementation, but CBA and QCA are performed only in the situation after the measure implementation.

Please note that the risk "before" is related to the existing, previously applied security measures, which also have costs, bring some benefits and have some qualitative restrictions. It would be better to analyze CBA and QCA parameters also before the security measure selection, to obtain a more detailed picture of the current situation. For this reason a differential approach is proposed. The gain related to the security measure selection will be defined more precisely, as a difference between the "before" and "after" situation. It is proposed to invoke the RRA, CBA and QCA components twice to analyze the current situation and the ex-post one.

Moreover, the RRA components should support explicit identification of the benefits related to the measures, which allow to elaborate more valuable input for the CBA analysis.

4. The ValueSec framework, based on rather simple risk model, has restricted possibilities to express more sophisticated relationships between different assets, threats and vulnerabilities, and in results to consider the cascading or escalating effects.

The analysis of this effects is important especially for the critical infrastructures. This limitation of the ValueSec methodology may disturb it dissemination in this domain of application. For this reason the more enhanced RRA components should be implemented and CBA and QCA properly enhanced.

Each of the four identified issues needs further researches to elaborate the useful enhancements of the ValueSec methodology.

In the next two sections one of these four issues will be shortly discussed, i.e. the issue No. 2, related monitoring the efficiency of the implemented security measures.

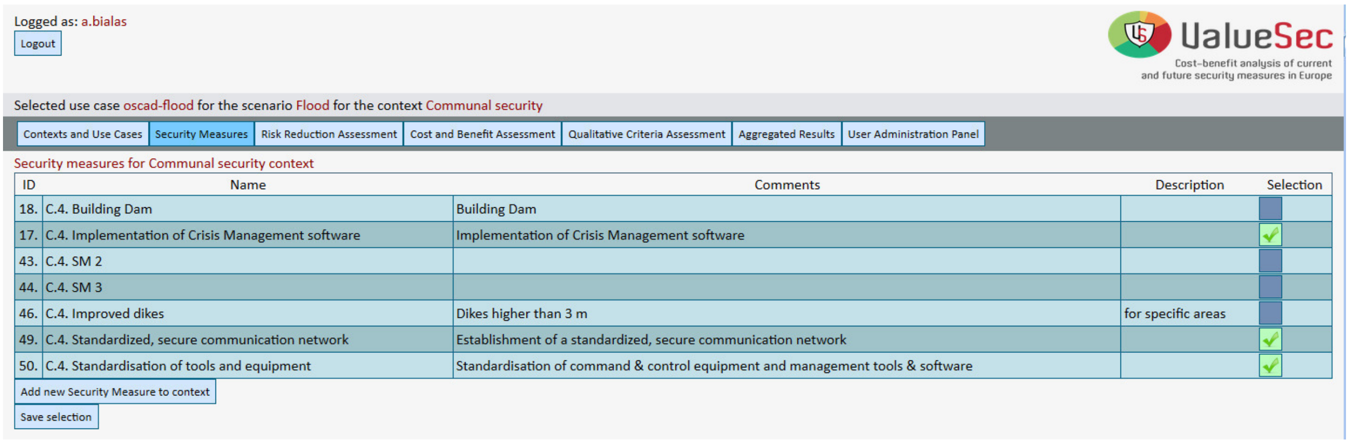
IV. The Current Security Measures Assessment Process

The current security measure assessment process ought to be shortly presented here.

At the beginning of the process the decision maker selects the context, e.g. "Communal security", the scenario, e.g. "Flood protection" and security measure to analyze, e.g.:

- "Implementation of crisis management software",
- "Establishment of a standardized secure communication network",
- "Standardization of command & control equipment and management tools & software".

Fig. 2 presents the general view of the ValueSec toolset and the three above selected items. Please note all main menu options, which are activated step by step, except "User Administration Panel". This menu option includes general managing functions of the tool.



Logged as: a.bialas

Logout

UalueSec
Cost-benefit analysis of current and future security measures in Europe

Selected use case oscad-flood for the scenario Flood for the context Communal security

Contexts and Use Cases
Security Measures
Risk Reduction Assessment
Cost and Benefit Assessment
Qualitative Criteria Assessment
Aggregated Results
User Administration Panel

Security measures for Communal security context

ID	Name	Comments	Description	Selection
18.	C.4. Building Dam	Building Dam		☐
17.	C.4. Implementation of Crisis Management software	Implementation of Crisis Management software		☑
43.	C.4. SM 2			☐
44.	C.4. SM 3			☐
46.	C.4. Improved dikes	Dikes higher than 3 m	for specific areas	☐
49.	C.4. Standardized, secure communication network	Establishment of a standardized, secure communication network		☑
50.	C.4. Standardisation of tools and equipment	Standardisation of command & control equipment and management tools & software		☑

Add new Security Measure to context

Save selection

Fig. 2 The ValueSec toolset main menu – selecting security measures for assessment

The ValueSec analyses are performed with use of the components of three pillars (RRA, CBA, QCA). As a result, the decision maker is provided with a huge number of analytical data of different shapes.

The first step of analyses is the assessment how each of the considered security measures affects risk. In the flood protection scenario, the OSCAD software elaborated by EMAG was used [4], [6] as the RRA component. The OSCAD tool allows to analyze risk with respect to processes (e.g. preparedness, reaction, restoration processes) or assets (e.g. people, infrastructure, natural environment). The examples of the considered issues are: loss of lives, injuries, damages of business and technical infrastructure, damages in agriculture and natural environment, etc. The risk assessment is performed twice:

- before the implementation of any measure (inherent risk, current situation),
- when the considered security measure is selected for implementation.

The risk values before and after measure selection are transferred to the main component of the ValueSec toolset, as the key data for the decision maker. This component calculates the risk reductions caused by any measure and shows them in percentage.

Next the CBA analysis is provided. In the beginning, different analysis parameters are defined (monetary value, time horizon, discount rate, used cost live cycle model, budget limit, investment costs-, future costs- and benefits subcategories). Then the distributions of cost-benefits categories/subcategories in time are produced and different analytical parameters, like NPV, PVB, PVC, BCR are calculated.

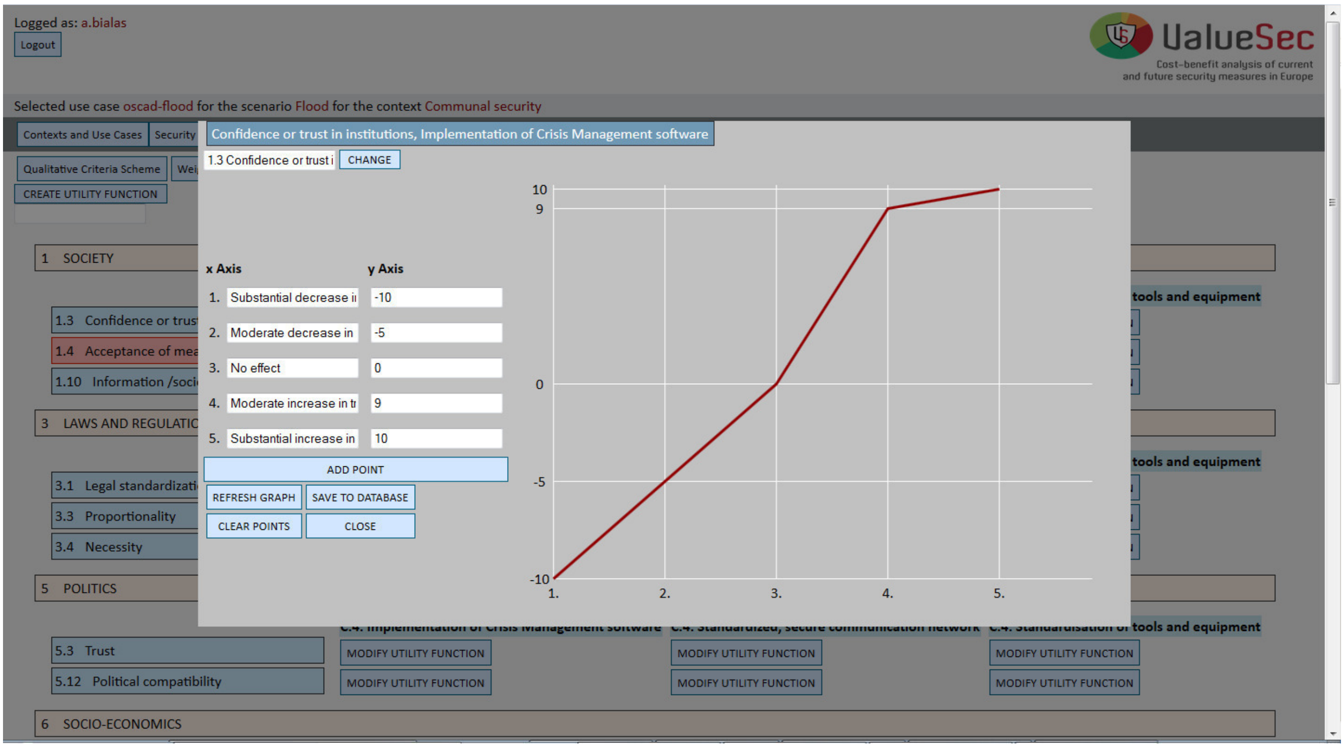
The next step is the QCA assessment of the proposed security measures. From the huge number of QCA categories/subcategories the decision maker selects these relevant for the context and scenario, eliminating cases called "overlappings" and "doublecounts", identifying dependencies between the selected items, defining weights and finally performing the evaluation.

For each QCA subcategory item the utility function can be defined, which expresses the item influence (linear or not) in a numerical way. The example of such function for the "Confidence or trust in institution" item, with respect to "Implementation of crisis management software", is shown in Fig. 3. For the five enumerative values placed on the X axis one can assign numbers of the range -10 to 10 in the Y axis.

The user can define the shape of this relationship.

Each pillar component produces its own data set which encompasses the detailed analysis results.

Moreover, the aggregated results in different kinds and shapes are provided to summarize the analysis.



Logged as: a.bialas

Selected use case oscad-flood for the scenario Flood for the context Communal security

Confidence or trust in institutions, Implementation of Crisis Management software

1.3 Confidence or trust i

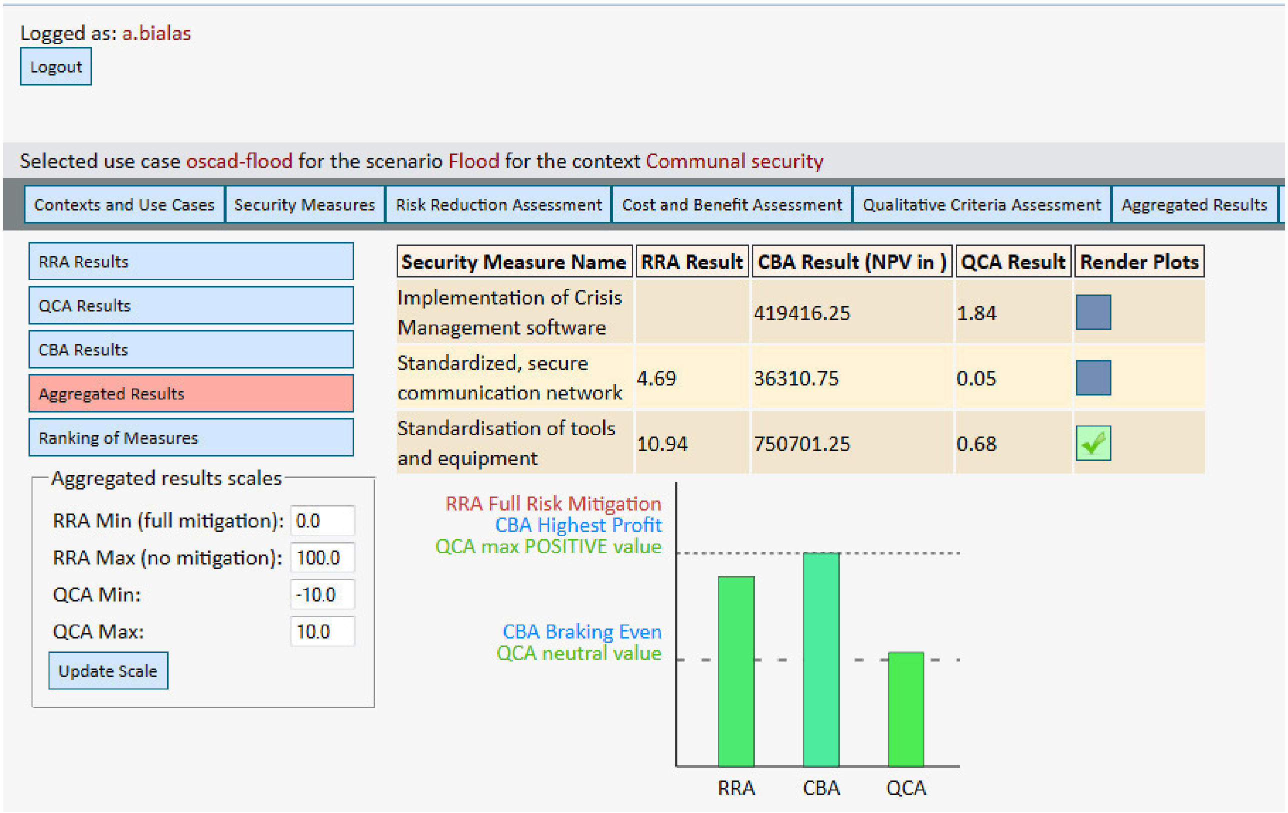
x Axis	y Axis
1. Substantial decrease i	-10
2. Moderate decrease in	-5
3. No effect	0
4. Moderate increase in tı	9
5. Substantial increase in	10

Fig. 3 Defining utility functions which transform analytical enumerative values into numbers expressing negative and positive impacts

Fig. 4 shows an example of data produced for three security measures considered in the scenario dealing with the flood protection. This an example of data called aggregated results.

Please note that the "Standardization of command & control equipment and management tools & software" security measure reduces risk by 10.94 %, has NPV: 750,701.25 Euro, and the middle QCA impact is 0.68. The detailed data interpretation is discussed in the project deliverables [1].

This short description of the ValueSec toolset shows that the decision maker obtains many diversified characteristics (tabular, diagrams) related to the selected measures. This allows to assess how the planned security measure should behave in the considered context.



Logged as: a.bialas

Selected use case oscad-flood for the scenario Flood for the context Communal security

Security Measure Name	RRA Result	CBA Result (NPV in)	QCA Result	Render Plots
Implementation of Crisis Management software		419416.25	1.84	🟦
Standardized, secure communication network	4.69	36310.75	0.05	🟦
Standardisation of tools and equipment	10.94	750701.25	0.68	✓

Aggregated results scales

RRA Min (full mitigation): 0.0

RRA Max (no mitigation): 100.0

QCA Min: -10.0

QCA Max: 10.0

Fig. 4 The ValueSec toolset main menu – selecting security measures for assessment

V. The Enhanced Security Measures Assessment Process

Please note that the presented here considerations end at the security measures selection, i.e. at the security planning. The above analyses do not provide any information or requirements related to the further implementation and maintenance of security measures. Will the properly selected measures be really effective? The stakeholders expect that security measures will not only be properly selected, but also effective when crisis situations occur – they simply expect certain assurance from the security system.

To ensure efficient monitoring of the implemented security measures (section III, issue No. 2), the following solutions can be implemented:

1. Extending the security measure specification by parameters related to the measure implementation and maintenance. Moreover, extending the ValueSec toolset by one menu option representing the security measure sensitivity analysis. The analysis is to show how the shortage of resources impacts the effectiveness of the security measure in the operational environment.
2. Implementing performance indicators which allow to check whether the security measures are effective when critical obstacles occur. The ValueSec toolset menu should be extended again.

To present the proposed solutions, a more precise specification is elaborated. Each security measure (SM) is represented by the `SecurityMeasure` class – a class of the ValueSec ontology. This ontology encompasses the project data and relationships. The `SecurityMeasure` class has many different parameters (ontology properties):

- `SMhasID` – unique identifier assigned to each measure;
- `SMhasName` – name of the measure, e.g. “Building dam” in Fig. 2 (please note: the context identifier “C.4” is not a part of this name but is concatenated to the security measure name);
- `SMhasDescription` – textual, informal description of the measure;
- `SMhasRiskBefore` – inherent risk value (risk before any considered security measure implementation);
- `SMhasRiskAfter` – assessed risk value when the considered security measure is applied;
- `SMhasInvCost` – points at the `InvestmentCost` class individual specifying investment costs (subcategories, their properties and parameters) in the cost-benefit analysis;
- `SMhasOperatCost` – points at the `OperatingCost` class individual specifying operating costs (subcategories, their properties and parameters) in the cost-benefit analysis;

- SMhasFutureBenefit – points at the FutureBenefit class individual specifying future benefits (subcategories, their properties and parameters) considered in the cost-benefits analysis;
- SMhasQCAimpact – points at the QCAimpact class individual specifying overall impacts identified in the QCA analysis (their subcategories, analytical parameters and relations with other items of the model).

Please note that the three complex classes:

- InvestmentCost,
- OperatingCost,
- FutureBenefit,

represent a data model of the CBA component, and the complex class QCAimpact expresses the QCA component.

To assess whether the security measures are properly implemented and maintained, two mechanisms (the tool functionalities) are proposed:

- the simple resource management; each security measure requires minimal resources for its implementation and operation; these resources should be monitored to control residual risk;
- the performance indicators allowing to check if the security measure is effective in a life cycle and/or when critical obstacles occur.

Both these mechanisms go beyond the range of the current ValueSec use because they concern implementation and operation, rather than security planning.

The simple resource management checks if proper resources are applied for the implementation and later for the operation. These resources are identified during the cost-benefit analysis. The investment costs and the operation costs subcategories can be the foundation of the security measure implementation- and operation plans (the risk treatment plans). The resources with respect to time horizons are specified in these plans. For this reason the currently used SecurityMeasure class can be extended by properties dealing with the resources:

- SMhasReqImplemResources – points at the ReqImplemResources class individual, specifying overall resources of different kinds, required for the proper security measure implementation;
- SMhasReqOperResources – points at the ReqOperResources class individual, specifying overall resources of different kinds, required for the proper security measure operation.

Both classes represent the minimal resources which assure proper behavior of the security measure, i.e. allowing to control the risk at the planned level (the residual risk), expressed by the SMhasRiskAfter property of the SecurityMeasure class.

It is assumed that insufficient resources cause that the risk level planned during RRA is not achieved in reality. It means that the insufficient resources increase the planned risk level.

Checking whether current resources are sufficient, and how their insufficiency may increase risk, is called here the Resources-Risk Sensitivity (RRS) analysis. RRS is closely related to CBA (this component provides information about required resources for risk treatment plans) and RRA. Decreased resources of different kinds can be considered as additional “vulnerabilities” which should be considered during risk assessment with use of the RRA component. The RRS component can be based on the modified RRA component. This issue needs further analysis.

The second proposed mechanism concerns the performance indicators which allow to check if the security measure is effective in its life cycle or in individual situations, when critical obstacles occur.

The implementation of performance indicators is rather difficult, especially when the planned security measures are used for a single application, e.g. to secure a specific mass event, organized occasionally. Data types and sources to feed the indicators variables are diverse. Therefore sampling a reasonable data set to derive sensible conclusions for the improvements and corrections requires time and effort. For the permanent operations of the proposed security measures the situation is more favourable. Here it is possible to acquire much information specifying how the security measures behave in a real environment. On this basis the different performance indicators (and statistics) can be defined. These indicators can be used in real time to react to the critical situation and to correct the protection system. Additionally, the indicators can be analyzed periodically to elaborate continual improvement actions. The indicators depend strongly on the domain of their application. The examples of indicators are:

- number of incidents (or losses) of a given type in the specified time period,
- mean time required to manage the incident of a given type,
- number of false alarms.

This mechanism can be supported, e.g. by certain verifications or tests of the implemented security measures, performed outside the ValueSec framework, not discussed in this paper.

To implement these both mechanisms, two main options should be added to the horizontal ValueSec menu shown in Fig. 2:

- Resources-Risk Sensitivity Assessment, encompassing the risk treatment plan elaboration and maintenance, required resources specification, performing the assessment with the use of the RRS component, etc.;
- Performance indicators, including: the indicators related to maintenance, alerting, statistics, etc.

To extend the existing ValueSec toolset prototype, the assumptions and functional project of the software enhancements should be developed. This undertaking goes beyond this paper. It can be considered by the ValueSec team task and needs proper organization and funds.

VI. Conclusion

The paper concerns improvements of the ValueSec methodology and toolset, based on the experiences gained during the project execution, especially during the validation of the project results. The validation was based on five scenarios in different application domains, called here contexts. The scenarios and their use cases (representative samples of security measures) meet the expectations of broad decision makers' needs.

The paper discusses the ValueSec methodology, presenting the decision framework and its implementation as the software tool prototype.

ValueSec uses the following principles of work. For the given context and with respect to the given scenario, the security measures – candidates for the implementation are selected. Their assessments are performed, based on three pillars:

- RRA pillar, responsible for the assessment how the analysed security measure effectively affects the risk,
- CBA pillar, designed to assess if the security measure candidate is effective with respect to the assumed cost-benefit model criteria,
- QCA pillar, responsible for the identification of any political, social, legal, etc. restrictions, which can decrease the security measure operations in the future, exclude them, or mitigate them before the measure implementation.

Further in the paper, the possible enhancements are discussed, born during validation experiments.

Four possible enhancements of the ValueSec methodology are proposed as the fields for further researches:

- better support of the decision process by means of specialized tools,
- extension of the methodology beyond the planning phase, i.e. to the security measures implementation and operation phases,
- improving the preciseness of the risk assessments,
- introducing more precise risk models, which allow to consider cascading and escalation effects, especially in critical infrastructures.

The discussions of these four issues go beyond a single paper. For this reason, a more detailed discussion is provided only for the second issue.

A solution is proposed which allows to monitor the decreased security measures effectiveness during the operation caused by the shortage of resources. Moreover, performance indicators allowing the corrections in the security system and its continual improvement are discussed. The new RRS component can be considered as the ValueSec fourth pillar. The RRS is an analytic tool used to assess how decreasing resources can reduce the security measures performance. It can be implemented on the basis of the RRA component. The main extension is related to the analyzed vulnerabilities. The new methodology element considers the different shortages of resources as an additional source of vulnerabilities. The RRS component needs validation and experimentations on the real RRA component with the use of, for example, OSCAD, elaborated by EMAG.

Acknowledgment

I wish to thank my colleagues from the ValueSec team for their co-operation in the course of the project.

References

[1] *ValueSec web page*: www.valuesec.eu accessed 6 March 2014.

[2] E. Adar, C. Blobner, R. Hutter, K. Pettersen, "An extended Cost-Benefit Analysis for evaluating Decisions on Security Measures of Public Decision Makers", *CRITIS 2012, 7th International Conference on Critical Information Infrastructures Security*, Lillehammer, September 17-19, 2012.

[3] E. Bjorheim Abrahamsen, T. Aven, K. Pettersen, T. Rosqvist, "A framework for selection of strategy for management of security measures", *Proc. PSAM11 & Esrel 2012 Int'l conference*, Scandic Marina Congress Centre, Helsinki, Finland, June 25-29, 2012, USB memory stick, pp. 18-Tu2-4.

[4] J. Baginski, "Software support of the risk reduction assessment in the ValueSec project flood use case", in: *New results in dependability and computer system*, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, J. Kacprzyk, Eds.: *Proceedings of the 8th Int. Conf. on Dependability and Complex Systems DepCos-RELCOMEX*, Brunów, Poland, September 9-23, 2013, *Advances in Intelligent and Soft Computing*, Vol. 224, 2013, Springer-Verlag: Cham, Heidelberg, New York, Dordrecht, London, pp. 11-24. http://link.springer.com/chapter/10.1007%2F978-3-319-00945-2_2#page-1 DOI: 10.1007/978-3-319-00945-2_2.

[5] *Risk management – Principles and guidelines*, ISO 31000:2009.

[6] A. Białas, "Risk assessment aspects in mastering the value function of security measures", in: *New results in dependability and computer system*, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, J. Kacprzyk, Eds.: *Proceedings of the 8th Int. Conf. on Dependability and Complex Systems DepCos-RELCOMEX*, Brunów, Poland, September 9-23, 2013, *Advances in Intelligent and Soft Computing*, Vol. 224, 2013, Springer-Verlag: Cham, Heidelberg, New York, Dordrecht, London, pp. 25-39. http://link.springer.com/chapter/10.1007%2F978-3-319-00945-2_3#page-1 DOI: 10.1007/978-3-319-00945-2_3.

[7] M. Räikkönen, T. Rosqvist, L. Poussa, M. Jähi, "A Framework for Integrating Economic Evaluation and Risk Assessment to Support Policymakers' Security-related Decisions", *Proc. PSAM11 & Esrel 2012 Int'l conference*, Scandic Marina Congress Centre, Helsinki, Finland, June 25-29, 2012, USB memory stick, pp. 18-Tu3-2.

12th Conference on Advanced Information Technologies for Management

We are pleased to invite you to participate in the 10th edition of Conference on "Advanced Information Technologies for Management AITM'2012". The main purpose of the conference is to provide a forum for researchers and practitioners to present and discuss the current issues of IT in business applications. There will be also the opportunity to demonstrate by the software houses and firms their solutions as well as achievements in management information systems.

Topics

The topics of interest include but are not limited to:

- Concepts and methods of business informatics
- Business Process Management and Management Systems (BPM and BPMS)
- Management Information Systems (MIS)
- Enterprise information systems (ERP, CRM, SCM, etc.)
- Business Intelligence methods and tools
- Strategies and methodologies of IT implementation
- IT projects & IT projects management
- IT governance, efficiency and effectiveness
- Decision Support Systems and data mining
- Intelligence and mobile IT
- Cloud computing, SOA, Web services
- Agent-based systems
- Business-oriented ontologies, topic maps
- Knowledge-based and intelligent systems in management

Event Chairs

Dudycz, Helena, Wrocław University of Economics, Poland

Dyczkowski, Mirosław, Wrocław University of Economics, Poland

Korczak, Jerzy, Wroclaw University of Economics, Poland

Program Committee

Abramowicz, Witold, Poznan University of Economics, Poland

Ahlemann, Frederik, University of Duisburg-Essen, Germany

Andres, Frederic, National Institute of Informatics, Tokyo, Japan

Brown, Kenneth, Communigram SA, France

Chmielarz, Witold, University of Warsaw, Poland

Cortesi, Agostino, Università Ca' Foscari, Venezia, Italy

Czarnacka-Chrobot, Beata, Warsaw School of Economics, Poland

De, Suparna, University of Surrey, Guildford, United Kingdom

Dufourd, Jean-François, University of Strasbourg, France

Franczyk, Bogdan, Universitat Leipzig, Germany

Kannan, Rajkumar, Bishop Heber College (Autonomous), Tiruchirappalli, India

Kersten, Grzegorz, Concordia University, Montreal, Poland

Kowalczyk, Ryszard, Swinburne University of Technology, Melbourne, Victoria, Australia

Ligęza, Antoni, AGH University of Science and Technology, Poland

Ludwig, André, University of Leipzig, Germany

Maciaszek, Leszek, Wroclaw University of Economics, Poland and Macquarie University ~ Sydney, Australia

Magoni, Damien, University of Bordeaux – LaBRI, France

Michalak, Krzysztof, Wroclaw University of Economics

Pankowska, Malgorzata, University of Economics in Katowice, Poland

Stanek, Stanislaw, General Tadeusz Kosciuszko Military Academy of Land Forces in Wroclaw, Poland

Teufel, Stephanie, University of Fribourg, Switzerland

Tsang, Edward, University of Essex, United Kingdom

Zanni-Merk, Cecilia, Universite de Strasbourg, France

Ziemba, Ewa, University of Economics in Katowice, Poland

Position papers of the 2014 Federated Conference on
Computer Science and Information Systems pp. 211–217

DOI: 10.15439/2014F47
ACSIS, Vol. 3

Service Design and Distributed System Reliability in Intelligence Information System Based on Service-Oriented Architecture

Jugoslav Achkoski
Military Academy “General Mihailo Apostlski” –
Skopje, associate member of “Goce Delchev”
University - Shtip str. Vasko Karangelevski NoN,
1000 Skopje, Macedonia
Email: jugoslav.ackoski@ugd.edu.mk

Vladimir Trajkovik
Ss. Cyril and Methodius University, Faculty of
Computer Science and Engineering str. Ruger
Boskovik 16, 1000 Skopje, Macedonia
Email: vladimir.trajkovik@finki.ukim.mk

***Abstract*—This paper presents the model of Intelligence Information System (IIS) based on a Service-Oriented Architecture. In this paper we propose the new service’s model, based on the Intelligence cycle and other systems which are necessary for gathering Intelligence information and data.**

The paper is mostly focused on the system architecture and services design as a mainstream for definition of the services. Furthermore, additional attention is dedicated on the Distributed System Reliability (DSR).

I.Introduction

INTELLIGENCE Information System Model gives contribution in Homeland Security and Civil Military Emerging Risks assessment through the possibility of providing information in an appropriate way, by implementing pushing and pulling mechanisms into information systems, selection of data and creation of information from raw data that can be used in creating intelligence products and dissemination reports for the authorities.

In the Intelligence Information System, which is based on SOA, are applications written in different programming language. The service design should provide interoperability between applications. It indicates that the services written in different programming languages are capable to communicate. In this connotation, processing elements (web servers, application’s servers, sensors for collecting data and so on) can create distributed environment for sharing information. SOA based multi-tier approach provides legacy systems to be hooked up in a new infrastructure where the new systems and legacy systems can communicate without complexity in communication protocol (SOAP messages).

In the phase of creating distributed systems, it is crucial to have metric for distributed system reliability. It provides appropriate distribution of the system’s components, because introduced algorithm for distributed system reliability shows where the gaps in the systems are. Also, the designers of a system can achieve higher level of system reliability using the distributed system reliability metric. In the distributed system, each node can present service and each service can present system, subsystem or processing element. Consequently, each service in the architecture has certain value of reliability. These values of reliability are very important for system designers.

The paper is organized as follows. Section 2 presents related work about the research presented in the paper. Section 3 is dedicated to architecture of the system. In the Section 3 are presented different levels of the system architecture and how they are connected. Section 4 presents service design where it is mostly focused on service interface. Section 5 demonstrates the algorithm for computing distributed system reliability and we implement GEAR as an algorithm for the system reliability. Finally, in the Section 6, concluding remarks of the paper are presented.

II.Related work

In [3], the quality attributes of loose coupling and autonomy for services in the context of service-oriented architecture are given. In order for services to be influenced by these quality attributes, an evaluation should be done during the phase of development of service design. According to [3], the recent research is focused on the textual description of the desired quality attributes and the thereby resulting formalized metrics require more information than the already available, or are based on theoretical models that hamper their applicability. In this paper, we present quality indicators for unique categorization, loose coupling, discoverability and autonomy. Formalized metrics is created for each quality indicator, in order to measure service candidates and service design in the Service oriented architecture Modeling Language (SoaML) [4], the standardized language for modeling service – oriented architecture. To illustrate the metrics and to verify their validity, service candidates and service designs of a campus guide system as developed at the Karlsruhe Institute of Technology, are evaluated.

In [6], a study of service reliability and availability for distributed systems is presented. The study gives an application example in order to explain usefulness of the GEAR algorithm. Furthermore, in the paper is presented research about reliability of modeled centralized heterogeneous distributed system (CHDS). Also, in the paper is studied implementation of availability function of virtual machine.

III.Architecture of Intelligence Information System

General architecture of Intelligence Information System prototype is presented on Figure 1. As a result of system complexity, the solution is presented as a layer model of architecture.

On the lowest level, IIS prototype has distributed system which consists of heterogeneous databases. In this case, most important database for IIS is database which holds data for users who use it. Intelligence center has responsibility for this database.

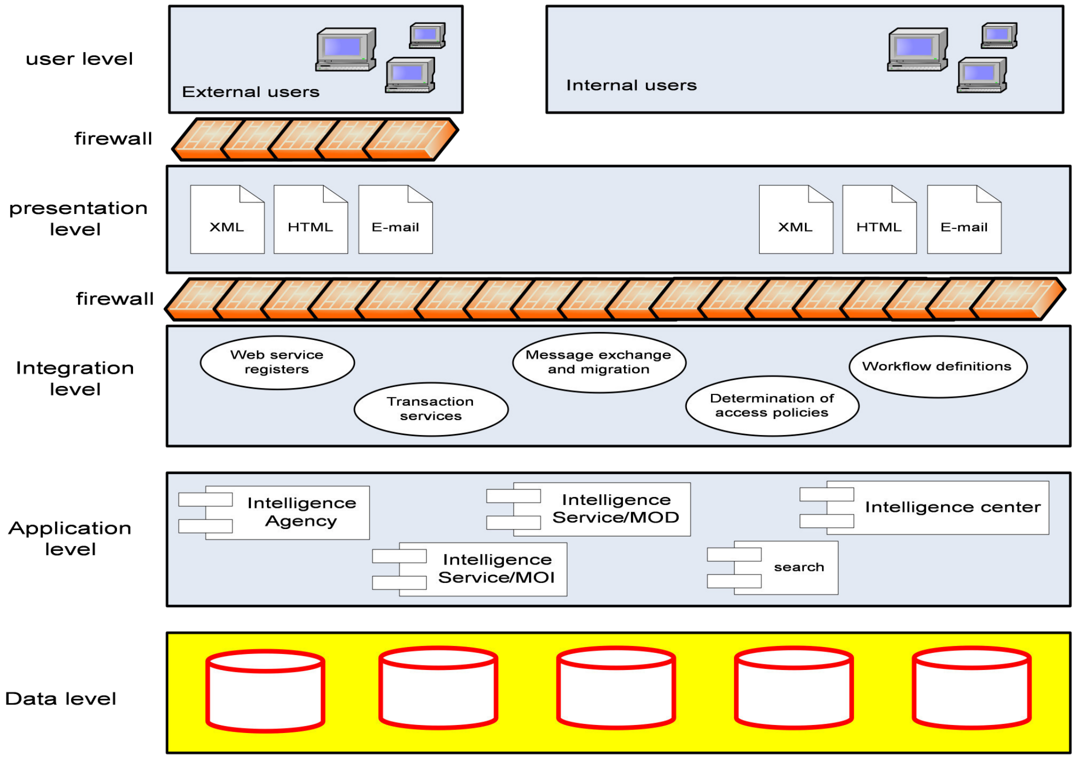


Fig 1. Prototype of Intelligence Information System Architecture

Access to a separate database will be made with application logic of module, which is part of internal information systems on government institution. This application should provide interfaces to the integration logic level [8], [9].

Integration level is a key level for our IIS model. That level should provide services through workflow which will be connected with modules of internal information systems and their transformation into web services. As a result of provided web services, integration level should exposed them into appropriate web services registers depending of security level. This level also govern security polices and polices for exchanging and adopting messages from different sources, in case of usage in comparable format. Finally, this level is taking care for governance of the services offered by IIS in a way of transactions when it is needed. With one sentences, this level is providing the functionality of the services in IIS.

The services should be available for different categories of users. For the purposes of protecting Intelligence Information System, firewall should be installed behind this level, which is followed by the level of presentational logic.

The presentation level can be implemented in a form of portal, which can offer: list of web services over approach to service registries, integration of web services with e-mails or directly as far procedure call of the applications (RPC) in a standard format (XML), but also as a ordinary HTML text for separated union of services – users. Exchanging information with external information systems is achievable through communication network, where IIS model is protected with another additional firewall. In this way, maximum protection from unexpected system failures is accomplished [8], [9].

IV.Service Design

In the process of service design it is possible services to be implemented in an existed platform or in the new platform, which will be created as a state-of-the art solution with straightforward purpose. This distinction is important, because our model allows web service to be implemented in both of these cases. Exploiting the services in such a way allows easy building of novel modules within information systems architecture and additionally, allows taking a pace with a contemporary ICT technology.

We propose coupling as a way for measuring service design. In the Intelligence Information System achieved desired level of coupling allows integration of subsystems and sensors as service providers with minimum number of connections between services. For example, if new sensor is added to the system, the communication that should be established between sensor and application server or other processing elements in the system does not imply that every server should establish communication with the new sensor [11].

In terms of service granularity (scope of functionality exposed by the services), it is the most convenient to create coarse-grained interfaces that implement a complete business process.

The coarse-grained interface should provide access to the data from different software artifacts and processing elements in the system depending of the user requirements [7]. It indicates that in the IIS sensors and other hardware components, which should be connected, have to exchange information in order to provide information for the senior decision makers or other end users. These components are based on different programming language (C++, JAVA, C and so on) and if the service's interface is not implemented in the applications, they could not exchange their data types. For instance, if application is written in JAVA and interface for this application is cre-

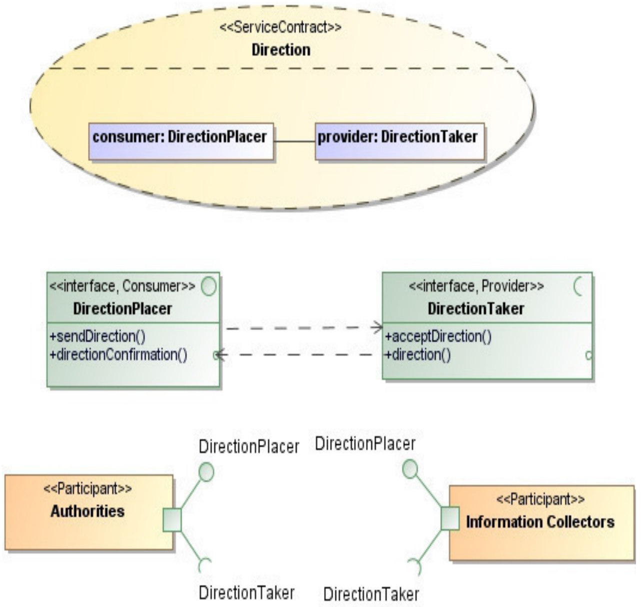


Fig 2. Specification of the Direction service, consisting of two roles, their respective consumer and provider interface type, and the corresponding ports on the participants

ated in JAVA, then application written in C++ could not use this JAVA interface, because data types (string, integer, float and so on) in JAVA and C++ are not treated in the same way. In our approach, the service's interface is based on XML, because applications written in different languages can exchange their data using WSDL (Web Service Definition Language). WSDL provides interoperability between applications.

A.Service contract based approach

The Service oriented architecture Modeling Language (SoaML) specification defines UML profile and metamodel for designing services within service-oriented architecture. Goals of SoaML refer to support activities at the stage of modeling and designing services and invoke them in model-driven development approach (MDA). It should support SOA in business and IT perspectives [1], [2].

SoaML specification defines three different types of approaches for specifying services:

- The simple interface based approach uses an UML interface to specify a one-way service interaction [1] [2].
- The service contract based approach extends an UML collaboration to specify a binary or n-ary service interaction [1], [2].
- The service interface based approach extends a UML class to specify a binary or n-ary service interaction [1], [2].

Different SoaML approaches recommend usage of divided UML parts which mean that reading SoaML specification is not understandable. Because of the reasons previously mentioned, problems in designing information systems emerge in software engineering [1].

A service contract based approach defines service specifications that define functions of service stakeholders (consumer and provider) and interface that implements these functions. In order to fulfill services' tasks interfaces must implemented services' function. Interfaces are types of ports in service-oriented architecture that requires each stakeholder to accomplish its task in the appropriate service contract [1].

The service contract based approach increases the UML collaboration in the model that present structured part of services' interactions. It can be used for specifying services that include contractual obligation, i.e. an agreement between two or more parties, which is relevant for circumstances of already established interaction patterns between the participants. These interaction patterns are used for exchanging messages and specifying interfaces between participants [1].

In order to demonstrate service contract based approach we are using services that make part of Intelli-

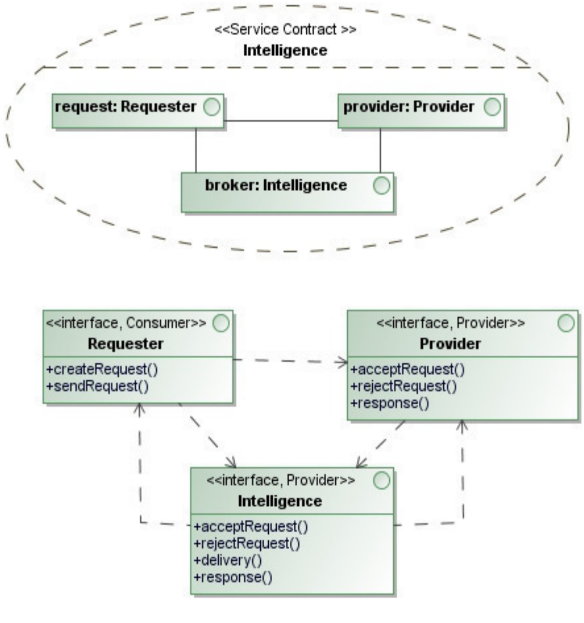


Fig 3. Specification of the Intelligence service contract, consisting of three roles, the respective consumer interface and the two provider interface types

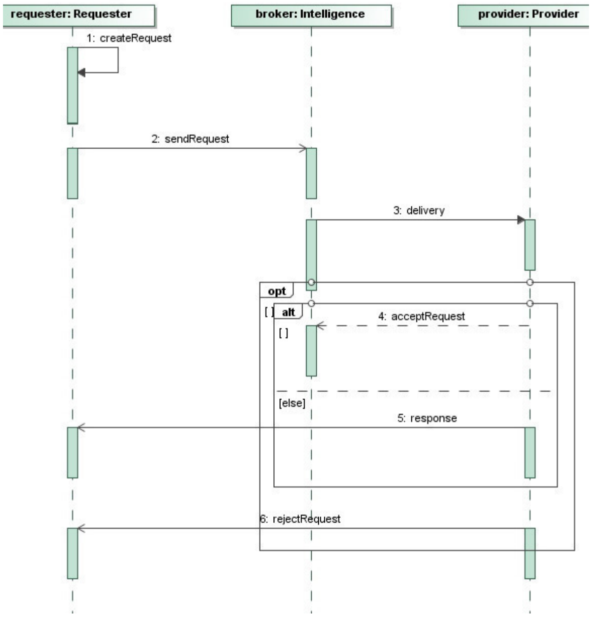


Fig 4. Specification of the Intelligence service choreography

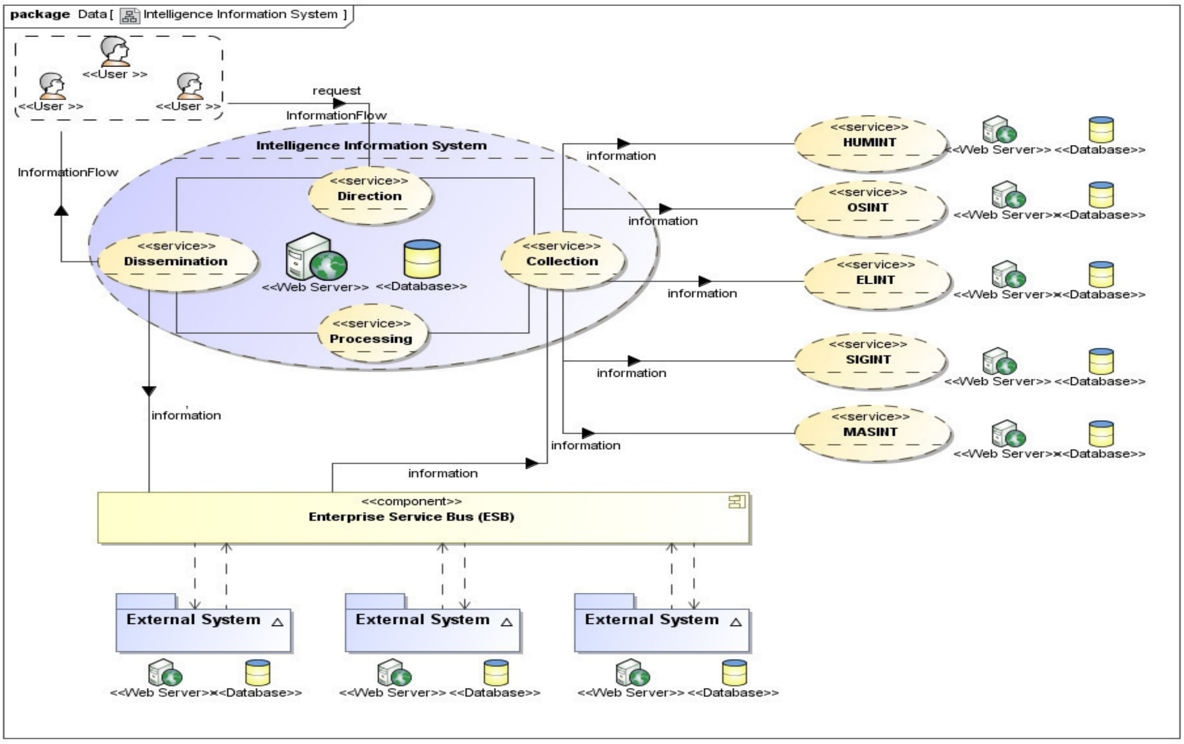


Fig 5. The dataflow in the Intelligence Information System - based on SOA

gence Information System. Services should contribute in defining a binary service contract, a multi-party service contract and a compound service contract that contributes to explaining service contract based approach. First, we suppose that Direction service contract can be modeled as two independent service contracts. One of them should specify an interaction for placing directions and another one should specify an interaction for taking directions in the process of intelligence information collection. Figure 2 shows specification of Direction service contract, consisting of two roles, namely their respective consumer and provider interface: DirectionPlacer and DirectionTaker [1].

The service contract shows that there is dependency between these two types of interfaces and they have to be modeled with UML dependencies. Participants use interaction in service contract and fulfill their tasks through appropriate interface. The binding between these interfaces is established by ports. From the role bindings in the services architecture we deduce that the Authorities have a request port typed by the DirectionPlacer interface, and the Information Collector has a service port typed by the DirectionTaker interface.

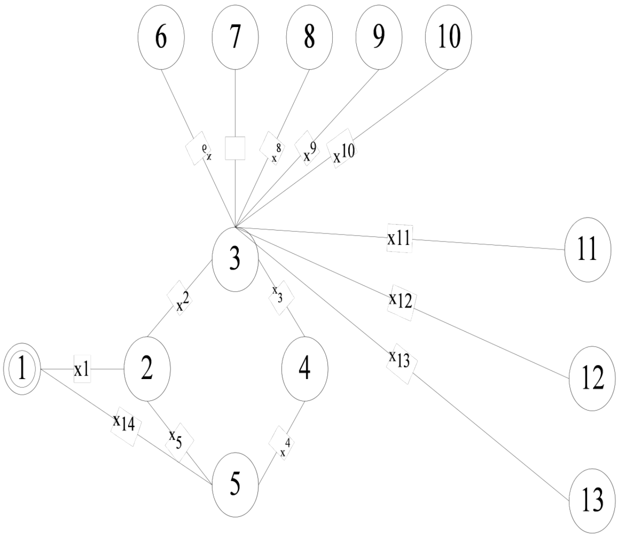


Fig 6. The graph of the Intelligence Information System - based on SOA

In this example service contract presents packing of two interfaces, providing that two interfaces are part of one service specification and not specified as a two independent service specification as two separate interface. Furthermore, it is recommended that a behavior on service contract is specified, i.e. a service choreography or a service protocol. Actually, there is a disagreement on whether a specification of service choreography should be used for understanding design of service interface in order to support exchanging message. SoaML is agnostic with regards to behavioral modeling and basically states that any UML behavior, e.g. interaction models, activity models or state machines, can be used [1].

The service contract based approach is convenient for specifying interaction between two or more roles that are introduced for establishing an agreement such as, for example, a message exchange. Service contract can be also applied as a reusable specification element, which can be re-used during the design time for connecting different stakeholders. In addition, this approach supports modeling of multiparty service contracts including three or more participants, as well as modeling of compound service contracts where the existing service contract can be used for defining several granular service contracts [1].

Figure 5 can be used to elaborate multiparty service contract. In our service model, we use Intelligence service contract where the interaction between the requester and

Table I.
Relations b/w an edge, a service name and vertices

A Node	A service name	A link	node vectors
1	end - user	x_1	1-2
2	Direction	x_2	2-3
3	Collection	x_3	3-4
4	Analyzing	x_4	4-5
5	Dissemination	x_5	5-2
6	HUMINT	x_6	3-6
7	OSINT	x_7	3-7
8	ELINT	x_8	3-8
9	SIGINT	x_9	3-9
10	MASINT	x_{10}	3-10
11	External System	x_{11}	3-11
12	External System	x_{12}	3-12
13	External System	x_{13}	3-13
14	Dissemination	x_{14}	5-1

the provider of information is mediated by Intelligence broker.

Figure 3 shows the specification of Intelligence service contract with three roles: information requester, information provider and intelligence information broker. These three roles have independent types of consumer and provider interfaces called Requester, Provider and Intelligence. The dependencies between the interfaces are explicitly modeled using UML dependencies. Stakeholders also have ports whose function is to connect services in service-oriented architecture.

Figure 4 shows the specification of the service choreography using UML interaction. Here we should notice that this is multiparty service contract, since the requester interacts directly with the information provider through delivering of messages. Except for the direct message delivering interaction, all other interactions pass through Intelligence broker. Service interaction starts with requirements for intelligence information toward the intelligence broker. At a later time a delivery is made which is either accepted or a grievance is sent to the broker and forwarded to the provider, who may file a justification in order to clarify whether to accept or ignore the requirements.

V.The Distributed System Reliability of the Intelligence Information System

The purpose of the GEAR algorithm is to compute accuracy of distributed computer system, which actually is composed of memory units, processing elements, and other hardware and software. Probability of application or service to be accurately executed in a distributed system is called availability of the distributed system.

In one distributed system, the nodes can present memory units, processing elements and programs (see Figure 5). The nodes can exchange data through a communication network in order to execute a program from separate nodes. Failures of the communication links or failures of the services in the distributed computer system, decrease the level of system performance and availability of the system. The level of success of one program in one node in distributed computer system depends on availability (successful program execution) of all other nodes, which indicates that the node have to be accessible from all other programs required for the appropriate program execution and also, communication links should be available without failures.

To compute DSR for IIS, we select GEAR, which is dedicated for computing reliability of links in computer networks. We introduce following assumptions:

- The RV maintains the information about links in the computer network [6];
- If link is operational, it has value 1 and if not, it has value 0 (faulty) in reliability expression;
- The “d” represents the link when we do not know whether link is operational or faulty and it is not computed in reliability expression;
- The LV has information about the edges that are traversed in the subnetwork [6];
- Each service presents self-contained processing element, and we can assume that they are self-contained node;
- The structure of the IIS is modeled as a graph and the graph does not have any loops [6], [9];

On Figure 6 is presented graph topology of distributed system, which is based on schema of Figure 5.

According to above mentioned assumptions, the dataflow in Figure 5 and the graph in Figure 6, in order to compute DSR of Intelligence Information System, we have to combine these pieces as in Table I. To be more precise, we can use the table to update RV and LV, because it gives a clear explanation how services, links and nodes are connected b/w each other.

In order to be computed reliability of computer distributed system, the GEAR algorithm requires both vectors RV and LV to be updated in the every iteration at each node. In order to be computed these two vectors in the GEAR algorithm are implemented simple rules without complexity.

To be updated Reliability Vector, we have to follow next two rules [5], [6]:

1. The updated RV value about new edge is obtained from the value of parent node and the value of vertices where the link is traversed from its parent edge.

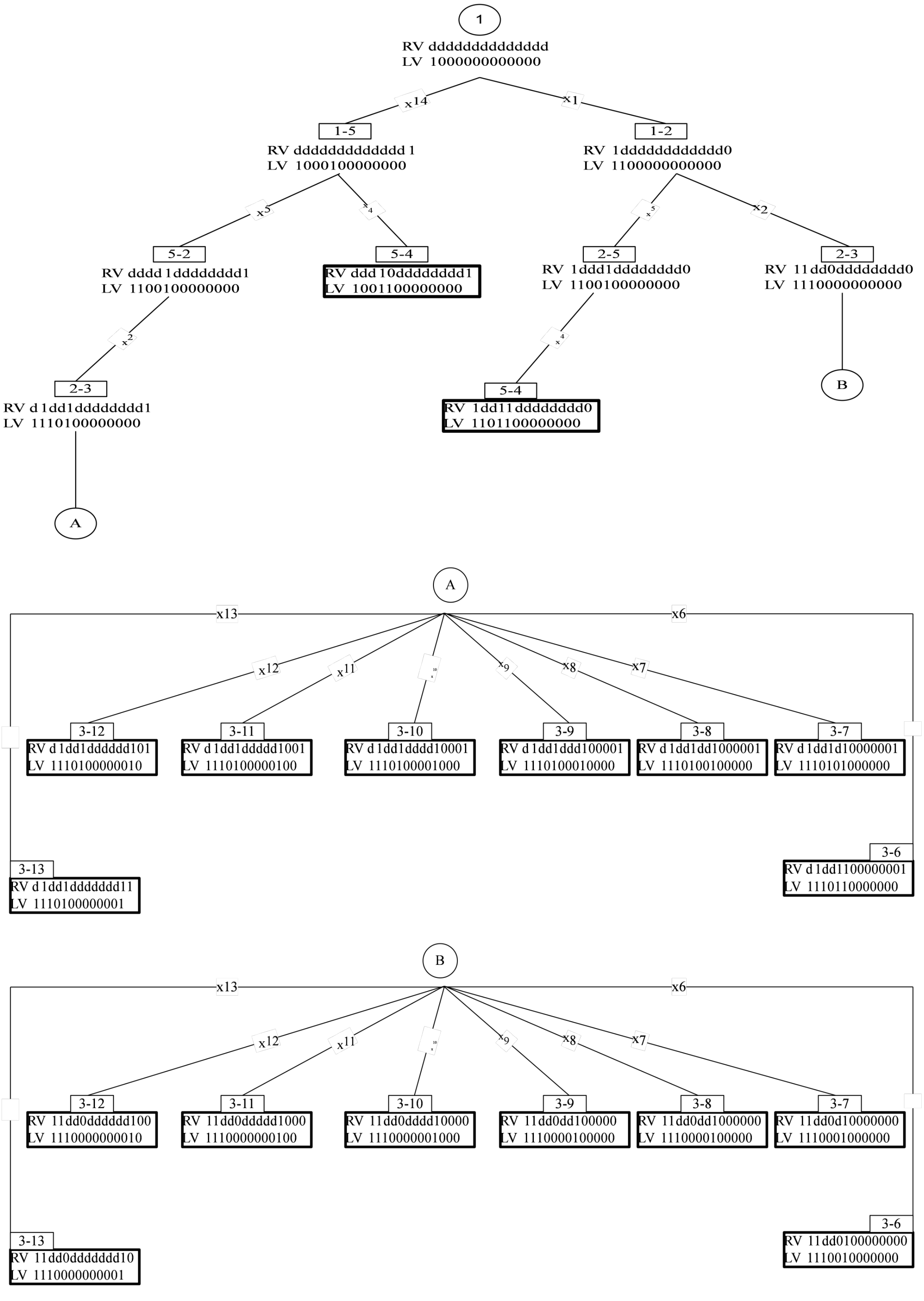


Fig 7. A complete tree for evaluating DSR in IIS

2. The edges, which are on the left side of the first updated edge (previous statement), are updated with the value of the edge, which is on their left side, and it is value 0 about the vertices and value 1 about vertices, which is traversed from its parent edge.

The intention of the LV is to avoid loops in the algorithm, which implies that one node is not traversed more than once [5], [6]. The updating of this vector is simple, which means that every node has value from the parent node in the tree and its value is represented with 1. Other nodes have value 0 and these nodes are not connected to the parent nodes and the vertices of the updated node.

In order to show, where in the tree the GEAR algorithm stop, we set up bold rectangles (Figure 7). The tree shows that the vertices in the graph from the Figure 5 have ending edges with the following numbers: 4, 6, 7, 8, 9, 10, 11, 12, 13. The starting edge is number 1.

In the each ending edge (bold rectangle), we can note the symbols as 1, 0 or d. In the ending edge where the value is 1, we replaced p and everywhere in the edge where the value is 0, we replaced q. The symbol d is not concerned with the computation expression because it is not involved in the reliability expression. It allows the expression (1) to be created for DSR.

$$\mathbf{DSR} = p_4q_5p_{14} + p_1p_4p_5q_{14} + p_1p_2q_5p_{13}q_{14} + p_1p_2q_5p_{12}q_{13}q_{14} + p_1p_2q_5p_{11}q_{12}q_{13}q_{14} + p_1p_2q_5p_{10}q_{11}q_{12}q_{13}q_{14} + p_1p_2q_5p_9q_{10}q_{11}q_{12}q_{13}q_{14} + p_1p_2q_5p_8q_9q_{10}q_{11}q_{12}q_{13}q_{14} + p_1p_2q_5p_7q_8q_9q_{10}q_{11}q_{12}q_{13}q_{14} + p_1p_2q_5p_6q_7q_8q_9q_{10}q_{11}q_{12}q_{13}q_{14} + p_2p_5p_{13}p_{14} + p_2p_5p_{12}q_{13}p_{14} + p_2p_5p_{11}q_{12}q_{13}p_{14} + p_2p_5p_{10}q_{11}q_{12}q_{13}p_{14} + p_2p_5p_9q_{10}q_{11}q_{12}q_{13}p_{14} + p_2p_5p_8q_9q_{10}q_{11}q_{12}q_{13}p_{14} + p_2p_5p_7q_8q_9q_{10}q_{11}q_{12}q_{13}p_{14} + p_2p_5p_6q_7q_8q_9q_{10}q_{11}q_{12}q_{13}p_{14} \qquad (1)$$

In Figure 7, DSR is calculated with the computation of p, and q. These two coefficients formulate the probability of every computer network link to be available for transferring data between services with probability p=0.9 (q=0.1) [6]. According to the expression (1) and replacement of the values p and q, we can obtain the result for DSR of Intelligence Information System.

$$DSR=0.8898$$

In conclusion of this section, we have to stress that obtained result about DSR is not on an appropriate level for the above mentioned system. Our intention is to create the system based on Service Oriented Architecture where we can obtain results of 0.977 out of 100% [10].

Furthermore, it is possible to increase the level of distributed system reliability, but we have to pre-plan the graph of computer network infrastructure with different ways of connections between edges and vertices. In addition, we have to pre-plan the ending edges where algorithm stops.

VI.Conclusion

The implementation of service-oriented architecture in Intelligence Information System increases the intelligence efficiency. Establishing a developmental methodology and designing the model can serve as a basis for building efficient information system.

The System architecture is explained in order to show the general concept of the system in terms of connectivity between system components, and relationship between layers. About the system architecture, we can firmly conclude that level of integration logic is a basis of the Intelligence Information System. Although, in the system architecture is diverse levels, only the level of integration logic is most significant for Service-Oriented Architecture. At the level of integration logic are set up most important services for appropriate system functioning.

The service design will contribute to the building Intelligence Information System based on an SOA platform because software artifacts can achieve a certain level of interoperability. Therefore, diverse hardware and software can exchange their data types in order to satisfy Intelligence functions. As a conclusion about service design, we can stress that core of interoperability relies on XML because WSDL and SOAP are based on XML.

The DSR provides the system's metric for reliability where many information systems rely on this metric. This metric provides reliable values for connecting nodes (processing elements, applications, I/O devices etc.) in distributed systems and it can be exploited in the early stage of information system development. Furthermore, DSR could be used for gathering testing data on further stage of system development. The obtained results from the tests will be taken from the equations for general metric about quality of service (QoS).

References

[1] B. Elvesæter, A.-J. Berre, A. Sadovykh, "Specifying Services using the Service oriented architecture Modeling Language (SoaML): A baseline for Specification of Cloud-based Services," in Proc. *1st International Conference on Cloud Computing and Service Science (CLOSER 2011)*, 7-9 May 2011. http://closer.scitevents.org/

[2] M. Gebhart, M. Baumgartner, S. Oehlert, M. Blersch, and S. Abeck, "Evaluation of Service Designs based on SoaML," in Proc. 5th *International Conference on Software Engineering Advances (ICSEA)* pp. 7-13, 2010, doi: 10.1109/ICSEA.2010.8

[3] M. Gebhart, S. Abeck, "Metrics for Evaluating Service Designs based on SoaML," *International Journal on Advances in Software*, vol. 4(1&2), 2011, pp. 61-75. http://iariajournals.org/software/

[4] OMG, "Service oriented architecture modeling language (SoaML) – specification for the UML profile and metamodel for services (UPMS)," Version 1.0 Beta 1, 2009

http://www.uio.no/studier/emner/matnat/ifi/INF5120/v10/undervisningsmateriale/09-12-09-SoaML.pdf

[5] A. Kumar, D.P. Agrawal, "A generalized algorithm for evaluating distributed-program reliability," *IEEE Trans. Reliability*, vol.42, Issue 3, pp. 416 – 426, Sep. 1993. doi: 10.1109/24.257825.

[6] Y.S Dai, M. Xie, K.L. Poh, G.Q. Liu "A study of service reliability and availability for distributed systems," *Elsevier, Reliability Engineering & System Safety*, vol. 79, Issue 1, 1 January 2003, pp. 103–112, http://dx.doi.org/10.1016/S0951-8320(02)00200-4

[7] M. P. Papazoglou, W.-J. van den Heuvel "Service-Oriented Design and Development Methodology," *International Journal of Web Engineering and Technology (IJWET)*, vol. 2 Issue 4, July 2006, pp.412-442.

[8] J. Achkoski, V. Trajkovik, and D. Davcev, "Service-Oriented Architecture Concept for Intelligence Information System Development," in Proc. *3rd international conferences on advanced service computing service computation 2011 (IARIA)*, Rome, Italy, September 25 - 30, 2011.

[9] J. Achkoski, V. Trajkovik,"Intelligence Information System (IIS) with SOA-based Information Systems," in Proc. *33rd International Conference on INFORMATION TECHNOLOGY INTERFACES, IEEE*, Cavtat/Dubrovnik, Croatia, June 27 - 30, 2011.

[10] J. Hurwitz, R. Bloor, C. Baroudi, and M. Kaufman, "Service Oriented Architecture (SOA) For Dummies," Hoboken, NJ 07030-5774: John Wiley & Sons. 2007.

[11] Priyantha, Nissanka B., et al. "Tiny web services: design and implementation of interoperable and evolvable sensor networks." Proceedings of the 6th ACM conference on Embedded network sensor systems. ACM, 2008.

DOI: 10.15439/2014F470
ACSIS, Vol. 3

A Fractal Measure for Comparing the Work Effort of Human and Artificial Agents Performing Management Functions

Matthew E. Gladden
Georgetown University, Washington, DC, USA; NeuraXenetica LLC, Indianapolis, IN, USA
Email: matthew.e.gladden@gmail.com

Abstract—**Thanks to the growing sophistication of artificial agent technologies, businesses will increasingly face decisions of whether to have a human employee or artificial agent perform a particular function. This makes it desirable to have a common temporal measure for comparing the work effort that human beings and artificial agents can apply to a role. Existing temporal measures of work effort are formulated to apply either to human employees (e.g., FTE and billable hours) or computer-based systems (e.g., mean time to failure and availability) but not both. In this paper we propose a new temporal measure of work effort based on fractal dimension that applies equally to the work of human beings and artificial agents performing management functions. We then consider four potential cases to demonstrate the measure's diagnostic value in assessing strengths (e.g., flexibility) and risks (e.g., switch costs) reflected by the temporal work dynamics of particular managers.**

I. The Need for a Common Temporal Measure of Work Effort

THE increasing power and sophistication of artificial agent technology is allowing businesses to employ artificial agents in a growing number of roles. Artificial agents are no longer restricted simply to performing logistical functions such as resource scheduling, but are now capable of more complex interpersonal workplace behavior such as using social intelligence to effectively manage the limitations, abilities, and expectations of human employees [1], recognizing and manifesting culture-specific behaviors in interactions with human colleagues [2], and assessing the performance of human members of virtual teams [3]. It is thus gradually becoming more feasible to design artificial agents capable of performing the four key functions carried out by human managers, which are planning, organizing, leading, and controlling [4].

As a result of such recent and anticipated future advances, businesses will increasingly be faced with concrete decisions about whether, for example, the manager of a new corporate call center should be an experienced human manager or the latest artificial agent system. Such decisions will be shaped by a large number of strategic, financial, technological, political, legal, ethical, and operational factors. One particular element to be taken into account is that of temporal work effort: i.e., how much time would a human manager actually be able to dedicate to carrying out the necessary work functions, given the fact that physiological, cultural, legal, and ethical constraints limit the number of hours per week that a human being is capable of working? Similarly, how much time would an artificial agent be able to dedicate to carrying out the necessary work functions, given the fact that scheduled maintenance or unscheduled outages can limit the uptime of computer-based systems? Knowing how much time per day (or week, or other relevant time interval) a manager will be available to carry out his or her functions of planning, organizing, leading, and controlling becomes especially relevant in an interconnected age when global businesses operate around the clock, and managers are expected to be available to respond to inquiries and make decisions at almost any time of the night or day.

In the case of human professionals, temporal measures such as 'full-time equivalent' (FTE) [5] and 'billable hours' are often used to quantify one's work effort. Computer-based systems, meanwhile, often use temporal measures such as 'availability' and 'reliability.' In the following sections, we will analyze such existing measures and then develop a new fractal-dimension-based temporal measure for work effort that has at least two notable advantages: it is applicable to the work effort of both human and artificial agent managers, and it provides valuable diagnostic insights into the strengths and dangers of an individual manager's temporal work dynamics that are not provided by existing measures.

II. Measures of Work Effort for Computer-based Systems

A. Availability and Reliability

A computer's reliability is often quantified as the mean time to failure (MTTF), the average length of time that a system will remain continuously in operation before experiencing its next failure [6]. The mean time to repair (MTTR) is the average length of time needed to detect and repair the failure and return the system to operation. A computer's steady-state availability A is the likelihood that the computer is operating at a particular moment, and is related to MTTF and MTTR in the equation [6]:

$$A = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}$$

A standard requirement for commercial computer systems is 99.99% availability over the course of a year [7].

Availability has traditionally been understood in a binary manner: a system is either 'up' or 'down.' Rossebeø et al. argue that a more sophisticated measure is needed that takes qualitative aspects into account and suggest recognizing a range of intermediate qualitative states between simply 'up' and 'down' [8]. As we explain below, the measure proposed in this paper takes a different approach: its unique diagnostic value comes not from adding a qualitative component but from considering more carefully the fineness and resolution of the time-scales on which measurements are being made.

B. Time-scales for Measuring Computer Performance

A computer performs actions across a vast range of time-scales. As Gunther notes, if a typical computer's CPU cycle were 'scaled up' so that it lasted one second, then using that same scale, a DRAM access would take about one minute, a single disk seek would require roughly 1.35 months, and a tape access would more than a century [7]. He explains that when measuring performance, "Only those changes that occur on a timescale similar to the quantity we are trying to predict will have the most impact on its value. All other (i.e., faster) changes in the system can usually be ignored.… In modeling the performance of a database system where the response time is measured in seconds, it would be counterproductive to include all the times for execution of every CPU instruction."

In a business context, artificial agents performing certain logistical or data-analysis tasks can operate at speeds constrained only by the laws of physics and availability of needed resources. However, an artificial agent manager whose role involves planning, organizing, leading, and controlling the activity of human colleagues should have its work effort measured within a corresponding time-scale. Thus for our present purposes there is no need to consider phenomena such as metastability that have major implications for computer design and functionality but are only directly relevant at the smallest temporal scale [7].

Viewed from the microscopic end of the temporal spectrum, the regulator of all activity within a computer system is the 'clock tick' or 'fundamental interval of time' created by an interrupt sent from the system's hardware clock to the operating system's kernel; in a Unix system, this tick interval is often set at 10 ms [7], during which time roughly 3.2×10^5 CPU cycles might occur. For an artificial agent system operating on a serial processor architecture, there is no need to adopt a temporal measure capable of resolving each individual CPU cycle, as that would not provide information that is directly relevant to the tasks in which the agent's work performance will be evaluated and which take place over a much longer time-frame. For example, an artificial agent manager might interact with human colleagues by generating text or images displayed on a screen. Assuming a screen refresh rate of 60 Hz, this yields a single work interval (or frame) of roughly 17 ms. Writing output data to disk would require a minimum work interval of roughly 3.50 ms for a disk seek [7]. If the artificial agent is generating speech or other audio to be heard by human beings, a standard sampling rate of 48,000 Hz would yield a single work interval of roughly 0.02 ms.

At the macroscopic end of the time-scale, it is not unknown for servers to run for several years without rebooting or a moment of downtime [9]. If we view an artificial agent manager as a form of enterprise software, we might expect its lifespan to average around 9 years and to be no shorter than 2 years [10]. Thus while a coarser or finer temporal resolution is possible, our proposed temporal measure for work effort should prove sufficient for artificial agent systems as long as it can encompass time-scales ranging from 0.02 ms up to several years.

III. Measures of Work Effort for Human Employees

A. The Significance of a Year as a Temporal Unit

We can now consider the case of a human manager. In principle, the longest possible macroscopic time-scale of work effort that one can utilize for a human employee is a biological lifespan. In practice, though, the relevant time-scale is obviously much shorter. In the United States, a typical managerial employee only remains with his or her current employer for about 5.5 years before moving to a new organization [11]. The 'year' has significant historical and conceptual value as a fundamental measure of human work activity. Just as enterprise system availability is often cited in terms of uptime per operating year, productivity figures for human workers are typically based on an annual time-frame [12].

Having taken the year as our initial frame of reference, how do we quantify the portion of a given year that a human employee actually spends on his or her work? For this purpose, the largest relevant subunit is that of a single week, as professional workers regularly assess a job's fringe benefits according to how many weeks of vacation they receive each year, and government agencies and researchers often track this data. The number of weeks worked per year varies significantly across nations and cultures [13].

B. The Significance of an Hour as a Temporal Unit

Even if we know that two employees both 'work' the same number of weeks per year, this fact does not yet tell us much about their relative levels of work effort, as it is possible for the employees to differ vastly in how many hours they work each week. Here, too, there is significant variation across nations and cultures and between specific jobs [13]. For example, an American law firm will likely expect attorneys to work more than 50 hours per week [14], while employees of high-tech Silicon Valley firms are routinely expected to work over 100 hours per week when project deadlines are approaching [15].

C. How Much Work in an Hour of Work?

The hour, though, is certainly not the smallest quantifiable interval of employee work effort. Two employees may consider themselves to have just spent an hour 'working,' but the number of minutes of work actually performed by each can differ greatly. In some professions, it is common to track work effort in sub-hour intervals. For example, attorneys with law American law firms typically track their work time in six-minute intervals and sometimes record and bill clients for work that took as little as one minute. For every hour that an attorney spends 'at work,' an average of roughly 45 minutes will count toward billable hours [14].

In other professions, employers have given up any attempt at precisely measuring how much time an employee is putting into their work, as the advent of new communications technologies has caused 'work time' and 'personal time' to meld into an indistinguishable blur [15]. The rise of multitasking and 'continuous partial attention' drives human workers to constantly monitor emails, texts, and instant messages, even while in the middle of meetings or conversations [16]. For knowledge workers, this continual checking of email can consume up to 25% of their workday [17]. While much of this nonstop communication activity is work-related, the existence of workplace phenomena such as shirking, social loafing, and job neglect means that a significant number of these electronic interruptions do not relate to work at all, but are purely personal. In particular, younger employees of the Millennial generation are less fond of email and tend to prefer text messaging, instant messaging, [18] and other forms of micro-communication that produce shorter but more frequent non-work interruptions to their work activities. Because professional employees can alternate between work-related and personal actions at such a rapid rate (once every few seconds, if not faster), it is now "very hard to tell when people are working and when people are not working," as a Silicon Valley executive reported in Shih's study [15]. In an effort to counteract this constant stream of distractions, some Extreme Programming (XP) teams employ the Pomodoro Technique, a time-boxing strategy in which physical timers are used to enforce a steady pace consisting of 25 minutes of focused work followed by a brief break [19].

D. Identifying a Minimum Time Unit of Work by Human Managers

Within a given period of 'work,' there may be alternating periods of work and non-work that are measured in seconds, not minutes. However, in attempting to identify the minimum unit of work of which humans are capable, it is valuable to consider time-scales even much smaller than a second. For example, scholars have estimated that the human brain is capable of between 10^{14} and 10^{16} calculations per second [20], [21], or roughly 6.6×10^{16} FLOPS [22], although the massively distributed parallel processing architecture of the brain [21] means that many calculations are taking place simultaneously, and the duration of a single calculation cannot be determined by simply dividing one second by, say, 10^{15}. In attempting to estimate the duration of a single 'calculation' performed by the brain, scholars have alternately cited the fact that an individual neuron can fire as often as 1,000 times a second [21], that "synapses carry out floating point operations … at a temporal resolution approaching about 1000 Hz" [23], that a neuron is capable of firing roughly once every 5 ms [24], or that the brain operates at a rate of speed of "around 100 cycles per second" [25]. These estimates yield a range of 1-10ms for the brain's smallest temporal unit of work activity.

It is helpful, though, to refer once more to Gunther's position on the measurement of computer performance: we can essentially ignore activity taking place within a system on a time-scale shorter than that of our work-relevant inputs and outputs, as it is "more likely to be part of the *background noise* rather than the main theme" [7]. In the case of a human being considered *qua* employee, the firing of a single synapse does not directly constitute 'work.' The work of planning, organizing, leading, and controlling for which human managers are employed typically involves more complicated inputs and outputs such as engaging in conversation or reading and creating documents. The smallest temporal unit of work would be the smallest unit relevant in the performance of such tasks.

That unit appears to be an interval of roughly 50 ms. Studies have shown that if one alternates too quickly between two tasks that require the same cognitive resources, one's performance on both tasks will be negatively impacted [26], as shifting from one mental task to another incurs a 'switch cost' of both a temporal delay and an increased error rate [27], which lowers productivity [28]. In particular, the human brain needs around 120 ms to fully allocate its attention to a new stimulus [29]. Marchetti cites diverse studies supporting the claim that the minimum 'integration time' needed for the brain to meld disparate sensory input into a conscious perception of a single event or experience is roughly 50-250 ms, with a median of about 100 ms [30]. These findings make it unlikely that a human manager would be capable of performing individual instances of work that need to be measured using a time-frame shorter than 50 ms. If one attempted to alternate between tasks faster than once every 50 ms, one's brain would not even have time to focus attention on a new task before abandoning it for yet another task.

E. Durations of Particular Work Inputs and Outputs

This minimum interval of roughly 50 ms is supported by the fact that the kinds of inputs and outputs that human managers typically utilize when performing work-related functions do not have durations shorter than this interval. For example, Hamilton notes that human beings can think at a rate of 400-800 words per minute, while we typically speak at 100-175 words per minute (with each spoken word comprising an average of 4-5 phonemes [31]). Optimal listening comprehension occurs when a speaker speaks at 275-300 words per minute, which gives a listener's mind less time to become distracted or daydream between each of the speaker's words [32]. Adult native speakers of English typically read 200-250 words per minute [33]. Regarding

TABLE I.
AVERAGE TIME NEEDED BY THE HUMAN BRAIN TO PERFORM WORK-RELATED INPUT, PROCESSING, AND OUTPUT FUNCTIONS

Activity	Average Time
Fully allocating attention to a new stimulus	120 ms
Consciously perceiving a single coherent experience or event	50-250 ms
Hearing one spoken phoneme (4.4 phonemes per word)	45-50 ms
Hearing one spoken word	200-220 ms
Reading one printed word	200-250 ms
Thinking one word	75-150 ms
Speaking one phoneme (4.4 phonemes per word)	20-140 ms
Speaking one word	90-600 ms
Typing one character (5.5 characters per word)	≤ 75 ms
Typing one word	≤ 400 ms
Writing one word in shorthand	≤ 170 ms

work output, the fastest sustainable typing rate is roughly 150 words per minute [34], with each word comprising an average of 5-6 characters; the fastest known shorthand writing speed is roughly 350 words per minute [35]; and the fastest known human speaker is able to clearly articulate more than 650 words per minute [36]. When these rates are converted into milliseconds, they yield the intervals seen in Table I.

F. The Fractal Self-similarity of Human Work Cycles

As we have seen, for artificial agent managers, the time-scales relevant to their work effort range from several years down to about 10 ms, while for human managers they range from several years down to around 50 ms. Within this range, there are multiple relevant time-scales and activity cycles of different lengths that demonstrate an interesting degree of self-similarity: within a given year of work, a typical human manager will spend many consecutive weeks working, interrupted periodically by non-work weeks of vacation. Within a given week of work, he will spend spans of several consecutive hours working, followed by non-work hours when he is asleep or out of the office. Within a given hour of work, his spans of minutes spent working will be followed by non-work intervals when he is daydreaming or writing a personal email. The roughly self-similar nature of this temporal dynamic opens the door to understanding a human manager's work activity as a fractal time series.

The fractal nature of our typical human work dynamics is not at all surprising: as Longo and Montévil note, fractal-like dynamics are "ubiquitous in biology, … in particular when we consider processes associated with physiological regulation" [37]. Lloyd notes that when an organism's biological processes operating on multiple time-frames displaying fractal temporal coherence, it creates a scale-free system with "robust yet flexible integrated performance" in which the oscillatory dynamics with long memory allow the organism to predict and respond to long-term environmental conditions such as tidal, seasonal, and annual cycles, while the short-term cycles coordinate internal processes such as organ functioning and cellular division [38].

IV. Calculating the Fractal Dimension of Work Effort

A. Significance of the Fractal Dimension

One of the most important and meaningful attributes of a fractal time series is that it possesses a *fractal dimension* that one can calculate and which captures valuable information about the series' temporal dynamics. The calculation of the fractal dimension of biological phenomena has varied practical applications. For example, analysis of the fractal dimension of EEG data can be used to quantify the level of concentration during mental tasks [39], and fractal analysis has demonstrated that healthy hearts display greater rhythmic complexity than diseased hearts [37].

The fractal dimension of empirically observed natural phenomena can be described by the equation $D = 2 - H$, where H is the Hurst exponent of the time series as graphed in two-dimensional Cartesian space. In this approach, an x-coordinate is the time at which a value was measured, and the y-coordinate is the value measured at that time [40]. The case $0 < H < ½$ represents a dynamic that is variously described as antipersistent, irregular, or trend-reversing: if the value in one moment is greater than the mean, the value in the next consecutive moment is likely to be less than the mean. The case $H = ½$ represents a random-walk process such as Brownian motion, in which the value in the next consecutive moment is equally likely to move toward or away from the mean. The case $½ < H < 1$ is described as persistent or quasi-regular: the value at the next consecutive moment in time is likely to be the same as the value in the previous moment [40], [41]. In this case, we can say that the dynamic has long memory.

B. Work Effort as a Time Series of Binary Values

Graphing a time-series in two-dimensional space is useful for natural phenomena such as earthquakes that occur at different times with different intensities [42]. However, in the case of developing a temporal measure for quantifying the work effort of human and artificial managers, we suggest that a different approach is warranted. Graphing an agent's work effort in two-dimensional space would be useful if the work effort displayed by a human or artificial agent manager at a particular instant of time were able to range across a continuous spectrum of values. However, in this case we have only a binary set of possible values: at any given instant, an agent is either focusing its attention on its work, or it is not. Marchetti draws on research from several areas of psychology to show that the human mind is incapable of dividing its attention between two different scenes, attitudes, or 'observational levels' at the same instant in time. (We would suggest that the same will likely be true for any artificial agent whose cognitive capacities are modeled closely on those of the human brain's neural network, as well as for any artificial agent governed by a computer program in the form of executable code.) As we saw above, the brain's attention mechanism is capable of alternating attention between two different thoughts or scenes with great rapidity (as in cases of so-called 'multitasking'), however in any given instant of time, our attention is allocated to at most one of those thoughts or scenes. This means that work effort cannot be quantified by saying, for example, that "At moment t, 70% of the agent's attention was dedicated to its work." Instead, one would say that "For all of the indivisible instances of attention that took place during time interval $[a, b]$, in 70% of those instances the agent's attention was focused on its work."

Mandelbrot notes that if the fractal dimension of a time series graphed in two-dimensional space is represented by the equation $D=2-H$, then the zero set (or any other level set) of the graphed time series would have fractal dimension [40]:

$$D=1-H \cdot$$

We can use this equation to relate the fractal dimension and Hurst exponent for work effort when we understand work effort as graphed on a one-dimensional line segment. The length of the entire segment represents the entire time available (such as a year, week, or hour) during which an agent can potentially be performing work. Those instants of actual work form the set that is graphed on the line segment, while instants of non-work do not belong to the set. With this binary approach, we can envision the depiction of an agent's work effort across time as a series of instances of work and non-work graphed on a line segment that resembles a generalized Cantor set in which the moments of work are those points contained in the set and moments of non-work are portions of a deleted interval. Because this is a graph of an empirically observed natural phenomenon rather than a purely mathematical object, it would have a minimum fineness and resolution: if our minimum unit of time is 10 ms and we graph a line segment representing one hour, it would comprise 3.6 × 10^6 such units of work or non-work.

In this context, the Hurst exponent takes on a different (and perhaps even counterintuitive) meaning. For a two-dimensional graph of a time series with $H \approx 0$, successive y-values alternate antipersistently around the mean, and the graphed line fills up a relatively large share of the two-dimensional space. For a one-dimensional graph of a binary time series, one might visualize the set as though it contains a single point that is able to slide back and forth along the x-axis to occupy many different x-values simultaneously, thus forming the set. For a set with high persistence ($H \approx 1$), the point may be locked to a single x-value, reflecting a process with long memory. For a set with low persistence ($H \approx 0$), the point 'forgets' where it is and is free to move up and down the line segment, occupying many different x-values. This conceptualization reflects the fact that the two-dimensional graph of an antipersistent time series will cross the horizontal line determined by the mean y-value at many different places, whereas the graph of a persistent process might only cross it once, and the graph of a random-walk process can intersect it either one or many times.

V. Formulating our Fractal Measure

A. Advantages of the Box-Counting Method

Different methods exist for calculating fractal dimension. A number of scholars prefer the Minkowski-Bouligand or box-counting dimension over alternatives such as the area-perimeter or power spectrum methods for estimating the fractal dimension of natural phenomena as diverse as seismic activity, electrical activity in the brain, and physical surface features at the nanometer scale [42], [39], [43]. Longo and Montévil argue that while it lacks some of the mathematical import found in other definitions of fractal dimension such as the Hausdorff dimension, the box-counting dimension has an advantage in that it can easily be applied to empirically observed phenomena [37].

In order to develop our comparative fractal measure of work effort for human and artificial agent managers, we have thus employed the box-counting method to estimate the temporal dynamics' fractal dimension. The box-counting dimension D of set F can be calculated as:

$$D = \lim_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta}\,.$$

Here $N_\delta(F)$ is the smallest number of sets of diameter δ that cover the set F [44]. When using the box-counting method to estimate the fractal dimension of natural phenomena, this can be done by calculating the average value of D that results when one empirically determines for multiple values of δ [45].

B. Calculation and Notation of our Fractal Measure

When we applied this approach to calculate the box-counting fractal dimension D for the work effort of particular hypothetical human and artificial agent managers, it yielded insights that could be useful for understanding, comparing, and enhancing the temporal work dynamics of such agents.

To accomplish this, we considered an agent's typical work effort as viewed across on three different time-scales or levels: 1) The set F_1 includes those weeks worked within a span S_1 of five years (or 260 weeks), for which the covering sets used for the box-counting estimation were δ_a = 4 weeks, δ_b = 2 weeks, and δ_c = 1 week. 2) The set F_2 includes those hours worked within a span S_2 of one week (or 168 hours), for which the covering sets used for the box-counting estimation were δ_a = 4 hours, δ_b = 2 hours, and δ_c = 1 hour. 3) The set F_3 includes those minutes worked within a span S_3 of one hour, for which the covering sets used for the box-counting estimation were δ_a = 1 minute, δ_b = 30 seconds, and δ_c = 15 seconds. Using the box-counting method, we calculated D_1, D_2, and D_3 for the time-scales F_1, F_2, and F_3, respectively, and averaged those values to produce a mean value of $D = \left(D_1,\ D_2,\ D_{3,}\right)$ for a particular agent. We then calculated the estimated value for the Hurst exponent for that agent's temporal dynamic with the equation $H = 1 - D$.

Drawing on the data considered in previous sections for the typical temporal performance of human professionals and artificial agents (envisioned as hardware and software systems), we present four specific hypothetical cases and the values of D and H calculated for each.

VI. Applying our Measure to Particular Cases

A. Temporal Dynamics of Human Manager A

Consider a hypothetical Human Manager A whose work effort approaches the maximum of which contemporary human beings are capable. This manager does not take any weeks of vacation during the five years worked in his position (S_1 = 260 weeks, $N_\delta(F_1)$ = 260 weeks). He concentrates exclusively on his career, working an average of 90 hours per week (S_2 = 168 hours, $N_\delta(F_2)$ = 90 hours). During the work day, he avoids all possible distractions and, relying on an approximation of the Pomodoro Technique, spends only 5 minutes of each 'work hour' not performing work-related functions (S_3 = 60 minutes, $N_\delta(F_3)$ = 55 minutes). We graphed each of these situations on a line segment that we then considered at three different temporal resolutions. Within the graph of the time series, a moment of work is indicated with a colored vertical slice, and a moment of non-work is indicated with an unshaded interval. A graph of the temporal work dynamics of Human Manager A is seen in Fig. 1 below. For an agent with these characteristics, we have calculated D = 0.962, H = 0.038, and availability (understood as the likelihood that any randomly-selected

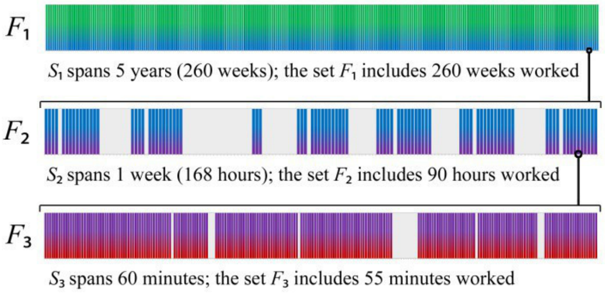


Fig. 1: Human Manager A's periods of work and non-work

instant of time will fall during a moment of work rather than non-work) as A = 49.1%.

B. Temporal Dynamics of Human Manager B

Hypothetical Human Manager B represents the opposite end of the spectrum: his time commitment approaches the lowest amount possible for someone who is fulfilling a management role with an organization. We suppose that Human Manager B spends only half of the weeks in the year working (S_1 = 260 weeks, $N_\delta(F_1)$ = 130 weeks). Even during those weeks when he is working, the manager dedicates only 10 hours of effort to this particular position (S_2 = 168 hours, $N_\delta(F_2)$ = 10 hours). Moreover, during each hour of 'work,' the manager spends only a third of the time focused directly on work-related tasks, with the rest of the time representing distractions or non-work-related activities (S_3 = 60 minutes, $N_\delta(F_3)$ = 20 minutes). A graph of the temporal work dynamics of Human Manager B is seen in Fig. 2 below. For an agent with these characteristics, we have calculated D = 0.532, H = 0.468, and A = 1.0%.

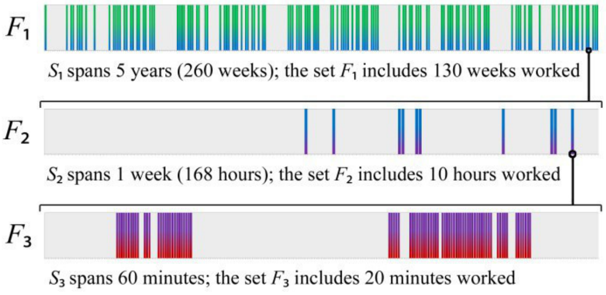


Fig. 2: Human Manager B's periods of work and non-work

C. Temporal Dynamics of Artificial Agent Manager A

Next consider a hypothetical Artificial Agent Manager A in the form of a software program running on a computer with a typical serial processor architecture. We suppose that during a given five-year operating period, there may be brief service outages for scheduled maintenance or updates but that there are no extended outages (S_1 = 260 weeks, $N_\delta(F_1)$ = 260 weeks). Each week, there is a scheduled maintenance window of one hour, when software updates are applied and the system is rebooted (S_2 = 168 hours, $N_\delta(F_2)$ = 167 hours). The software program and hardware substrate for Artificial Agent Manager A have no non-work-related functions and are not capable of being 'distracted' in the way that a human manager is, thus during a typical hour period of work, Artificial Agent Manager A does not dedicate any minutes to non-work-related functions (S_3 = 60 minutes, $N_\delta(F_3)$ = 60 minutes). A graph of the temporal work dynamics of Artificial Agent Manager A is seen in Fig. 3 below. For an agent with these characteristics, we have calculated D = 0.999, H = 0.001, and A = 99.4%.

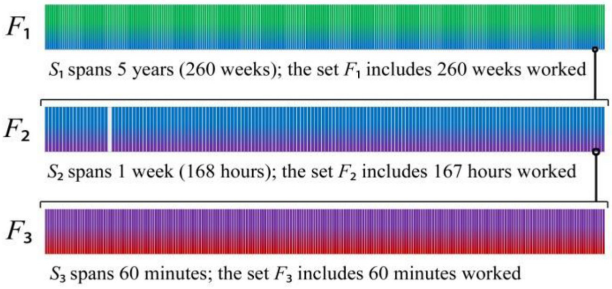


Fig. 3: Artificial Agent Manager A's periods of work and non-work

D. Temporal Dynamics of Artificial Agent Manager B

Finally, consider the hypothesized future scenario of Artificial Agent Manager B, an artificial general intelligence with a distributed neural network architecture that is modeled on the human brain and displays human-like motivations, emotions, and learning capacity [21]. While Artificial Agent Manager B enjoys its job, every two years it must spend a week away from work for a period of psychological assessment, maintenance, and relaxation, to reduce the likelihood of professional burnout (S_1 = 260 weeks, $N_\delta(F_1)$ = 258 weeks). Moreover, during each week of work, its neural network architecture requires it to spend two hours daily in a 'sleep' mode in which any new external stimuli are shut out, in order to facilitate the assimilation of the day's experiences into long-term memory. In order to maintain its capacity for creativity, satisfy its intellectual curiosity, and avoid the development of cyberpsychoses, it must also spend two hours daily exploring spheres of experience unconnected to its work-related tasks (S_2 = 168 hours, $N_\delta(F_2)$ = 126 hours). Because Artificial Agent Manager B reflects the full constellation of human-like cognitive and social behaviors, it spends five minutes of each hour on functions other than work, such as

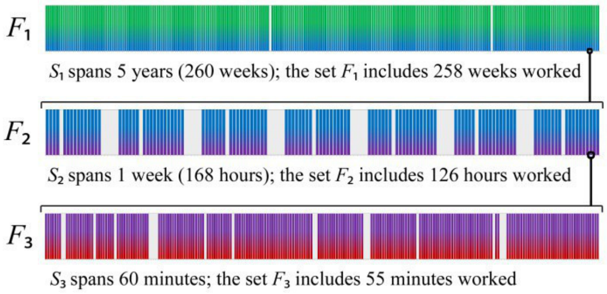


Fig. 4: Artificial Agent Manager B's periods of work and non-work

TABLE II.
AGENTS' WORK EFFORT AS CHARACTERIZED BY FRACTAL DIMENSION, HURST EXPONENT, AND AVAILABILITY

Agent	*D*	*H*	*A*
Artificial Agent Manager A	0.999	0.001	99.4%
Human Manager A	0.962	0.038	49.1%
Artificial Agent Manager B	0.945	0.055	68.2%
Human Manager B	0.532	0.468	1.0%

cyberloafing, following news stories, and communicating with friends (S_3 = 60 minutes, $N_\delta(F_3)$ = 55 minutes).

A graph of the temporal work dynamics of Artificial Agent Manager B is seen in Fig. 4 below. For an agent with these characteristics, we have calculated D = 0.945, H = 0.055, and A = 68.2%.

VII. ANALYSIS AND DISCUSSION

A. Comparison and Analysis

Table II below gives the values of *D*, *H*, and *A* for all four agents, ranked from the highest value of *D* to the lowest.

We may note the following conclusions:

1) Artificial Agent Managers A and B and Human Manager A all display similar values of $H \approx 0$ (antipersistence), while Human Manager B displays a value of $H \approx ½$ (randomness). While more study is required to verify this supposition, it seems likely that managers with low persistence (as understood in the mathematical sense defined above) would be free from high switch costs, as their work intervals last longer, and they spend a smaller share of their work time transitioning into or out of periods of work.

2) The managers displaying high values for *D* possess 'flexibility' in the sense that they are ready and available to work in almost every possible moment. However, they may simultaneously display 'inflexibility,' in the sense that they are *used to working* in every possible moment, thus unexpected interruptions may be more likely to derail the work of this sort of manager. Meanwhile, managers with a lower value for *D* possess 'flexibility,' insofar as they are already used to working only sporadically and juggling intervals of work amidst many other activities, thus unexpected interruptions to their work may not greatly faze them. On the other hand, they might simultaneously display 'inflexibility,' insofar as the bulk of their time may already be filled with non-work-related activity, leaving only brief, sporadic slivers of time available for work. If an unexpected distraction prevents them from working during one of these windows, it may be quite some time before another window of availability for work appears.

3) The values of *A* and *D* are neither directly nor inversely proportional to one another. Artificial Agent Manager A possesses the highest values for both *A* and *D*, while Human Manager B displays the lowest values for both. However, in the middle of the table, Human Manager A displays a higher value for *D* than Artificial Agent Manager B but a lower value for *A*. This means that if one only utilizes a simple measure such as availability in assessing (and ranking) the temporal work dynamics of human and artificial agents, one will miss out on additional information that the fractal dimension and Hurst exponent can provide. While availability is a useful measure, it can potentially be misleading if not complemented by more sophisticated measures such as fractal dimension.

B. Avenues for Future Research

Further steps that we have identified to advance this research include:

1) Gathering empirical data about temporal work dynamics from a sample of real-world human managers and artificial agent systems to verify the appropriateness and value of this fractal-dimension-based model. Analysis of such data could aid in predicting the temporal dynamics of future artificial agent systems (for which empirical data is not yet available) and designing more advanced artificial intelligence systems that will be capable of carrying out a wider range of business management roles.

2) Adding data for a time-scale S_4 that captures the work activity of human and artificial agent managers as viewed in intervals as small as 10 milliseconds. The ability to capture such data for the neural activity of a human manager exceeds the temporal resolution available with current fMRI technology, but it may be possible using EEG or MEG techniques (perhaps in conjunction with fMRI).

3) Attempting to identify correlations between the values of *D* and *H* for a particular manager's temporal dynamics and traits identified in established models of managerial motivation and behavior.

In conclusion, we hope that if this paper's proposal for a single fractal temporal measure of work effort that is applicable to both human and artificial agent managers proves useful, it might in some way contribute to the development of a new perspective in which an organization's human resources management and its management of artificial agent systems are seen not as two disconnected spheres, but rather as two aspects of a new, integrated discipline of human *and* artificial agent resource management.

REFERENCES

[1] M. Williams, "Robot social intelligence," in *Social Robotics*, Lecture Notes in Computer Science 7621, 2012, pp. 45-55, http://dx.doi.org/10.1007/978-3-642-34103-8_5.

[2] M. Rehm, Y. Nakano, E. André, T. Nishida, N. Bee, B. Endrass, M. Wissner, A. A. Lipi, and H. Huang, "From observation to simulation: generating culture-specific behavior for interactive systems," in *AI & Society*, vol. 24, no. 3, pp. 267-80, October 1, 2009, http://dx.doi.org/10.1007/s00146-009-0216-3.

[3] M. Nunes and H. O'Neill, "Assessing the performance of virtual teams with intelligent agents," in *Virtual and Networked Organizations, Emergent Technologies and Tools*, Communications in Computer and Information Science 248, 2012, pp. 62-69,[3] http://dx.doi.org/10.1007/978-3-642-31800-9_7.

[4] R. L. Daft, *Management*, 10th edition. Stamford, CT: Cengage Learning, 2011, pp. 7-8.

[5] Full-Time Equivalent (FTE). European Commission – Eurostat [Online]. Available [5] http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:Full-time_equivalent.

[6] M. Grottke, H. Sun, R. M. Fricks, and K. S. Trivedi, "Ten fallacies of availability and reliability analysis," in *Service Availability*, Lecture Notes in Computer Science 5017, pp. 187-206, 2008, http://dx.doi.org/10.1007/978-3-540-68129-8_15.

[7] N. J. Gunther, “Time—the zeroth performance metric,” in *Analyzing Computer System Performance with Perl::PDQ*, pp. 3-46, Berlin: Springer, 2005, http://dx.doi.org/10.1007/978-3-540-26860-4_1.

[8] J. E. Y. Rossebeø, M. S. Lund, K. E. Husa, and A. Refsdal, “A conceptual model for service availability,” in *Quality of Protection*, Advances in Information Security 23, pp. 107-18, 2006, http://dx.doi.org/10.1007/978-0-387-36584-8_9.

[9] Cool solutions: uptime workhorses: still crazy after all these years. (January 12, 2006). *Novell Cool Solutions* [Online]. Available http://www.novell.com/coolsolutions/trench/241.html.

[10] T. Tamai and Y. Torimitsu, “Software lifetime and its evolution process over generations,” in *Proceedings of 1992 Conference on Software Maintenance*, pp. 63-69, 1992, [10] http://dx.doi.org/ 10.1109/ICSM.1992.242557.

[11] Employee tenure summary. (September 18, 2012). United States Department of Labor, Bureau of Labor Statistics [Online]. Available http://www.bls.gov/news.release/tenure.nr0.htm.

[12] Annual Hours Worked. Organisation for Economic Co-operation and Development [Online]. Available [12] http://www.oecd.org/els/emp/ ANNUAL-HOURS-WORKED.pdf.

[13] L. Golden, “A brief history of long work time and the contemporary sources of overwork,” in *Journal of Business Ethics*, vol. 84, no. 2, pp. 217-27, Jan. 1, 2009, http://dx.doi.org/10.1007/s10551-008-9698-z.

[14] The Truth about the Billable Hour. Yale Law School [Online]. Available[14] http://www.law.yale.edu/studentlife/cdoadvice _truthaboutthebillablehour.htm.

[15] J. Shih, “Project time in Silicon Valley,” in *Qualitative Sociology*, vol. 27, no. 2, pp. 223-45, June 1, 2004, http://dx.doi.org/ 10.1023/B:QUAS.0000020694.53225.23.

[16] C. Sellberg and T. Susi, “Technostress in the office: a distributed cognition perspective on human–technology interaction,” in *Cognition, Technology & Work*, vol. 16, no. 2, pp. 187-201, May 1, 2014, http://dx.doi.org/10.1007/s10111-013-0256-9.

[17] A. Gupta, R. Sharda, and R. A. Greve, “You’ve got email! Does it really matter to process emails now or later?”, in *Information Systems Frontiers*, vol. 13, no. 5, pp. 637-53, November 1, 2011, http://dx.doi.org/10.1007/s10796-010-9242-4.

[18] A. Hershatter and M. Epstein, “Millennials and the world of work: an organization and management perspective,” in *Journal of Business and Psychology*, vol. 25, no. 2, pp. 211-23, June 1, 2010, http://dx.doi.org/10.1007/s10869-010-9160-y.

[19] F. Gobbo and M. Vaccari, “The Pomodoro Technique for sustainable pace in extreme programming teams,” in *Agile Processes in Software Engineering and Extreme Programming*, Lecture Notes in Business Information Processing 9, pp. 180-84, 2008, [19] http://dx.doi.org/ 10.1007/978-3-540-68255-4_18.

[20] R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books, 2006. Kindle version, locations 2080-2120.

[21] J. Friedenberg, *Artificial Psychology: The Quest for What It Means to Be Human.* Psychology Press, 2011, pp. 24-26, 179-91, 199-200.

[22] A. Llarena, “Here comes the robotic brain!”, in *Trends in Intelligent Robotics*, Communications in Computer and Information Science 103, pp. 114-21, 2010, http://dx.doi.org/10.1007/978-3-642-15810-0_15.

[23] J. L. McClelland, “Is a machine realization of truly human-like intelligence achievable?”, in *Cognitive Computation*, vol. 1, no. 1, pp. 17-21, March 1, 2009, http://dx.doi.org/ 10.1007/s12559-009-9015-x.

[24] R. Kurzweil, *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Penguin Books, 2000. Kindle version, location 1754.

[25] L. F. Abbott and P. Dayan, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge: MIT Press, 2001. As cited in A. Llarena, “Here comes the robotic brain!”, in *Trends in Intelligent Robotics*, Communications in Computer and Information Science 103, pp. 114-21, 2010, [25] http://dx.doi.org/ 10.1007/978-3-642-15810-0_15.

[26] S. W. Brown and S. M. Merchant, “Processing resources in timing and sequencing tasks,” in *Perception & Psychophysics*, vol. 69, no. 3, pp. 439-49, April 1, 2007, http://dx.doi.org/10.3758/BF03193764.

[27] T. D. Wager, J. Jonides, and E. E. Smith, “Individual differences in multiple types of shifting attention,” in *Memory & Cognition*, vol. 34, no. 8, pp. 1730-43, December 1, 2006, [27] http://dx.doi.org/ 10.3758/BF03195934.

[28] M. C. Schippers and R. Hogenes, “Energy management of people in organizations: a review and research agenda,” in *Journal of Business and Psychology*, vol. 26, no. 2, pp. 193-203, [28] http://dx.doi.org/ 10.1007/s10869-011-9217-6.

[29] P. U. Tse, J. Intriligator, J. Rivest, and P. Cavanagh, “Attention and the subjective expansion of time,” in *Perception & Psychophysics*, vol. 66, no. 7, pp. 1171-89, October 1, 2004, [29] http://dx.doi.org/ 10.3758/BF03196844.

[30] G. Marchetti, “Observation levels and units of time: a critical analysis of the main assumption of the theory of the artificial,” in *AI & Society*, vol. 14, no. 3-4, pp. 331-47, September 1, 2000, [30] http://dx.doi.org/ 10.1007/BF01205515.

[31] W. Levelt, “Models of word production,” in *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 223-32, 1999, [31] http://dx.doi.org/10.1016/ S1364-6613(99)01319-4.

[32] C. Hamilton, *Essentials of Public Speaking.* Stamford, CT: Cengage Learning, 2014, p. 60.

[33] M. J. Traxler, *Introduction to Psycholinguistics: Understanding Language Science.* Hoboken, NJ: John Wiley & Sons, 2011, Ch. 10.

[34] World’s fastest typer. *Chicago Tribune* [Online]. Available http://www.chicagotribune.com/sns-viral-fastest-records-pictures-018,0,193476.photo.

[35] “New World’s Record for Shorthand Speed: Nathan Behrin Transcribed 350 Words in a Minute With Only Two Errors,” *New York Times*, December 30, 1922.

[36] Fastest talker. Guinness World Records [Online]. Available http://gwrstaging.untitledtest.com/world-records/1/fastest-talker.

[37] G. Longo and M. Montévil, “Scaling and scale symmetries in biological systems,” in *Perspectives on Organisms*, Lecture Notes in Morphogenesis, Berlin: Springer, 2014, pp. 23-73, http://dx.doi.org/10.1007/978-3-642-35938-5_2.

[38] D. Lloyd, “Biological time is fractal: early events reverberate over a life time,” in *Journal of Biosciences*, vol. 33, no. 1 pp. 9-19, March 1, 2008, http://dx.doi.org/10.1007/s12038-008-0017-8.

[39] O. Sourina, Q. Wang, Y. Liu, and M. K. Nguyen, “Fractal-based brain state recognition from EEG in human computer interaction,” in *Biomedical Engineering Systems and Technologies*, Communications in Computer and Information Science 273, pp. 258-72, 2013, http://dx.doi.org/10.1007/978-3-642-29752-6_19.

[40] B. B. Mandelbrot, *The Fractal Geometry of Nature.* London: Macmillan, 1983, pp. 353-54.

[41] J. M. Valverde, A. Castellanos, and M.A.S. Quintanilla, “Looking for memory in the behavior of granular materials by means of the Hurst analysis,” in *Of Stones and Man: From the Pharaohs to the Present Day*, J. Kerisel, Ed. London: Taylor & Francis Group, 2005, pp. 817.

[42] L. Telesca, V. Cuomo, V. Lapenna, and M. Macchiato, “On the methods to identify clustering properties in sequences of seismic time-occurrences,” in *Journal of Seismology*, vol. 6, no. 1, pp. 125-34, January 1, 2002, http://dx.doi.org/ 10.1023/A:1014275509447.

[43] Y. Zhang, Q. Li, W. Chu, C. Wang, and C. Bai, “Fractal structure and fractal dimension determination at nanometer scale,” in *Science in China Series A: Mathematics*, vol. 42, no. 9, pp. 965-72, September 1, 1999, http://dx.doi.org/10.1007/BF02880388.

[44] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications.* Hoboken, NJ: John Wiley & Sons, 2004, pp. 41-44.

[45] B. Wahl, P. Van Roy, M. Larson, and E. Kampman, *Exploring Fractals on the Macintosh.* Reading, MA: Addison-Wesley Professional, 1994, pp. 75-108.

9th Conference on Information Systems Management

THIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from two complimentary directions: management of information systems in an organization, and uses of information systems to empower managers. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in an organization. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome.

Topics

The areas and topics of interest include, but are not limited to two groups:

- Management of Information Systems in an Organization:
  - Modern IT project management methods
  - User-oriented project management methods
  - Business Process Management in project management
  - Managing global systems
  - Influence of Enterprise Architecture on management
  - Effectiveness of information systems
  - Efficiency of information systems
  - Security of information systems
  - Privacy consideration of information systems
  - Mobile digital platforms for information systems management
  - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers:
  - Achieving alignment of business and information technology
  - Assessing business value of information systems
  - Risk factors in information systems projects
  - IT governance
  - Sourcing, selecting and delivering information systems
  - Planning and organizing information systems
  - Staffing information systems
  - Coordinating information systems
  - Controlling and monitoring information systems
  - Formation of business policies for information systems
  - Portfolio management,
  - CIO and information systems management roles

Event Chairs

Arogyaswami, Bernard, Le Moyne University
Chmielarz, Witold, University of Warsaw, Poland
Karagiannis, Dimitris, University of Vienna, Austria
Kisielnicki, Jerzy, University of Warsaw, Poland
Ziemba, Ewa, University of Economics in Katowice, Poland

Program Committee

Bialas, Andrzej, Institute of Innovative Technologies EMAG, Poland
Christozov, Dimitar, American University in Bulgaria, Bulgaria
Csikosova, Adriana, The Technical University of Košice, Slovakia
DeLorenzo, Gary, California University of Pennsylvania, United States
Dima, Ioan Constantin
Espinosa, Susana de Juana, University of Alicante, Spain
Gafni, Ruti, The Academic College Tel-Aviv-Yaffo, Israel
Geri, Nitza, The Open University of Israel, Israel
Grabara, Janusz, Czestochowa University of Technology, Poland
Jelonek, Dorota, Czestochowa University of Technology, Poland
Kersten, Grzegorz, Concordia University, Montreal, Poland
Kobyliński, Andrzej, Warsaw School of Economics, Poland
Kohun, Frederick, Robert Morris University, United States
Koohang, Alex, Middle Georgia State College, United States
Lasek, Mirosława, University of Warsaw, Poland
Levy, Yair, Nova Southeastern University - Graduate School of Computer and Information Sciences (GSCIS), United States
Modrak, Vladimir, The Technical University of Košice, Slovakia
Niedźwiedziński, Marian, University of Lodz, Poland
Pańkowska, Małgorzata, University of Economics in Katowice, Poland
Pastuszak, Zbigniew, Maria Curie-SKlodowska University, Poland
Phusavat, Kongkiti, Kasetsart University in Bangkok, Thailand
Rizun, Nina, Alfred Nobel University, Dnipropetrovs’k, Ukraine
Rouibach, Kamel, Kuwait University, Kuwait

Ruzic-Dimitrijevic, Ljijana, Higher Education Technical School of Professional Studies, Novi Sad, Serbia

Schroeder, Marcin, Akita International University, Japan

Skovira, Robert, Robert Morris University, United States

Stanek, Stanisław, The General Tadeusz Kościuszko Military Academy of Land Forces in Wrocław, Poland

Świerczyńska-Kaczor, Urszula, Jan Kochanowski University in Kielce, Poland

Travica, Bob, University of Manitoba, Canada

DOI: 10.15439/2014F463
ACSIS, Vol. 3

Functionality Assessment and Requirements of the IT System in an Education Institution

Ljiljana Ruzic-Dimitrijevic
Higher Education Technical
Schoolof Professional Studies,
Novi Sad,Serbia
Email: ljdimitrijevic@gmail.com

Milorad Stevic
Higher Education Technical
School of Professional Studies,
Novi Sad, Serbia
Email: milorad.stevic@live.com

Abstract—**The information technology (IT) system in one higher education institution which serves as a support to many processes, with special accent on teaching process is described in this paper. The objective of the paper is to underline the significance of the IT system for an education institution as well as the required changes in this system due to the implementation of "Bologna" study rules. Apart from description of the system and advantages allowed by this system, difficulties and problems which occur in further development and maintaining of this system are indicated**

I. Introduction

EACH company has many resources important for its business. Today, the information system has become very important and irreplaceable resource for successful business operation and management. IT (information technology) is a strategic resource. Managers have to be involved in design of information system [1]. The IT system of a higher education institution is particularly important. It must be used as integral part of all processes of this institution: teaching, enrolment, finances, administration, marketing, researches etc.

New study rules, resulting from "Bologna" system, have brought a lot of positive moving toward a much better quality in the teaching process and its outcomes. In order to achieve the goals of Bologna, the teaching staff had to raise the awareness about the using of new contemporary educational methods and on the important role IT takes in it.

In this paper, we will focus only on the part in relation to the teaching, from the students' side, as well as from the side of the teaching staff. With our transition to the "Bologna" curricula and syllabi [2] and with changes in the Law on higher education [3], it was necessary to introduce new categories in managing the higher education institution [4]. One of the important items is completely new information system which can follow the teaching and studying processes by the new rules.

The emergence of this IS is represented in the papers [5] and [6], while this paper will try to assess the implementation of that system and its contribution to the more efficient teaching process, greater success of the students, as well as mutual satisfaction of the users (administrative staff, teachers and students).

We will stress the way of developing the IT system and advantages of such an approach. Our goal in this paper is not to investigate project designing of the information system, but to point out its importance and benefits in educational processes. In addition, we will mention some problems regarding the system maintenance.

Development of IS was launched by the IT professors engaging several advanced students in their final study year. Over time, the IT team of students at specialist studies was formed, and led by the assistant, who was on the postgraduate studies IT and who, by his enthusiasm and practical experience, transferred to advanced students, succeeded to create atmosphere where the mere entrance into his team [7] represented the matter of prestige.

The thing we need to point out as the advantage of this working method is firstly the fact that the software is developed by the teaching staff. Considering all processes and their connection, with continual consultation with the staff of student service, the managing structure was directing and coordinating development of the system [6]. The implementation of the new study rules through IS improved understanding and accepting of the new rules by the teachers and staff.

Thanks to this working method, our school is most likely has made the greatest step forward in following the new rules of studies "according to Bologna" in Serbia.

Since the IT team in charge for developing the system is consisted of students and the teaching staff, it means that all of them are participants in the processes and that they have deeper insight in them and have greater knowledge on their requirements than certain software houses, which produce typical software, often with many unnecessary elements, and without possibility for efficient use of certain necessary functions specific for the individual company.

On the other side, insufficiencies of such work can be seen exactly in a composition of the team, which is variable, because students mainly leave the school after graduation and the team gets the new students all the time, but they start to work on the system which is already developed.

II. FUNCTIONALITY ANALYSIS

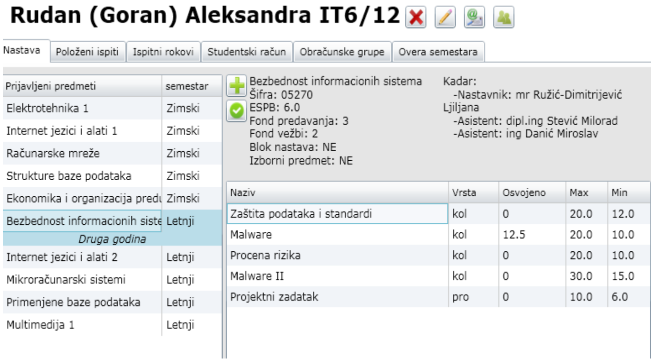
The risk assessment has become a very current and important for successful business as well as for successful realization of a project.

The risk of our project can be particularly analyzed. Different methods of risk assessment can be used, but what has become the fact proven in practice, it is that the new IS has mitigated and speeded up the work for the teaching staff of the School, students meet their requirements far more easier, and the teachers carry out their obligation with higher quality.

In order to gain deeper insight in the area of IS (information system) implementation in the School, we need to consider all processes which are carried out within that institution. We will focus only on the teaching processes [6]:

- Students' enrolment of the school year with registering of subjects in each semester
- Teaching
- Students assessing during tuition delivery and exams
- Tuition fee charges
- Issuing of certificates, degrees and students record

Each of these processes is included in the information system. Upon students' enrolment, the student's record is created so the student can access to it over the Internet and see its content with all data relevant for his activities. Students can register for the exams and subject they want to attend by online system. Their record contains all data about previously passed exams, payments, debts, as well as about improvement during semester for each subject and points they win after fulfilling pre-exam obligations. There are no previous crowds and a long line of people in front of office any more.



Rudan (Goran) Aleksandra IT6/12

Prijavljeni predmeti	semestar
Elektrotehnika 1	Zimski
Internet jezici i alati 1	Zimski
Računarske mreže	Zimski
Strukture baze podataka	Zimski
Ekonomika i organizacija predu	Zimski
Bezbednost informacionih siste	Letnji
Druga godina	
Internet jezici i alati 2	Letnji
Mikroračunarski sistemi	Letnji
Primenjene baze podataka	Letnji
Multimedija 1	Letnji

Bezbednost informacionih sistema
Šifra: 05270
ESPB: 6.0
Fond predavanja: 3
Fond vežbi: 2
Blok nastava: NE
Izborni predmet: NE

Kadar:
-Nastavnik: mr Ružić-Dimitrijević Ljiljana
-Asistent: dipl.ing Stević Milorad
-Asistent: ing Danić Miroslav

Naziv	Vrsta	Osvojeno	Max	Min
Zaštita podataka i standardi	kol	0	20.0	12.0
Malware	kol	12.5	20.0	10.0
Procena rizika	kol	0	20.0	10.0
Malware II	kol	0	30.0	15.0
Projektni zadatak	pro	0	10.0	6.0

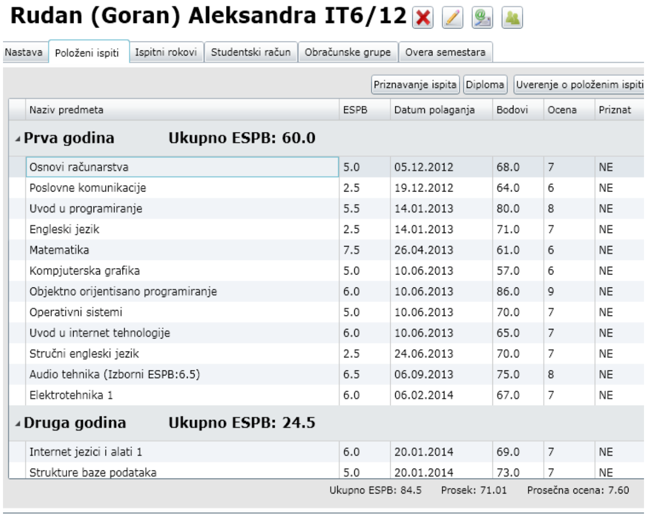
Fig. 1 Student's record – pre-exam obligations

Work of the student service is far more efficient, the possibilities for mistakes are much lesser and easier to notice, because of possibility for both monitoring of students activities, on one side, and activities of administration, on the other side, by simple access to the data in system. Relationship with students is free from tension and misunderstanding in that way. The newest function developed in this system is creating of diploma supplement, which has been done separately till now and requested special engagement of the staff of student service with great probability for mistakes, and that is the reason for all previous multiple verifications.

At first glance, teaching staff may have gained more obligations because of the new system for monitoring of students' work and their grading during entire semester, and not only on exam at the end of teaching. This system introduces an order and reduces subjectivity in grading.

The teacher must give clear rules of working on his subject and regularly and precisely fill in pre-exam obligations which student fulfills. In that way, the final grade will be more realistic and reflect student's work during entire semester. The teacher can simply communicate with the students registered for listening of the subject, by generating group mail via e-mail.

The system is posted on the school intranet and teachers have access from their own cabinet computers. For now, access over the Internet is offered only to a few teachers, although it would be more comfortable to have possibility for access out of the intranet in the school. The main reason is insecurity of user in terms of unawareness of the importance of protection, but it is a part of the future plan when condition for the higher level of security will be present.



Rudan (Goran) Aleksandra IT6/12

Naziv predmeta	ESPB	Datum polaganja	Bodovi	Ocena	Priznat
Prva godina **Ukupno ESPB: 60.0**					
Osnovi računarstva	5.0	05.12.2012	68.0	7	NE
Poslovne komunikacije	2.5	19.12.2012	64.0	6	NE
Uvod u programiranje	5.5	14.01.2013	80.0	8	NE
Engleski jezik	2.5	14.01.2013	71.0	7	NE
Matematika	7.5	26.04.2013	61.0	6	NE
Kompjuterska grafika	5.0	10.06.2013	57.0	6	NE
Objektno orijentisano programiranje	6.0	10.06.2013	86.0	9	NE
Operativni sistemi	5.0	10.06.2013	70.0	7	NE
Uvod u internet tehnologije	6.0	10.06.2013	65.0	7	NE
Stručni engleski jezik	2.5	24.06.2013	70.0	7	NE
Audio tehnika (Izborni ESPB:6.5)	6.5	06.09.2013	75.0	8	NE
Elektrotehnika 1	6.0	06.02.2014	67.0	7	NE
Druga godina **Ukupno ESPB: 24.5**					
Internet jezici i alati 1	6.0	20.01.2014	69.0	7	NE
Strukture baze podataka	5.0	20.01.2014	73.0	7	NE

Ukupno ESPB: 84.5 Prosek: 71.01 Prosečna ocena: 7.60

Fig. 2 Student's record – exam results

What is particularly useful for the teaching process, apart from introducing pre-exam obligations for the students and monitoring of their work, is statistical review of pass-rate of students which every teacher can generate for his own subjects. These data are extremely useful for the analysis of the teachers work and their self-evaluation. The percent of students which fulfill their obligations and pass exam represent special indicators. By analysis of these data, the teacher can estimate if the pre-exam obligations are too easy to be done, if great number of students fulfill pre-exam obligations and considerably smaller number of students pass exam, as well as if they are too hard, if very small number of students fulfill the obligations and don't have right even to take the exam. By constant correction, teacher can improve quality of his work.

Besides, in order to achieve even better results, the teacher will also do his best to give good lectures and teaching material to his students, so they can become more successful. Distance learning system which is used in certain programs opens possibility for the teachers to use hybrid system posting the teaching material online and asking the students to get introduce with it before lecture.

The open source software is used in this part, but the idea for developing our own software emerges within our project, which would support distance learning system and grading fulfilling the characteristics of our teaching system.

III. Architecture of information system

The Information system of the Higher Education Technical School of Professional Studies in Novi Sad is projected and realized as multilayer, service oriented architecture. The classical layers can be identified in the architecture: the data management layer, service layer with realized web services and client layer with realized client interface.

Each of these layers keeps on to decompose further, so that the data management layer contains a relational database management system, that is to say, system for the structured data management and system for the unstructured data management; the service layer contains two layers within itself – one for accepting and realization of requirement for manipulation over the structured data and one for accepting and realization of requirement for manipulation over unstructured data, while client interface for the access by PC computer is developed on the client side, but client interface is developed also for the access to the system over the new generation devices – smart phones and tablets.

In this way described multilayer service oriented architecture is shown in Table 1. The data management layer takes care of two types of data which are stored in the information system, structured and unstructured data.

For the first one, it is characteristic to have database scheme and this type of data is modeled by relational language for description and data management based on first-order predicate logic. Practical implementation is carried out using Microsoft SQL Server system for the relational database management system.

For the second type of data, it is characteristic that they don’t have database scheme while management system for these data is relied on new type of system for the unstructured data management-NoSQL. Practical implementation is carried out by MongoDB system for the unstructured data management.

A special attention is dedicated to the binary files, because it is hard to predict capacities for storing such type of data, and because of unpredictability of their number, structure and size; and that’s why horizontal scalability of the system for the binary files management is so important.

For the practical implementation, the same platform as for the unstructured data is used, Mongo DB database, because this platform has special specification intended for these structures-GrisFS management. This specification is suitable for practical implementation because of the feature such as a secure horizontal scalability, tolerance on interrupted work of projected number of servers, as well as optimization of server load used in the system.

Service layer is divided in accordance to division which is also carried out on layer for data management, on service layer which deals with manipulation over structured data, and on service layer that deals with manipulation over unstructured data. Practical implementation of the first service layer is carried by use of .NET technology and C# programming language. By using of these technologies, the services based on SOAP protocol were developed, and their purpose is to, on the request of user over the client interface, carry out manipulation over structured data on a layer of data of the information system.

The second type of service is developed to serve for different application – to allow standardized manipulation over unstructured data. In accordance to that fact, technology and protocols used for this purpose are also different. Practical implementation of this service layer is carried out by using Python programming language and these services are based on a REST protocol which is more efficient for work with unstructured data and with greater number of data, which is useful in a case of manipulation over binary files. It is possible to carry out the manipulation over data stored in a SQL Server Database, but also in a MongoDB database from any service layer.

Client layer of the information system is developed according to expected users of the information system. The most frequent users are users who use PC computers in order to realize interaction with the information system and for these users the application based on C# programming language and Silver light/Pipelight technology is developed.

TABLE I.
CONCRETE MULTILAYER SERVICE ORIENTED ARCHITECTURE OF THE INFORMATION SYSTEM

Layer	**Data type/Device type/Client**	**Practical implementation**	**Technology/OS**
Data management layer			
	Structured data	SQL Server	RDBMS/Windows
	Unstructured data and binary files	MongoDB	NoSQL/Linux
Service layer			
	Structured data	C#	SOAP/Windows
	Unstructured data and binary files	Python	REST/Linux
Client layer			
	Desktop, web	C#, Silverlight, Pipelight	.NET/Plugin
	Smart devices (phones, tablets)	AngularJS	JavaScript/HTML

The number of second group of users is also growing, and these are the users of smart devices, phones and tablets, who need to establish interaction with information system. For these users, different type of user interface is developed, practically realized by using AngularJS programming language and HTML technology. As in the case of service layer, it is possible to realize communication with both service layers from any client interface.

IV. PROTECTION OF THE INFORMATION SYSTEM

The information system is projected and realized in such a way to offer high level of data protection. Data protection is carried out on levels which are the subject of studying in literature: there is protection of data at rest, data in motion/data in transit and data in use. Protection of data at rest is carried out all the time.

Both databases use techniques of replication and creation of back up data at regular time intervals. Then, these copies are automatically copied on protected location. This part of protection is done due to potential physical damages on servers.

Protection of data in motion is done by using the encryption, that is to say, by using of https protocol.

Data in use are protected by projecting and realizing of the subsystem for approval to access to data, which present advanced technique of data access based on group of users.

Instead of this access, the access based on data classification on sensitive and less sensitive is used, as well as on the user account management in order to define precisely which data are accessible for each user. It allows that only authorized users can manipulate data which are approved for manipulation. Such approach allows also efficient monitoring of user activities in the system because of later analysis of these activities and eventual identifying the misuse of data.

V. THE PROBLEMS IN MAINTAINING AND FURTHER DEVELOPMENT OF THE INFORMATION SYSTEM

Conceived and realized as the ultimate IT product, with respect for the most modern standards, rules and recommendations which exist in the world, the system has already been working for three years with very modest equipment without interruption, crash and security data violation. Apart from all previously mentioned, this information system is ill-equipped because of several problems:

Lack of regularly engaged engineers in maintaining and continual development

- Only one person is engaged in developing and maintaining of the information system all the time, but partly, and mostly in teaching activities and a few individuals more who are engaged occasionally, but partly on these tasks, with contracts which will expire in next several months
- Syndrome of concentrating all business activities in the organization on IT center, that is to say, syndrome of considering the information system as a whole which will individually adjust itself to the business needs of organization.
- Lack of open minds for the new ideas and business application in further development of the information system
- Unwillingness to accept the information system as very important product for the main activity of the

organization, as a direct and indirect generator of financial gain for the organization.

VI Conclusion

The application of “Bologna” system in teaching and learning processes is very poor if there is no support of a suitable information system. Besides supporting these processes, the information system following Bologna curricula offers more quality work. Both teachers and students have to respect the required policy that is implemented in the information system and everything is transparent. The opportunity of electronic tracking of students’ success provides a possibility of enhancing both teaching and learning

This paper describes the problems that may arise in the development and maintenance of the information system. However, it has helped us to have a better insight into all flows of information and required new commitments of teachers and student services regarding the supply and use of data. The increased concern about students and a better communication with them is just a small detail from an incredibly large contribution to their success in general.

References

[1] B.Gates, Business @ the Speed of Thought,. Warner Books, Inc., New York, 2001

[2] The Bologna Declaration. (2000). On the European space for higher education: An explanation Retrieved November 9, 2012, from http://ec.europa.eu/education/policies/educ/bologna/bologna.pdf

[3] Zakon o visokom obrazovanju (Law on higher education), (2005). Službeni glasnik RS, (Official Gazette of Republic of Serbia) 76/2005, Belgrade, Serbia.

[4] J. Dakic, “Uloga i značaj menadžmenta u visokom obrazovanju” (The role and importance of management in higher education), SYMORG; XIII International Symposium Innovative management and business performance. 2012

[5] Z. Lovrekovic, L. Ruzic-Dimitrijevic, and B. Nikolic, “Information system implementation based on process approach at higher education institutions“, Proceedings of the 2007 Computer Science and IT Education Conference, Mauritius,2007, pp. 454-461.

[6] L. Ruzic-Dimitrijevic, and B. Nikolic, “Designing and building an information system for a higher education institution“, Proceedings of the Informing Science & IT Education Conference 2008, Bulgaria, pp. 283-300.

[7] B.Nikolic, J. Dakic, L. Ruzic-Dimitrijevic, “Contemporary management in a higher education institution in Serbia“, Online Journal of Applied Knowledge Management Volume 1, Issue 1, 2013 pp72-81

3rd Workshop on Information Technologies for Logistics

THE main purpose of the workshop is to provide a forum for researchers and practitioners to present and discuss current issues concerning use of ICT in logistic applications (hardware and software). There will be also an opportunity for hardware integrators, software developers and logistics companies to demonstrate their solutions, as well as achievements, in different logistic systems.

Topics

The topics of interest include but are not limited to:

- Innovations in information systems supporting logistics and its management (WMS, SCM, TMS, LIS, VMI, CRP, PLM, and others)
- Innovative technologies in warehouse management: RFID, Voice Picking, Image Recognition, Pick Radar, etc.
- Logistics process modeling, including influence of warehouse automatic
- Optimization of logistics processes:
  - optimal vehicle routing and management, boundary conditions
  - optimal picking routing (global optimization, fast search, collision prediction and prevention)
  - shared mobility systems
  - day-to-day dynamic traffic assignment models
  - effective methods of picking (multi picking, batch picking ect.)
  - relationships between picking efficiency and products decomposition in warehouse area
- Environmental protection (for example carbon-aware transportation)
- Artificial intelligence systems and decision support systems in logistics
- BI, data mining and process mining in logistics
- Quality management algorithms and methods
- Material Flow Theory and applications

Event Chairs

Gontar, Beata, Uniwersity of Łódz, Poland

Program Committee

Azavedo, Susana, University of Beira Interior, Portugal

Balicki, Jerzy, Gdansk University of Technology, Poland

Banaszak, Zbigniew, Warsaw University of Technology, Poland

Bobkowska, Anna, Gdansk University of Technology, Poland

Bruzda, Jaonna, Nicolaus Copernicus University, Poland

Duran-Grados, Vanesa, University Cadiz, Spain

Fosner, Maja, Faculty of Logistics, University of Maribor, Slovenia

Franczyk, Bogdan, Universitat Leipzig, Germany

Goh, Thong Ngee, National University of Singapore

Gontar, Zbigniew, University of Lodz, Poland

Jedliński, Mariusz, University of Szczecin, Poland

Korczak, Jerzy, Uniwersytet Ekonomiczny we Wrocławiu, Poland

Langviniene, Neringa, Kaunas University of Technology, Lithuania

Lent, Bogdan, University of Bern, Switzerland

Liao, Da-Yin, National Chi Nan University, Taiwan

Malavasi, Gabriele, University of Rome, Italy

Montemanni, Roberto, University of Applied Sciences of Southern Switzerland, Switzerland

Pamuła, Anna, University of Łódź, Poland

Patasiene, Irena, Kaunas University of Technology, Lithuania

Ricci, Stefano, SAPIENZA Università di Roma, Italy

Semenov, Louri, West Pomeranian University of Technology, Poland

Shinkevich, Алексей Иванович, Kazan National Research Technological University

Sitek, Pawel, Kielce University of Technology, Poland

Tipi, Nicoleta, University of Huddersfield, United Kingdom

Zielinski, Jerzy, University of Lodz, Poland

DOI: 10.15439/2014F130
ACSIS, Vol. 3

Cloud-based Cooperation of Logistics Service Providers in Logistics Cluster Organisations

Jan Oberländer
Netzwerk Logistik Leipzig-Halle e.V.,
Terminalring 13, 04435 Leipzig/Halle Airport,
Germany
Email: jan.oberlaender@logistik-mitteldeutschland.net

Bogdan Franczyk
University of Leipzig,
Information Systems Institute,
Grimmaische Straße 12, 04109 Leipzig, Germany
E-Mail: franczyk@wifa.uni-leipzig.de

Abstract* — In today's logistics market, small and medium-sized logistics companies are exposed to increasing cooperation pressure to use rationalization potentials, win new groups of customers and assert themselves in the competition with big logistics companies yielding a distinct market power. In the past, numerous SME-size logistics companies have joined to form regional logistics associations in order to assess possible cooperation and develop cooperation potentials. In a parallel process, logistics cloud service providers develop new types of cooperation-supporting logistics cloud services. This paper describes a method of how regional logistics associations can apply a cloud-based cooperation development process. The method was developed as part of a doctoral thesis and evaluated first hand by expert interviews in several logistics association in relation to the application involved. The results of the evaluation suggest that a suitable approach was found by which regional logistics associations can tap the potentials of logistics cloud services for their member companies.

I. Introduction

Globalization and the new internet-based capabilities of ready informational networking among companies impose and enable new value-added structures known as bottom-up economy. The structure and process related nature of the bottom-up economy is dramatically different from the top-down economy of the past in that it follows a logic of cooperation among smaller, locally based value-added units flexibly combining to form larger structures to generate complex products and services [1]. This is referred to as open production by production managers and suggests that new technical opportunities might give rise to structural changes also in the logistics sector in future.

From the angle of small and medium-size logistics companies the challenge is to be able to forge ties of cooperation with other logistics service providers quickly and with as little input of resources as possible in a situation in which the cooperative business processes must be handled efficiently with the aim of providing joint logistics services in the market. The present stock of software used by small and medium-size logistics companies systematically supports isolated internal functions and is not made for easy and quick integration with the software applications of complementary partners in the value-added process. Innovative, cooperation-supporting and multitenant cooperation-enabling logistics cloud services, resp. hold out the promise of new opportunities for the short-term establishment and termination of logistics ties of cooperation without the need or risk of investment associated with conventional software applications. Meanwhile the integration issue between cooperation partners on the IT level can essentially be cleared up by the usage of the same cloud software installation.

The problem at present is less that of the availability of the suitable logistics cloud services than the introduction of new approaches for the shared use of software among SME logistics companies. Logistics associations can provide new technical approaches as source of innovation and as multiplier. The degree to which this can be realistic and the way in which such a cloud-based cooperation development process can be designed will be described below.

The argumentation structure of this paper and the underlying doctiral thesis aims on the following thesis: Logistics associations can contribute to increasing the competitiveness of small and medium-sized logistics enterprises by using the described method. This can be achieved by increasing the cooperation ability of the small and medium-sized logistics enterprises. Therefore multitenant cooperation-enabling logistics cloud services are recommended for the cooperation development that would not be used by the logistics enterprises on its own.

These thesis can be derived from the following basic consideration: A part of the small and medium-sized logistics enterprises is exposed to an *cooperation pressure* or rather an *cooperation exigency* in the logistics market to survive in the long term. A part of these logistics enterprises has a lack of ability to cooperate with other logistics companies because of insufficient in-house software products.

Some of these logistics enterprises would be able to cooperate at lower costs and in a shorter time by using multitenant cooperation-enabling logistics cloud services. However, a part of these logistics enterprises does not know about the advantaged and capabilities of logistics cloud services. They need an external suggestion to get know about cloud computing and an external support for the implementation of logistics cloud services in their own companies.

A part of the existing logistics associations in Germany and of course in other countries could provide this external support. In this case the logistics associations cloud get the role of logistics cloud service intermediaries between the logistics enterprises as cloud service users and the cloud service providers. However, the problem is that the logistics associations themselves need an adequate approach for the introducing of logistics cloud services together with the logistics companies. At this point, the lack of a scientifically based method is apparent which represents a such adequate approach.

The method described in this paper was developed using the paradigm of design-oriented business informatics. For this purpose, several phases have been processed: 1. analysis phase, 2. design phase (draft of the method), 3. evaluation phase (by means of expert interviews with cluster managers of logistics associations), 4. revision phase (correction of the method design) and 5. diffusion phase (is currently taking place, for example by writing of this paper). In the following sections the results of each phase are described.

II. State of Cooperation Capabilities of small and medium-sized Logistics Service providers

Logistics SMEs often lack the willingness or ability for cooperation and as a consequence of this they have to put up with a competitive disadvantage [2]. Schwinger and Wäscher analysed the cooperative skills of 133 SMEs in the logistics industry focusing on technical, personnel and organisational aspects of cooperation [3]. They found that 94 of the 133 logistics SMEs studied were willing to cooperate and would like to work with other companies in the industry better and more frequently whereas only 14 of the 94 companies willing to cooperate had very good cooperative skills. The two main reasons for the lack of cooperative skills are the lack of confidence in other logistics companies (social/personnel cooperative skills) and the lack of availability of IT applications capable of cooperation (technical cooperative skills) [3].

The relative importance of both obstacles to cooperation can be diminished or essentially eliminated in the long term by involving the logistics SMEs in logistics associations. Regional logistics associations can act as a suitable platforms on which social relations can grow and trust among logistics SMEs can be strengthened. Besides, these platforms can act as cloud enablers and improve the technical cooperative skills of the logistics SMEs by introducing logistics cloud services. These two aspects will be discussed in the next section.

III. State of the Art regarding Logistics Clusters and collaboration Types of medium-sized Logistics Service Providers

A. *State of the Development of Logistics Clusters*

Numerous statistical surveys have shown that the logistics industry in Germany is dominated by small and medium-size companies; they account for approximately 97 per cent of the total number of logistics companies in this country [4], [5]. The overwhelming majority of logistics SMEs are voluntary members at least of a regional, state-wide or nationwide logistics association [6]. Reference [6] shows a quantitative survey of about 275 logistics associations in Germany. Logistics association, for the purpose of this study, is defined as follows:

“A logistics association is any form of non-contractual cooperation of several logistics companies by which the logistics companies involved try to attain an economic advantage. The logistics association can be institutionalized as a legal entity in which case a regional logistics association forms the cluster organization of a regional logistics cluster.”

The logistics association can be considered as ”precursor“ of a higher developed, more complex and contract-based form of cooperation among logistics companies that emerges from the logistics association in the course of time. The logistics association can be viewed as umbrella organization under which the logistics companies develop ties of cooperation.

Starting in 2000, the development of regional logistics cluster organizations driven by regional economic actors gathered momentum[7]. On the basis of the quantitative survey by Oberländer, 64 regional logistics associations were defined in Germany that can be referred to as regional logistics cluster organizations.

B. *Regional Logistics Associations as “Technology rollout Networks” for Logistics Cloud Services*

The study by [6] identified the main tasks of the 64 regional logistics associations found. These include the marketing of the logistics companies of the regional logistics association (52 of 64), the exchange of information and know-how among the member companies (44 of 64), the recruiting of personnel (36 of 64) and location marketing (33 of 64) [6].

IT-related aspects of the regional logistics associations are of significance to this publication. Totally 22 of 64 regional logistics associations are active in the fields of innovation, IT, technique, and R&D projects; 8 of 64 provide online services, web services or Software-as-a-Service (SaaS) applications to the member logistics companies. Cloud computing is gaining in importance among the logistics associations, which was confirmed by the interviews conducted (see VI). Reference [8] names ... the identification of future trends and the utilization of new technologies particularly for SMEs as the main task of technology foresight. SMEs often lack the resources and access to the required information. It would be an advantage for many similar companies in an industry to cooperate in the field of technology elucidation so that available resources can be bundled and joint innovation projects in cloud computing and logistics cloud services initiated. Reference [9] shows that many SMEs make no use of cloud computing because the lack the competence and the capacity of taking advantage of these potentials. A suggested solution would be a cloud enabler and trust builder for SMEs in order to unlock the potentials of that technology.

These roles can be played by regional logistics associations in that they act as technology elucidators for logistics companies as far as cloud services are concerned.

C. *Bottom-up Development of Fourth-Party-Logistics Providers based on regional Logistics Associations*

Several different versions of the emergence of fourth party logistics (4PL) providers (or similar types of logistics integrators) are known [10]:

i. an original equipment manufacturer (OEM) or another large industrial / commercial company forms a subsidiary company as 4PL provider or assigns the tasks of a 4PL provider to an existing subsidiary company;

ii. a consulting firm takes over the role as 4PL provider for and on behalf of the value-added partners;

iii. an IT service provider takes over the role as 4PL provider for and on behalf of the value-added partners;

iv. a third party logistics (3PL) provider revises its business model and thereby becomes a 4PL provider;

v. the value-added partners (e.g., industrial or commercial company and logistics provider) form a cooperation, e.g., a joint-venture company, which takes on the role as 4PL provider;

vi. several logistics SMEs join to form a 4PL provider.

The versions (i. to v.) can be regarded as top-down development approaches or a mixture of top-down and bottom-up approaches. The combination of small and medium-size logistics companies fuels the interest in the bottom-up development of 4PL providers.

The emergence of regional logistics associations marks a first step in that direction. Some logistics companies are expressly formed with the aim of providing logistics SMEs with a platform on which they can cooperate. Reference [6] shows that ... about 20 % of all regional logistics associations systematically deal with the mediation of logistics orders to the member companies or manage the handling of large logistics orders by several member companies. So some of the regional logistics associations already act as 4PL providers, for which logistics cloud services have hardly been used or are only in the nascent state.

IV. State of the Art regarding Cloud Services, Software and Platform for Logistics Service Providers

A. *Application Area of Logistics Cloud Services*

The german “Bundesverband Informationswirtschaft, Telekommunikation und neue Medien” (BITKOM) and other institutions/authors describe different scenarios for the redesign of a company’s internal IT landscape due to the use of cloud services. The scenarios of using cloud services most frequently cited include the following [11]:

i. Higher flexibility of software use by the faster and simpler provision of software applications via the internet (on-demand) or outsourcing of software applications so far used internally, e.g., order management, enterprise ressource planning (ERP), supply chain management (SCM) and customer relationship management (CRM) systems;

ii. Improvement of the cost variability of software use by more flexible accounting models (pay-per-use)

iii. Further development and augmentation of the internal business model by the ability of offering complementary IT services to available company services (e.g., easier IT integration of customers and value-added partners by linking them to the cloud service);

iv. Protection of critical business processes, business applications and business data by shifting them to a cloud environment with very high security requirements.

Many of the application scenarios of cloud services discussed so far aim at providing internal/company-focused advantages of cloud usage. In contrast with that, explicitly cooperation focused application scenarios are discussed much less frequently today. The benefits of the multitenant capability of SaaS applications (e.g., cost sharing, risk sharing, internet-based access, and constant availability) are in the focus of attention whereas multitenant cooperation by means of SaaS applications has received less attention. The joint use of multitenant cooperation logistics cloud services could mean that logistics SMEs willing to cooperate develop a higher degree of ability to work together than with the applications they have been using so far.

Unlike local integration of all with all applications (n:m, “mesh *networking”) among all network partners,* the investment in IT for integration and interface development in the case of a central SaaS application with shared use is limited to the number of applications needing integration (1:n, “hub-and-spoke network”).

B. *Cloud Collaboration Platforms for Logistics Service Providers*

Major developments of cloud collaboration platforms for logistics service providers are undertaken, for example, by the EffizienzCluster Logistik Ruhr e.V. the Logistics Mall of the Fraunhofer Institute for Material Flow and Logistics (IML) [12] as well as at the Netzwerk Logistik Leipzig-Halle e.V. [16] under the European project LOGICAL [13], [14], [15]. Some logistics software vendors have recognized the trend towards cooperation-supporting cloud-based logistics software and reorient their software development in that direction [17].

The technique-driven approaches to the development of cooperation platforms and cooperation-enabling logistics cloud services require an application-related and practical procedure to be able to tap the potentials of novel logistics cloud services for logistics SMEs. The best logistics cloud platforms and applications will not be used if the logistics companies are unaware of the advantages of cloud applications and their possible use for improving the cooperative skills. The regional logistics association need structured directions for activities to take all actors and aspects of cloud-based cooperation among logistics SMEs into consideration. The fundamentals of such a method will be described in the next section.

V. Description of a method for Cloud-based Collaboration of Logistics Service Providers

A. *Meta Model*

The meta model of the method consists of the meta model elements *result*, *technique*, *activity* and *actor/role*, which are described in greater detail in the respective submodels.

Comprised of these four elements is the method in which structured sequences of activities by several actors or roles are performed that generate results and outputs by the application of techniques. The submodels developed by the author and evaluated in expert interviews will be described briefly in the next sections. The interview results will be discussed in detail later. The meta model is shown in Fig. 1. The description of the method in the form of models was performed with the aim to describe the method in a formal way. It is not provided to instance the models using BPML or other instantiation techniques. For the target group of logistics SMEs (as method users), it is important that the method and its components are easy to understand and easy to follow. By subsequent research and development activities there could be an instantiation of the different method models to obtain a software support of the method. This could be termed as cloud-based "computer aided cooperation development" (CACD). But this goes beyond the considerations in this paper.

B. *Role Model*

The role model comprises several roles on the higher level of legal entities (logistics service provider, logistics associations, logistics customers, etc.), as well as concrete roles within these legal entities (chief executive officers, commercial managers, cloud software endusers, cloud soucing project manager, logistics deciders etc.). Fig 2 shows the role model.

C. *Process Model for Logistics Collaboration Development based on Logistics Cloud Services*

The approach model comprises of totally six main activities each with three or four intermediate activities which, in turn, are divided into sub-tasks. Fig. 3 shows the main activities of the process model using the business process model and notation (BPMN). For reasons of space, only the main and intermediate activities can be presented here in order to provide an overview of the process model. The process model combines, in its activities, a multitude of procedures, techniques and methods of other engineering disciplines, e.g. business engineering, systems, engineering, software engineering, service engineering, integration engineering, and requirements engineering.

The assignment of roles, results and techniques to the activities cannot be detailed here for space reasons and is described in detail in reference [6].

D. *Result/Output Model and Technique Model*

The techniques and results of all activities and their subactivities are presented in [6]. For example, the techniques include presentation and workshop techniques to describe the possible application of logistics cloud services by the project manager or cloud service provider to decision-makers in the logistics association and the logistics companies.

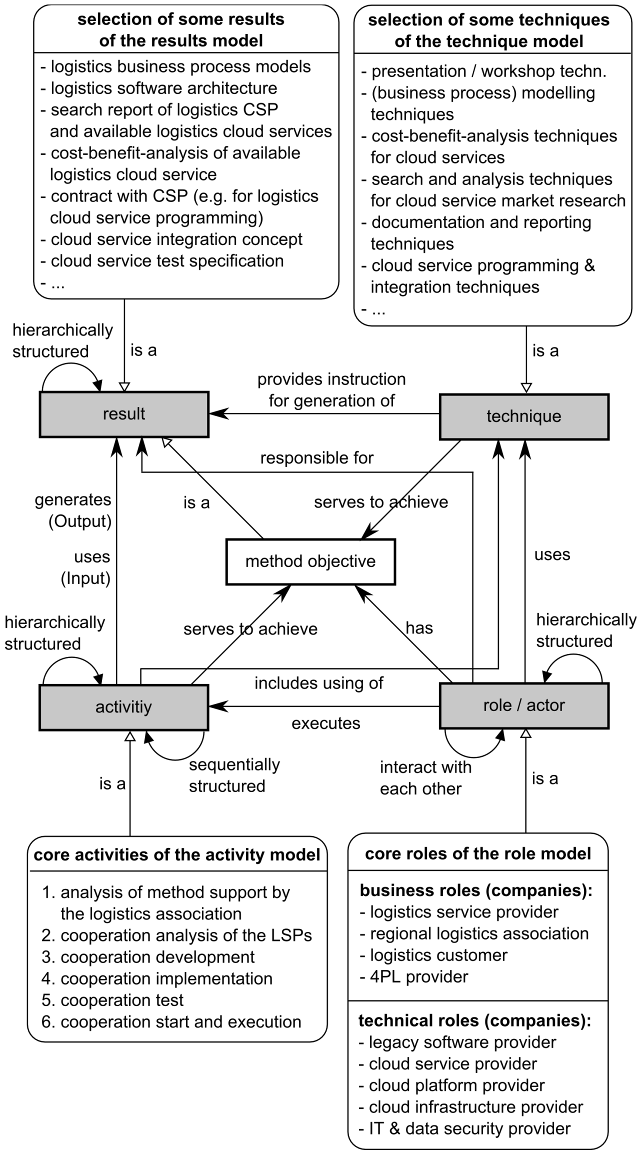

selection of some results of the results model
- logistics business process models
- logistics software architecture
- search report of logistics CSP and available logistics cloud services
- cost-benefit-analysis of available logistics cloud service
- contract with CSP (e.g. for logistics cloud service programming)
- cloud service integration concept
- cloud service test specification
- ...
selection of some techniques of the technique model
- presentation / workshop techn.
- (business process) modelling techniques
- cost-benefit-analysis techniques for cloud services
- search and analysis techniques for cloud service market research
- documentation and reporting techniques
- cloud service programming & integration techniques
- ...
hierarchically structured
is a
is a
provides instruction for generation of
result
technique
responsible for
generates (Output)
is a
serves to achieve
uses (Input)
method objective
uses
hierarchically structured
serves to achieve
has
hierarchically structured
includes using of
activitiy
role / actor
executes
sequentially structured
interact with each other
is a
is a
core activities of the activity model
1. analysis of method support by the logistics association
2. cooperation analysis of the LSPs
3. cooperation development
4. cooperation implementation
5. cooperation test
6. cooperation start and execution
core roles of the role model
business roles (companies):
- logistics service provider
- regional logistics association
- logistics customer
- 4PL provider
technical roles (companies):
- legacy software provider
- cloud service provider
- cloud platform provider
- cloud infrastructure provider
- IT & data security provider

Fig 1: Meta Model of cloud-based Logistics Cooperation Development

The result and output model contains the final results of each activity that should be attained by the respective activity. These include, for example, business process models, search and market analysis results, IT system descriptions, target and performance specifications, minutes of meeting, results of cost-benefit calculations, etc.

VI. Interview Findings in german Logistics Clusters with regard to the Method

The four submodels were evaluated and reviewed in totally seven expert interviews with board members or cluster managers of seven German regional logistics associations in spring 2014. These interviews were opportunities for a critical assessment of the method by potential users of it. The author received a direct feedback from the logistics industry which enabled the better evaluation and improvement of the four submodels. Individual face-to-face interviews were se-

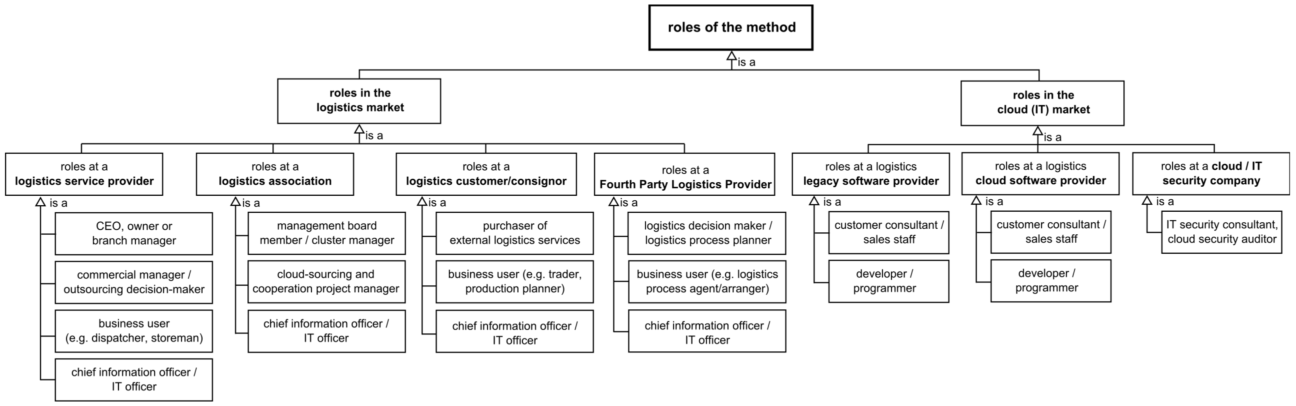

roles of the method
is a
roles in the logistics market
roles in the cloud (IT) market
is a
is a
roles at a logistics service provider
roles at a logistics association
roles at a logistics customer/consignor
roles at a Fourth Party Logistics Provider
roles at a logistics legacy software provider
roles at a logistics cloud software provider
roles at a cloud / IT security company
is a
CEO, owner or branch manager
commercial manager / outsourcing decision-maker
business user (e.g. dispatcher, storeman)
chief information officer / IT officer
management board member / cluster manager
cloud-sourcing and cooperation project manager
chief information officer / IT officer
purchaser of external logistics services
business user (e.g. trader, production planner)
chief information officer / IT officer
logistics decision maker / logistics process planner
business user (e.g. logistics process agent/arranger)
chief information officer / IT officer
customer consultant / sales staff
developer / programmer
customer consultant / sales staff
developer / programmer
IT security consultant, cloud security auditor

Fig 2: Role Model of cloud-based Logistics Cooperation Development

lected as evaluation method to give comprehensive explanations to the interviewee about the overall context and the objective of the approach. During the interviews, the models were extensively explained using the model charts and figures and revised using handwritten sketches. The complexity of the models is too large to get appropriate feedback exclusively by an quantitatively designed online questionnaire or another quantitative survey technique.

The results of the interviews will be discussed summarily in the following sections.

A. *Classification of the Interviewees*

The following persons were interviewed as representatives of their respective logistics associations. All interviewee hold managerial positions in their logistics associations. The number of member companies and the year in which the association was formed are added in brackets:

i. Prof. Dr. Uwe Arnold, Logistics Cluster Manager of the Network Logistics Leipzig-Halle e.V. (NLLH); 130 member companies, founded in 2008
ii. Mark Renner, Logistics Cluster Manager of LogistikNetz Berlin-Brandenburg e.V. (LNBB), 70 member companies, founded in 2007
iii. Michael Kluger, Logistics Manager at the House of Logistics and Mobility e.V. & HOLM GmbH & LogisticsCluster RheinMain, 220 member companies, founded in 2007
iv. Dr. Robert Schönberger, CEO of the Logistics Cluster Schwaben e.V. (LCS), 90 member companies, founded in 2011
v. Björn Geib, Head of Innovation at Logistik-Initiative Hamburg e.V. (LIHH), 530 member companies, founded in 2000
vi. Christian Prasse and Andreas Nettsträter, Effizienz-Cluster Logistik Ruhr e.V. (ECLR) / EffizienzCluster Logistik GmbH / Fraunhofer IML Dortmund, 170 member companies, founded in 2010
vii. Dr. Christian Huck, board member of the Logistics Network Thuringia e.V. (LNT), 40 member companies, founded in 2008

The interviews were conducted on the basis of a uniform questionnaire with entirely open questions and lasted for 90 to 160 minutes. The interview contained questions on the following subjects:

i. Problems with regard to the cooperative skills of the logistics SMEs member of the logistics association (see II.)
ii. Characterisation of the logistics association and its organisational structure (see III.)
iii. Activities and projects of the logistics association in cloud computing (see IV.)
iv. Feedback as to the quality of the method (see V.)

The results of the interviews will be described briefly below. Reference [6] contains the complete interview results.

B. *Interview Findings regarding the Cooperation Aspects*

All interviewees confirmed that the logistics SMEs are exposed to extrinsic cooperative pressure and the need for cooperation was high in view of the requirements of the logistics market.

Some logistics companies therefore decide to join a logistics association; small and medium-sized logistics companies tend to join a regional logistics association whereas large logistics companies and groups prefer membership in national or state-wide logistics associations. The spatial and social closeness among the logistics SMEs within a regional logistics association tends to improve their spirit or readiness for cooperation, with one of the major obstacles to cooperation, i.e., lack of confidence in potential cooperation partners, being alleviated in part.

All interviewees also confirmed that the cooperative skills of the logistics SMEs were not unlimited but still running significant deficits as a result of, for example, poor IT equipment (such as incompatible software applications or poor interfaces).

Thus, those interviewed confirmed the hypothesis of limited cooperative skills due to the lack of IT capabilities among small and medium-size logistics companies.

C. *Interview Findings regarding Cloud Computing and the Concept of Bottom-up-based 4PL Providing*

For the time being, cloud computing receives intensive attention only by some logistics associations. Only 2 of 7 interviewees said that projects in cloud computing had been

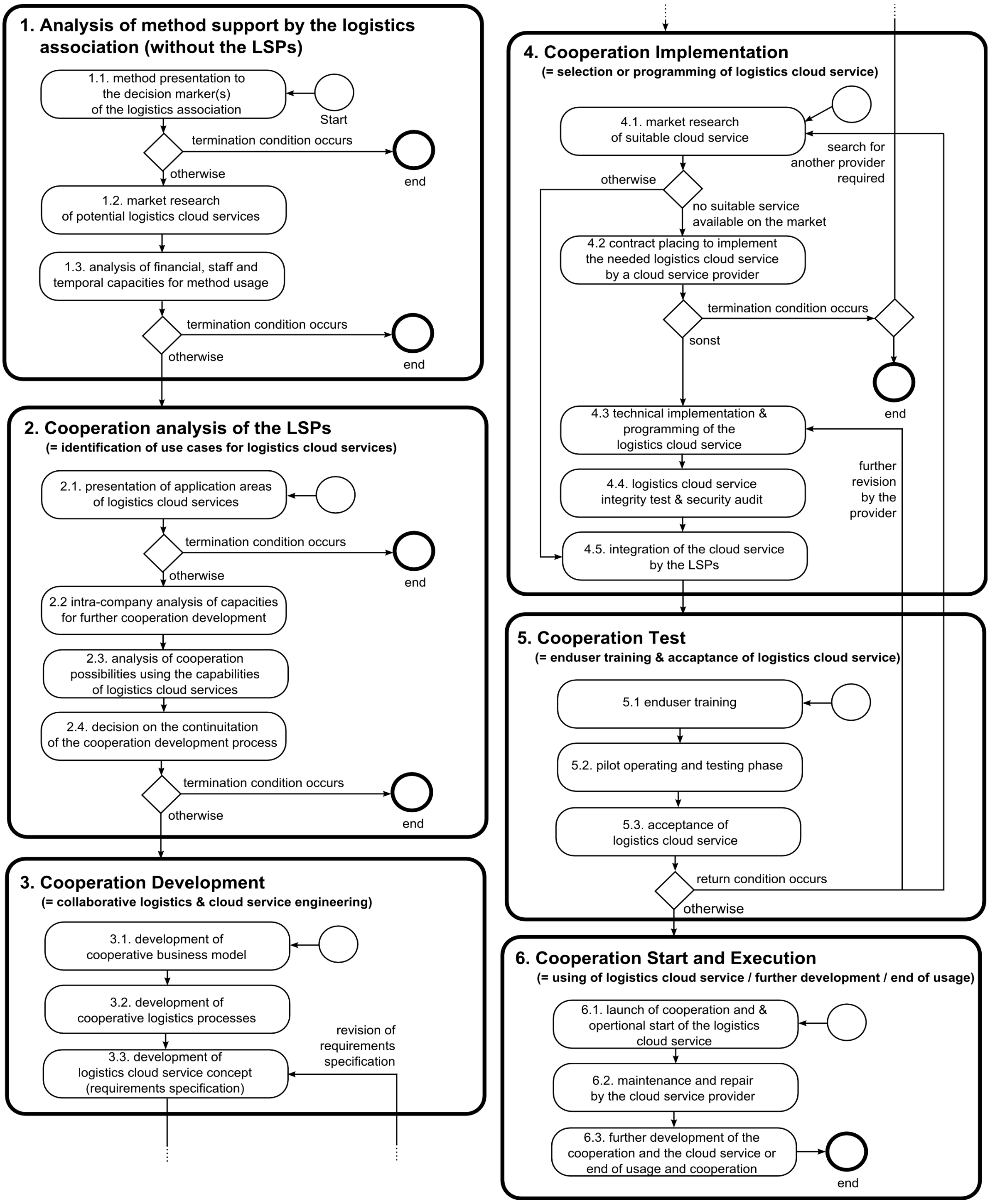

1. Analysis of method support by the logistics association (without the LSPs)
1.1. method presentation to the decision marker(s) of the logistics association
Start
termination condition occurs
end
otherwise
1.2. market research of potential logistics cloud services
1.3. analysis of financial, staff and temporal capacities for method usage
termination condition occurs
end
otherwise
2. Cooperation analysis of the LSPs
(= identification of use cases for logistics cloud services)
2.1. presentation of application areas of logistics cloud services
termination condition occurs
end
otherwise
2.2 intra-company analysis of capacities for further cooperation development
2.3. analysis of cooperation possibilities using the capabilities of logistics cloud services
2.4. decision on the continuitation of the cooperation development process
termination condition occurs
end
otherwise
3. Cooperation Development
(= collaborative logistics & cloud service engineering)
3.1. development of cooperative business model
3.2. development of cooperative logistics processes
3.3. development of logistics cloud service concept (requirements specification)
revision of requirements specification
4. Cooperation Implementation
(= selection or programming of logistics cloud service)
4.1. market research of suitable cloud service
search for another provider required
otherwise
no suitable service available on the market
4.2 contract placing to implement the needed logistics cloud service by a cloud service provider
termination condition occurs
sonst
end
4.3 technical implementation & programming of the logistics cloud service
further revision by the provider
4.4. logistics cloud service integrity test & security audit
4.5. integration of the cloud service by the LSPs
5. Cooperation Test
(= enduser training & accaptance of logistics cloud service)
5.1 enduser training
5.2. pilot operating and testing phase
5.3. acceptance of logistics cloud service
return condition occurs
otherwise
6. Cooperation Start and Execution
(= using of logistics cloud service / further development / end of usage)
6.1. launch of cooperation and & opertional start of the logistics cloud service
6.2. maintenance and repair by the cloud service provider
6.3. further development of the cooperation and the cloud service or end of usage and cooperation
end

Fig 3: Process Model of cloud-based Logistics Cooperation Development

initiated and the developments in this innovative field were followed and promoted actively. Another 2 interviewees at least accompanied workshops, lectures and other events on cloud computing, for example, to present the solutions of logistics cloud service providers and discuss them with the logistics companies. The 3 other interviewees had not so far dealt with cloud computing but were planning to do so in future.

The concept of 4PL providers was known to 4 of 7 interviewees and as such does not represent is common knowledge of potential forms of cooperation of logistics companies in the logistics industry.

Only 2 of 7 interviewees ever thought of developing a 4PL provider with regional activities within or parallel with the available regional logistics association, however none of 7 logistics associations interviewed actually set it up. After the theoretical concept had been explained, 5 of 7 interviewees said it was difficult to imagine at present but not unrealistic in the long term.

D. *Interview Findings regarding the Method*

To explain a potential solution to the current lack of technical cooperative skill among logistics SMEs, the possible application scenarios for logistics cloud services were explained and the method described with reference to the role model, procedure model, technique model and result model in the next step. These explanations took about 30 minutes. After that, those interviewed were asked whether the four submodels were complete, understandable and of practical relevance (applicable). The interviewees provided a very comprehensive and critical feedback and added further aspects to all four models, e.g.:

i. New roles (e.g., IT security advisor / auditor to verify the technical security of the logistics cloud service),
ii. New activities (acquisition of subsidies to support innovation for the introduction of cloud services),
iii. New techniques (e.g., a forecast calculation to assess the long-term benefits associated with the use of the logistics cloud services), and
iv. New result types (logistics cloud platform model, where several logistics cloud services are available to the members of a logistics association on a central platform.)

Reference to the additional roles and activities was already made in section V. The principal question of the discussion of the submodels was that of the possible application of the method in the regional logistics association each of the interviewees represented.

Totally 5 of 7 interviewees confirmed that the method could be applied to provide the associated logistics SMEs with a new technological roadmap for developing cooperation. 2 of 7 interviewees said the method was correct and made sense but their logistics association was not pursuing activities towards developing cooperation among members. However, these two interviewees were agreed that they might apply the method if the support of the development of cooperation was the task of the logistics association.

VII. Conclusions

The quantitative analysis of regional logistics associations in Germany has shown that the support of cooperation and IT/innovation are direct activities in which numerous associations engage but cloud computing and logistics cloud services are not sufficiently known yet.

The qualitative analysis of selected regional logistics association was able to show that the majority of the logistics SMEs is exposed to pressure for cooperation but only some logistics companies have the required cooperative skills due to the general lack of access to light-weight and low-priced software applications facilitating cooperation.

The investigation of logistics associations as cloud-enabler has not been carried out in Germany before and is described in this paper for the first time. To develop the method, the cloud migration approaches of other authors were used and applied to the special circumstances of logistics SMEs and the innovation potentials of logistics associations. Quantitative indicators for the usefullness of logistics cloud services especially for logistics SMEs were given by the detailed, quantitative company surveys of the Fraunhofer IML in Dortmund [20].

The overwhelmingly positive assessment of the method by those interviewed is encouraging and underlines the need for an application-oriented and comprehensive guideline of action for regional logistics association to provide cooperation support for their members by applying multitenant cooperation-enabling logistics cloud services.

VIII. Implications for further Research and Development

The development of cloud services for the logistics industry is making rapid progress. At the same time as cloud service providers develop cloud services, cloud platforms comprising the cloud services of several individual providers emerge. The same development can also be seen in the logistics sector, where the number of logistics cloud platforms is increasing. Several logistics association have developed their own logistics cloud platform strategy and try to offer their members in the logistics sector new technical opportunities of developing cooperative logistics business models. Areas in which further research is needed and new questions must be answered include the following:

i. The further development of multitenant logistics cloud services to multitenant cooperation logistics cloud services by the respective providers;
ii. The development of business models and legal frameworks of logistics cloud platform providers [18];
iii. Ensuring the integrability of different logistics cloud services of the same or different cloud service platforms on technical and semantic levels;
iv. Ensuring data protection, data security, confidentiality and integrity in cloud environments and of cloud service based business processes [19];
v. The development of transnational logistics cloud platforms because logistics processes have no regard for national borders;
vi. The integration of public (e.g., customs authorities) and quasi-public actors (e.g., seaports, airports, freight distribution centres, etc.) in the logistics industry in the logistics cloud environments (cloud-based logistics e-government)

The use of logistics cloud services in future could enable logistics SMEs to combine and form 4PL providers in the same way as bottom-up logistics cloud platforms are composed of the logistics cloud services of individual providers.

References

[1] T. Redlich: “Value creation in bottom-up economics” (Wertschöpfung in der Bottom-up-Ökonomie), Springer-Verlag, Berlin, 2011, pp. 1-5, DOI: 10.1007/978-3-642-19880-9.

[2] U. Arnold, J. Oberländer, S. Mutke: “Milestone Report 1.1 about Requirements and Test Szenarios for Logistics Services”, BMBF-Project InterLogGrid, Logistics Community of D-Grid Initiative, 2010.

[3] D. Schwinger, G. Wäscher: “Maturity for virtual companies” (Reif für das virtuelle Unternehmen?), in: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 58. Year, vol. 3, Verlagsgruppe Handelsblatt, 2006, pp. 390-424, DNB: http://d-nb.info/96784391X.

[4] BGL: Bundesverband Güterkraftverkehr, Logistik und Entsorgung e.V.: “Statistics about road haulage in Germany” (Statistiken zum gewerblichen Güterkraftverkehr in Deutschland), 2014. [Online] Available: http://www.bgl-ev.de/web/daten/index.htm.

[5] BAG: Bundesamt für Güterverkehr: “Structure of companies for german road haulage” (Struktur der Unternehmen des gewerblichen Güterkraftverkehr in Deutschland), Köln, 2012. [Online] Available: http://www.bag.bund.de/SharedDocs/Downloads/DE/Statistik/Unternehmen/Ustat/Ustat_2010.html.

[6] J. Oberländer: “Method for Development of web-based Cooperations of small and medium-sized Logistics Service Providers” (unpublished doctoral thesis, submission midyear 2014).

[7] W. Elsner, J. A. Hübscher, M. Zachzial: “Regional Logistics Clusters – Statistical survey, strengths and weaknesses, action potentials (Regionale Logistik-Cluster: Statische Erfassung, Stärken und Schwächen, Handlungspotenziale), Peter Lang Europäischer Verlag der Wissenschaften, Frankfurt/Main, 2005, pp. 3-21, DNB: http://d-nb.info/976630370.

[8] C. Mieke: “Technology introduction in company networks” (Technologiefrühaufklärung in Netzwerken), Deutscher Universitäts-Verlag, Wiesbaden, 2006, DOI: 10.1007/3-8350-5715-4.

[9] T. Haselmann: “Cloud services for small and medium-sized companies – benefits, approach and costs” (Cloud-Services in KMU – Nutzen, Vorgehen und Kosten), Verlagshaus Monsenstein und Vannerdat, Münster, 2012, pp. 263, DNB: http://d-nb.info/1024981452.

[10] F. Kasiske: “Road to the Management of the Supply Chain” (Wege zum Management der Supply Chain) in: H. Baumgarten, I.-L. Darkow, H. Zadek: “Supply Chain Management and Services – Management of Global Supply Chains by Logistics Service Providers”, Springer-Verlag, Berlin, 2004, pp. 151-156, DNB: http://d-nb.info/970085117.

[11] BITKOM: Cloud Computing: Evolution of Technology, Revolution of Business, 2009. [Online] Available: http://www.bitkom.org/files/documents/BITKOM-Leitfaden-CloudComputing_Web.pdf

[12] D. Daniluk, J. Rahn, M.-B. Wolf: “Logistics Mall – cloud-based logistics software” (Logistics Mall – Logistiksoftware aus der Cloud) in: Journal of Business Informatics and Management, vol. 1, Springer-Verlag, Berlin, 2014, pp. 28-37, DNB: http://d-nb.info/992716098.

[13] U. Arnold: “Cloud logistics – the user perspective” (Cloud Logistics – die Anwenderperspektive) in Journal of Business Informatics and Management, vol. 1, Springer-Verlag, Berlin, 2014, pp. 16-27, DNB: http://d-nb.info/992716098 .

[14] U. Arnold, J. Oberländer, B. Schwarzbach: “LOGICAL - Development of Cloud Computing Platforms and Tools for Logistics Hubs and Communities”. In: Proceedings of the Federated Conference on Computer Science and Information Systems, Wrocław, pp. 1083–1090 (2012), ISBN: 978-1-4673-0708-6.

[15] U. Arnold, J. Oberländer, B. Schwarzbach: “Advancements for Cloud Computing in Logistics”. In: Proceedings of the Federated Conference on Computer Science and Information Systems, Wrocław, pp. 1055-1062 (2013), ISBN: 978-1-4673-4471-5.

[16] U. Arnold, J. Oberländer: “Annual Report 2013 of Logistics Network Leipzig-Halle e.V.” (Jahresbericht 2013 des Netzwerk Logistik Leipzig-Halle e.V.), 2014 [Online] Available: http://www.logistik-leipzig-halle.net/uploads/tx_abdownloads/files/Jahresbericht_NLLH_2013_web.pdf, pp. 20-23, (2014).

[17] G. Teichmann: “Cloud-based collaboration in logistics” (Cloud-basierte Kollaborationen in der Logistik) in Journal of Business Informatics and Management, vol. 1, Springer-Verlag, Berlin, 2014, pp. 56-65, DNB: http://d-nb.info/992716098.

[18] A. Kawa, M. Ratajczak-Mrozek: “Cooperation between Logistics Service Providers Based on Cloud Computing” . In: Intelligent Information and Database Systems , Kuala Lumpur, pp. 458-467 (2013), DOI: 10.1007/978-3-642-36543-0_47.

[19] BMBF project : “Cloud Computing: efficient cooperation without security risks”, 2014 [Online] Available: http://www.vdivde-it.de/KIS/sichere-ikt/sicheres-cloud-computing/prestige.

[20] M. ten Hompel (Ed.): “Cloud Computing for Logistics 2” (Cloud Computing für Logistik 2), Fraunhofer Verlag, Stuttgart, 2013, DNB: http://d-nb.info/1041945531.

Position papers of the 2014 Federated Conference on
Computer Science and Information Systems pp. 245–249

DOI: 10.15439/2014F472
ACSIS, Vol. 3

Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management

Silva Robak
Uniwersytet Zielonogórski, ul.
prof. Z. Szafrana 4a, 65-516
Zielona Góra, Poland
Email: s.robak@wmie.uz.zgora.pl

Bogdan Franczyk
Uniwersytet Ekonomiczny we
Wrocławiu, ul. Komandorska
118/120, 53-345 Wroclaw,
Universität Leipzig, Germany
Email: franczyk@wifa.uni-leipzig.de,
bogdan.franczyk@ue.wroc.pl

Marcin Robak
XLogics Sp. z o.o., ul.
Kostrzyńska 4, 65-127 Zielona
Góra, Poland
Email: m.robak@xlogics.eu,
Uniwersytet Zielonogórski,WEIiT,
m.robak@weit.uz.zgora.pl

Abstract—**One of contemporary big challenges in information systems are the issues associated with coping with and utilization of the vast amounts of data. In this position paper we present few research problems emerging in association with occurrence of big data, with focus on domain of logistics and supply chains. We also reveal a number of influence factors on supply chains design and management which are related to big data. We show that the association of data science and predictive analysis with domain knowledge can build an important improvement for logistics and supply chain design and management. We point out how the domain problems in the key managerial business components could be solved better with big data applications for the stakeholders involved in the supply chains. We also highlight open problems in this domain and chances for the future.**

I. Introduction

IN THE contemporary world logistics companies have to face unprecedented challenges. As a result of globalization the amount of data arising in supply chains is raising, the competition is becoming fiercer and the customers often expect integrated services, what requires a close cooperation between involved organizations. The companies have to adapt to new business models and rethink their role and position in their value chain regarding the potential possibilities given by the utilization of big data to add value for their customers and suppliers. This requires changes from logistic companies in their way of thinking about the supply chain design and management, and at the same time in their information technology view to support the collaborative decision making.

The problem of the appropriate information technology environments for collaborative processes between business participants is still present, since appropriate IT infrastructures for utilization of big data are needed. There are data 'siloes' from diverse applications like ERP or customer relationship management systems (CRM), etc. and the collaboration between business partners may require undertaking steps for IT environment integration, such as one of the known enterprise application integration solutions or usage of the Web services [1].

In our previous papers we approached the problems of advantageous utilization of vast amounts of big data in supply chains and also the information integration issues in order to overcome the data silo problem. The proposal of possible utilization of the Linked Data [2] as an integration solution for business process management BPM in supply chains networks we have already presented in [3]. In [4] we investigated the appropriate IT architectures for big data used in association of cloud computing facilities [5] and the utilization of common (open stated) data format as it is offered by Linked Data for data silos integration purposes. In this position paper we recommend the utilization of big data in conjunction with Data Science and Predictive Analysis, as appropriate for logistics with emphasis on the value-adding partnerships in supply chains.

A supply chain is defined as a network that comprehends all the organizations and activities associated with the flow and transformation of goods, starting from raw material stage through the whole process, to the end user, as well as the associated information flow [6].

In the inter-organizational information systems, which link companies to their suppliers, distributors and customers, a movement of information through electronic links takes place across organizational boundaries between separately owned organizations. It requires not only electronic linkage in form of basic electronic data interchange systems, but also interactions between complex cash management systems or by accessing shared technical databases. The problems with sharing and exchange and also utilization of information are viable in supply chains contexts.

A business process in a supply chain consists of one or more than one related activities that combined together respond to the need for a business action. The processing steps in a workflow might go through numerous data transformations (geographic, technological, linguistic, syntactical and semantic transformations) [6]. We will show that data science and predictive analysis may leverage the utilization of big data for business processes in supply chains in business networks. For this aim the rest of the paper is organized as follows.

In Section 2 we explain the V-model and the key characteristics of big data and its storage formats appearing in supply chains. In Section 3 we reveal the concepts of predictive analysis and data science. We examine the possible added value resulting from their application for big data in supply chain design and management. In Section 4 we point out some open research problems in logistics and supply chain design management. In the last Section of this position paper we conclude our work.

II. BIG DATA IN LOGISTICS AND SCM

A. Big Data V-Model

Big data is referred to data that goes beyond the processing capacity of the conventional database systems. In addition to the aspect that it is big (e.g. a huge number of small transactions, or continuous data streams from sensors, mobile devices etc.) it may move too fast, or does not fit the structure of traditional (i.e. relational) database architectures. Big data also may have a low value for further usage before processing it [7].

When we denote a big amount of data as "big data" it has to cover the 3V model with three basic features such as: volume, velocity and variety [7].

Volume of big data denotes its massive character, i.e. a huge amount of information involved. The big volume of data can be beneficial for the data analysis aims. It may improve the analytics models by having more cases available for forecasts and increase the number of factors to be considered in the models making them more accurate. On the other hand, the volume bears potential challenge for IT infrastructures to deal with big amounts of data, especially when taking into account its second V-feature – velocity.

The velocity in which data flows into organization or the expected response time to the data is the second V-feature of big data. Big data may arrive quickly - in (near) real-time (i.e., near-time). If data arrives too quickly the IT infrastructures may be not able to respond timely to it, or even to store all of it. Such situations may lead to data inconsistencies.

The variety of big data comes from its many potentially different data sources. Therefore big data may have diverse structures and forms, not falling into the rigid relational structures of SQL databases without loss of information. Some of data may be saved as blobs inside traditional data bases. The IT infrastructures for big data are denoted as NoSQL (i.e., "not only SQL" [8]). Examples of diverse sources and kinds of data are standard business documents, transactional records and unstructured data in form of images, recordings, HTML documents, Web pages, text and e-mail messages, streams from meters and environment sensors, GPS tracks, click streams from Web queries, social media updates, data streams from machines' communication or wearable computing sensors, etc. [4].

In accordance with authors we add the fourth V-feature for big data, which is its value. The big data value V-feature denotes the need for processing it before using it in order to make it valuable for analysis purposes ([8] and [9]).

We claim that the fourth V-feature bears a special importance for logistics and supply chain design and management.

The big scale usage of available and generated data is made possible for organizations owing to cloud computing paradigms, such as Infrastructure as a Service IaaS, Storage as a Service SaaS, which revolutionized the way the computing infrastructures are used [5].

B. Big Data Characteristics (Data structures)

In the beginning of this Section the V-model for big data with its four characteristic V-features: volume, velocity, variety and value have been considered. Based on it we now define big data as:

"data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value." [10].

The variety of data streams in logistics and SCM may be regarded as structured, semi-structured, "quasi"-structured or unstructured. The structured data contains defined data types, formats and structures. Transaction data from Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) can be named as example.

The semi-structured data is represented by textual data files with discernable patterns, what enables data parsing. The examples of semi-structured data are self-describing XML data files defined by XML Schema.

The quasi-structured are irregular data formats that can be formatted, but only with additional effort, tools and time. An example of such data could be Web click-stream data.

As the unstructured we denote data with no inherent structure that can be stored in different types of files. The fact that part of data is unstructured, or rather, it lacks a structure appropriate for storage in conventional SQL databases, implies that other solutions are needed.

The conventional IT infrastructures in supply chains include structured data in form of OLTP and OLAP systems. The traditional OLTP systems support the transactional systems with highly structured SQL databases, whereas the OLAP systems contain aggregated historical data in form of cubes. The OLTP systems deliver simple reports, while OLAP systems (Data Warehouses) are suited for (traditional) business intelligence applications with reporting facilities on business statistics, performance, etc. on the basis of structured (analytical) historical data. On the opposite of partially low value big data the OLAP and OLTP provide only high quality data.

As stated above these both databases forms are unsuitable for all big data formats. The data stored there has to have fixed structure, which is conflicting with variety of big data. Hence the analytical OLAP systems contain only historical data; they are not suitable for big data applications. Another problematic issue may be due to the velocity of big data.

As the rigid SQL data structures are insufficient for big data applications, other OLT solutions in form of ‘key-value stores’ like “NoSQL OLTP” MongoDB, AmazonDynamo or Windows Azure Table Storage can be used [11].

For storing and analyzing massive data sets the “NoSQL warehousing” open-source Apache Hadoop [11] can be applied. It is a highly distributed and fault-tolerant framework for software development with its own highly distributed HDFS file system, and a MapReduce framework for writing and executing distributed algorithms.

The conventional IT structures may encounter problems with storing variety of data and immediately reacting to it. It is because of big data amounts on (also unstructured) data arriving in near-time. It is apparent that efficient dealing with big data in logistics requires for new data architectures, analytics sandboxes and tools. The IT infrastructure for the data platform will be preferably cloud-based.

III. Predictive Analysis and Data Science

In this Section we present the special role of data science and predictive analysis in logistics and supply chain development and management.

Data Science is the application of quantitative and qualitative methods to solve the relevant problems and predict outcomes. Its importance is due to the fact that with growing amounts of data the domain knowledge (here logistics) and the generalizable extraction of knowledge from data (data analytics) cannot be separated. Thus, the desirable professionals involved in supply chain design and management should posses both kinds of skills – the analytical skills and at the same time a deep understanding of the business domain and its management.

There are several roles to be covered by the practitioner of data science like: the analysts, the data professionals and also the technology and data enablers. Data scientists in the role of data analysts should possess a deep analytical talent and also an advanced training in quantitative disciplines like mathematics, statistics and machine learning.

The data savvy professionals (such as analysts and managers) are preferably people with basic knowledge of statistics and advantageously also machine learning, who should be able to define key questions that can be answered by using advanced analytics. The technology and data enablers are data scientists providing technical expertise needed to support analytical projects; their skills sets include computer programming and database administration knowledge.

The key activities of data scientists in logistics include providing services to other stakeholders like data engineers, data analysts, business analysts and the users in a line of business. Such activities comprise the reframing business challenges as analytics challenges, design, implementation and deployment of statistical models and also the data mining techniques, especially for big data. A crucial aspect in data science is thereby included in creating insights leading to actionable recommendations to help business to gain a competitive edge [12].

For supply chain design and management an advantageous skill set in the data science includes various skills, where each may have a different importance (weight). The disciplines to be covered are such as: statistics, forecasting, optimization and discrete event simulation, applied probability, analytical mathematical modeling, finance, economics, marketing, and accounting. For all the above mentioned disciplines, due to big data utilization, the focus will be different than in a traditional approach. More important become skills in conjunction with broad awareness of many different methods in comparison to the classical point of view. For instance, for the aims of forecasting the understanding the application of qualitative and quantitative methods will become more important than the understanding of the underlying stochastic processes.

Predictive analysis is a subset of data science. It encompasses a variety of statistical techniques that enable to take advantage of the patterns found in historical and transactional data. In logistics it could help to optimize business operations, to identify business risks (and security aspects), to predict new business opportunities and to fulfill the law or regulatory requirements. The business value of predictive analysis (data science) and data mining will be higher than gained from conventional business intelligence due to optimization, predictive modeling, forecasting and analysis of vast data resources (big data).

In logistics the predictive analysis uses both quantitative and qualitative methods to estimate the past and future behavior of the flow and storage of inventory, as well as the associated costs and service levels. On the other hand, the SCM predictive analysis also uses both quantitative and qualitative methods to improve supply chain design and competitiveness by estimating past and future levels of integration of business processes among functions or companies, as well as the associated costs and service levels. The value of predictive analysis is not to be scoffed at. Together with the appropriate analytics tools it may become a decisive competitive asset.

IV. Research Problems in Logistics and SCM

The open research problems in logistics and supply chain management can be viewed from the perspective of key managerial components of business logistics and on the other hand the different category of the stakeholder. The main business functions are such as forecasting, inventory management, transportation management and transportation and human resources. The stakeholders (user) such as carrier, manufacturer, retailer may be examined how they could benefit from usage of big data all the business functions in logistics and supply chain management.

For instance, considering the forecasting, the user carrier, by relating to forecasting, could predict a out the time delivery, etc. The manufacturer would be able to make an early response to extremely negative or positive customer sentiments, etc. The inventory manager may plan capacity availability in real time, etc. For the retailer it could mean an improvement in perpetual inventory system accuracy.

The similar approach may be conducted for the type of data and the management functions: inventory, transportation, customer and supplier relationship management to identify the problems to be solved with application of predictive analysis by using big data. Further consideration to the evolution of logistics [14] due to internet-based applications can be found in [15], [16] and [17].

V. Conclusion

This position paper has provided a research perspective on contemporary problem and chances in the domain of logistics, supply chain design and management in conjunction with data science, predictive analysis and big data. We have highlighted the various aspects of big data and data management in logistics and supply chains and pointed out how to efficiently use big data across the supply chains development and management.

We have categorized the potential applications components of big data in four functional categories: forecasting, inventory management, transportation management, and human resources and proposed activities based on big data for three categories of user (carrier, manufacturer, retailer). The insights won from individual components should be integrated into one global strategy on usage of predictive analysis and data science in logistics domain.

The discussion presented in this paper highlights the fact that it is important to consider using big data and predictive analysis and data science along with particular domain knowledge concerns for achieving enhanced results throughout a supply chain life cycle.

As we stated in [4], in the past vendors had to exploit earlier period's structured data in order to analyze stockholders needs and sentiments and to increase the optimal performance and the potential business value. We have examined the nowadays common solutions for data storage options for the decision making support. With the opportunity of usage of big data and cloud computing technologies, along with raising all-embracing connectivity with involved stakeholders in supply chains networks, results in the possibility of accessing current data in near-time and in getting a near-time feedback. This bears legitimate chances for almost immediate improvement of the relationships with the supply chain's stakeholders and therefore increases the agility and ability to react in real-time to the environmental changes.

One of the advantages of predictive analysis and data science is that they may provide better insights than it would be possible with traditional business intelligence systems. Thus, the high quality decision support becomes attainable, and usage of predictive analysis and data science methods will enable extending its application in decision making.

In this position paper we have presented an introduction in big data usage in logistics and supply chains. We expect that the application of predictive analysis and data science in this domain will shift the emphasis from traditional point of view of relevant aspects in many comparative disciplines like statistics, forecasting, optimization, discrete event simulation, applied probability, data mining, etc. The main criteria will become the speed of decision making, throughput and analysis flexibility.

In the future work further aspects like economic evaluation of applying big data and linked data concepts in supply chain management should be considered, also from the point of view of management theory.

References

[1] Web Services Activity, W3C Working Group, http://www.w3.org/ws

[2] W3C LinkedData, 2011, www.w3.org/wiki/LinkedData

[3] S. Robak, B. Franczyk, and M. Robak, Applying Linked Data concepts in BPM, *Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS*. Wrocław 2012. IEEE Conference Publications, ISBN: 978-1-4673-0708-6, pp. 1105-1110.

[4] S. Robak, B. Franczyk, and M. Robak, Applying Big Data and Linked Data concepts in Supply Chain management, *Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS*. Kraków 2013. IEEE Conference Publications, pp. 1215-1221.

[5] D. Agrawal, S. Das and A. E. Abbadi, „Big data and cloud computing: current state and future opportunities". *EDBT 2011*, March 22-24, 2011, Uppsala, Sweden. ACM 978-1-4503-0528-0/11/0003, DOI http://dx.doi.org/10.1145/1951365.1951432.

[6] M. P. Papazoglou, and P. M. A Ribbes, *E-business: organizational and technical foundations, John Wiley and sons*. London 2006, pp.88-90.

[7] E. Dumbill, "*What is big data? An introduction to the big data landscape*", Strata O'Reilly, 11 January 2012, http://strata.oreilly.com/2012/01/what-is-big-data.html

[8] S. Wrobel, „*Big Data – Vorsprung durch Wissen*", Fraunhofer-Institut für Intelligente Analyse- und Informationsverarbeitungsysteme IAIS. Presentation, www.iais.fraunhofer.de, 2012.

[9] I. Mitchell and M. Wilson, "Linked Data. Connecting and exploiting big data", Fujitsu Services Limited, March 2012, www.fujitsu.com.uk.

[10] B. Franczyk, "*Data science, predictive analysis and big data in logistic, SC design und management*". Universität Leipzig, Wirtschaftswissenschaftliche Fakultät. Institut fur Wirtschaftsinformatik, Presentation 2014.

[11] The Apache Hadoop Project. http://hadoop.apache.org/core/ , 2009.

[12] La Ponsie, Maryalene, "*Data scientiscs: the hottest job you haven't heard of*" online http://jobs.aol.com/articles/2011/08/10/data-scientist-the-hottest-job-you-havent-heard-of/

[13] D. C. Luckham, *Event processing for business: organizing the real-time enterprise*. Hoboken, New Jersey: John Wiley & Sons, Inc., p.3. 2012.

[14] H. Baumgarten, *Das beste der Logistik*. Springer Verlag, Berlin 2008.

[15] S. Chopra, and P. Meindl, "*Supply chain management: strategy planning, and operation*", 3rd ed., Prentice Hall, 2007, pp.427.

[16] A. Matopoulos, and E. -M. Papadopoulou, "The evolution of logistics service providers and the role of Internet-based applications in facilitating global operations" in *Enterprise Networks and Logistics for Agile Manufacturing*, L. Wang, and S.C.L. Koh, Eds., Springer, 2007, pp. 298-304, DOI 10.1007/978-1-84996-244-5_14.

[17] R. Sethuraman and S. K. Kundharaju, "Top 7 Tips for Big data to optimize Supply Chains. 5 Februar 2013, http://risnews.edgl.com/retail-trends/Top-7-Tips-for-Utilizing-Big-Data-to-Optimize-Supply-Chains86163.

Collaborative Human-Machine Intelligence 20th Conference on Knowledge Acquisition and Management and 2nd Workshop on Artificial Intelligence for Knowledge Management

KNOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference "Knowledge Acquisition and Management" and 2nd Workshop on "Artificial Intelligence for Knowledge Management" (KAM&AI4KM'14). The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we've joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The "Artificial Intelligence for Knowledge Management" Workshop was initiated by IFIP Group TC12.6 in 2012 as the separate event during European Conference on Artificial Intelligence in Montpellier (ECAI'2012). From the beginning the workshop aims in bringing together researchers and practitioners involved in Knowledge Management using the methods and techniques of AI for building and improving all aspects of KM and of knowledge flow, among others, improvement of the innovation process. This year both teams KAM and AI4KM, have decided to join efforts in the FedCSIS framework with common challenge: "Collaborative Human-Machine Intelligence" under IFIP supporting.

The aim of this common event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

Topics

The following group topics, concerning both theory and applications, will be included (unavoidably incomplete):

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks
- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

Event Chairs

Hauke, Krzysztof, Wroclaw University of Economics, Poland

Nycz, Małgorzata, Wrocław University of Economics, Poland

Owoc, Mieczyslaw, Wroclaw University of Economics, Poland

Pondel, Maciej, Wroclaw University of Economics, Poland

Program Committee

Abramowicz, Witold, Poznan University of Economics, Poland

Andres, Frederic, National Institute of Informatics, Tokyo, Japan

Chmielarz, Witold, Warsaw University, Poland

Christozov, Dimitar, American University in Bulgaria, Bulgaria

Goluchowski, Jerzy, University of Economics in Katowice, Poland

Helfert, Markus, Dublin City University, Ireland

Jelonek, Dorota, Faculty og Management of Czestochowa University of Technology

Ligęza, Antoni, AGH University of Science and Technology, Poland

Mach-Król, Maria, University of Economics in Katowice, Poland

Matouk, Kamal, Wroclaw University of Economics

Mercier-Laurent, Eunika, IAE Lyon3, France

Sobińska, Małgorzata, Wroclaw University of Economics

DOI: 10.15439/2014F476
ACSIS, Vol. 3

Managing Intellectual Capital in Knowledge Economy

Eunika Mercier-Laurent
University Jean Moulin
Lyon3 6 cours Albert Thomas 69008 Lyon, France
Email: eunika.mercier-laurent@univ-lyon3.fr

Abstract—**The Strategic Knowledge Management considers Intellectual Capital as roots of all organizations activities. The success of organizations strongly depends on the way they manage all facets of knowledge and skills. Artificial Intelligence brought some methods and techniques for handling intellectual assets of companies, expertise management, knowledge transfer and training. This paper presents a short overview of experiences and research in the field of intellectual capital management and gives some perspective for future.**

I. Introduction

SINCE over two decades the interest for managing intangible assets, including intellectual capital has been grown. However the roots of intellectual capital go far back in the history. In XX century the term of human capital has been probably re-introduced by the economist Theodore Schultz in 1961 [1]. He considers that the investment in human capital is crucial for the economic development and the education has a key contribution. Latter the term of "intellectual capital" have been introduces to cover larger field including patents and documents. Among training professionals, Karl Erik Sveiby [2] defined an Intangible Assets Monitor to drive the management of human capital.

Probably the first effort in applying the artificial intelligence techniques to managing skills in a given situation was the application developed for French police [3]. These principles were used in larger system for managing security of the Winter Olympic Games [4]. The techniques such as case-based reasoning can be useful for matching demand and offer (looking for a job or a skill). Organizations such as OECD [5] have been involved in defining a general methodology for measuring intangible investment since 1989.

The globalization changed the game of economic development. Intellectual capital has become an important asset and its assessment and management has turned to a priority for the Knowledge Economy. The intellectual capital is among the hot topics of conversations, conferences, magazines, scientific journals, books and reports. However companies and organizations are still measuring their success in term of financial capital and ROI (return on investment).

This paper presents key references related to the evolution of human capital, gives some elements of economic and environmental context and mentions current efforts in measuring of intellectual capital. It is followed by a presentation of a method and tools to manage this wealth differently and to stimulate a reflection on the role of this capital in the Knowledge Economy and in the Innovation Ecosystems.

II State of the Art

The issue of intellectual capital is complex. It involves various fields such as management, psychology, economy, sociology, health, intellectual property rights and recently sustainable development. Intellectual capital forms the basis of the successful development of companies and countries. Such a development requires the right way of managing the intangible wealth in connection with tangible ones.

Numerous publications provide a multidisciplinary view of the subject. According to Theodore Schultz [1], in charge of economic development, the education is the most important in managing of "human capital". Another economist Gary Becker [6] considers education, training, and health as the major investments in human capital.

Leif Edvinsson [7] points out the role of intellectual capital in the modern economy.

OECD [8] highlights the role of human capital in the well-being of nations.

According to Dixon [9], training, capacity building and learning are key enabling factors for "sustainability" seen as long term ability of individuals and organizations to produce innovations as a reaction and adaptation to changes in external conditions. It is the link between opportunities, projects, addressing the real needs, and building capacity or empowerment that ensures useful learning, innovation and an economically efficient process. Training supports the development of all phases of the project lifecycle (situation analysis, forecasting, planning, implementation and evaluation /measurement of impacts). Trained persons develop skills and produce methods, information and knowledge required for the success of the project. Training, combined with the development and implementation of projects on the local level, allows: i) increasing and mobilizing human and social capital ii) developing new activities and iii) creating interactions leading to collective dynamics to the invention of new rules and standards (institutional capital) needed to integrate new activities in formal economy.

Folke et al, [10] propose to develop an "adaptive capacity". The concept has been used in biology and in the context of climate change, but applies to a much broader range of issues. Adaptive capacity developed in poor countries is extremely important to be successful in XXI century. Persons able to adapt and to solve problems using individual and collective knowledge, as well as solutions from the past that work for current challenge, is able to survive and even lead in global dynamics – Mercier-Laurent [11]. Viability theory of Aubin [12] may be useful to control the balance of the ecosystems based on human capital as engine

According to Charles Savage [13] the 5th generation of managerial methods has to consider knowledge as asset. This statement has been enhanced by Debra Amidon [14] in The Innovation Strategy for the Knowledge Economy. *To know* is the opposite of *to have* attitude cultivated in today world and focus exclusively on quick business. From education point of view the most important is to learn how and what to learn.

These few references cover a large spectrum on human and intellectual capital themselves and the roles they play in economic development and the wellbeing of the nations

III. Economic and environmental context

The current economic situation in the developed countries and intensive industrialization in Asia generate new problems and needs – among them we can mention the industrial decline and unemployment in developed countries, exodus from regions to towns and the emergency of planet protection. In search of the cheapest work cost China has become the world factory. Goods travel all around the globe, generating pollution. Asian people are also studying abroad to increase their intellectual capital and sometimes bring it back to their respective countries.

In Europe the emphasis is on education and innovation is seen as a magic wand to renew industry. Despite the recommendations of the Lisbon treaty, the impact of education and innovation is still not measures in term of job creation and economic development of the regions and countries.

The intensive industrialization from the beginning of 20th century did not taken into account the impact of these activities – Lenkowa [15], Eckholm [16] on the planet. The recent alerts points out the extreme emergency – Arthus-Bertrand [17]. The Sustainable Development and Corporate Social responsibility movements focus on the use of local resources. While companies are concerned about carbon and recently about water footprint, less about raw materials, they seem not concern by biodiversity. In reality they still do not manage the human capital; the local skills and know-how are not taken into consideration, because of the lack of holistic approach. By consequence skilled people are travelling. Despite ubiquitous information and communication technology these movements remain significant even increase.

The appropriate management of human capital and the education of knowledge cultivators will certainly bring a contribution to planet ecosystems protection.

This challenge is among the most important of the 21 century. It is vital to understand, know and manage intellectual capital in connection with others tangible and intangible assets of companies and other organizations, of cities, regions and countries using a combination of holistic and system approaches – Mercier-Laurent [11].

IV. Managing human capital

While some thinkers state that the human capital is the most important asset, only few are measuring and managing it. The most important barrier in managing intellectual capital is the ignorance. Another one is a way of thinking.

The are a plethora of various data bases, even big data, but still built using traditional information processing methods, as a very limited number codes for professions[1], taking into account only traditional ones. Pole d'Emploi (national center for employment) and other public initiatives in France are supposed to help people in finding jobs, but their efficiency is very low, because they are not using the right methods and tools.

With the quick progress of technology and artificial intelligence, computers are able to process natural language instead of codes. This open programming approach allows including in real time new professions that appear every day.

There is also a lot of valuable paper and electronic reports about intellectual capital containing key information and complex charts, very useful to know the current status, but they are static.

[1] for example all information services are coded 721Z

For someone looking for a local know-how, it is not easy to find quickly a right person. Some social networks such as Viadeo in France or LinkedIn are trying to connect talents and those who are looking for. Google is certainly among the most efficient search engines, but its business model introduces an important "noise" (and intellectual pollution) due to the advertisement system management.

Know what we have

A concept of "knowledge trees" have been introduced by Michel Autier and Pierre Levy [18] and implemented in tools as Ginko (Trivium) and Selva (Ligamen), offering graphic representations of individual and collective skills as a tree.

The Figure 1 illustrates the skills of 10 people: the trunk represents common knowledge, branches the specialties, and leaves the unique skills. Such an image provides information on everyone's ability and helps to decide if the unique skills represented by the leaves are strategic.

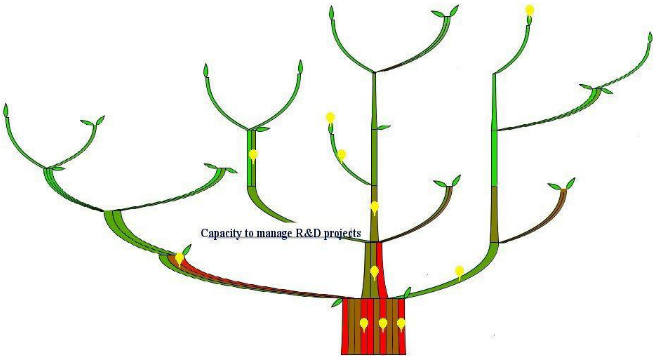


Figure 1 "Knowledge tree" created using the Ligamen software (http://www.ligamen.fr)

The part to the far right of the trunk, as well as the triple branch, indicate the position of a person in a group. Such visualization facilitates the identification of skills and helps to detect the lacks in relation to a required profile, which can be filled by training. Thus, we can build the competency tree of a company or a region and reason backwards: what projects can we achieve with such intellectual capital? We then need to search for the skills in a neighboring region or "rent" them to a partner.

In the international context and within a networked enterprise, it would be better to manage skills with a holistic perspective - on regional, national and international levels. This intangible wealth can grow through continuous learning from interaction with the environment, according to corporate strategy.

The training department is in charge of making this capital grow. It is involved for now, because in the global Knowledge Management approach all knowledge cultivators are constantly learning.

The training department could be a guardian for the transmission and preservation of the essential knowledge and know-how of retiring, especially when this is the knowledge of a long-life and is a strategic product for the company. Collaboration between several professionals facilitates the skills management.

Find the right profile

When we know what we have and what we are looking for, one of artificial intelligence techniques - case-based reasoning [19] could be very helpful. The built-in analogy engine works by matching demand (I am looking for) and offer (base of existing skills and know-how) to find instantaneously the profile we want, if such a profile is registered. If not, a set of similar profiles that could be adapted to the expected ones by training is proposed to the user. We can imagine a World Knowledge Base including Talent Bank equipped with such an engine.

The various methods of measuring the value of human capital of a company, city, region or country provide the information on what we have. But it is certainly most useful to know what can be done with for future development, what new activities and companies may be created. This purpose is illustrated in Figure 2.

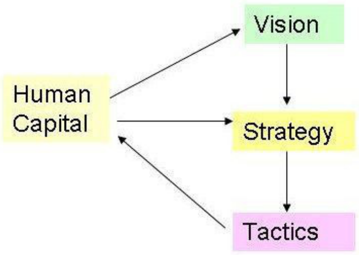


Figure 2 Human capital dynamics

A company/city/region need to elaborate a clear vision for the future. At this point the skilled persons able to envision it are needed. This vision will be "translated" into corporate strategy and a tactics (actions to achieve the strategic goals). The intellectual capital of professional working on accomplishing various tasks grown and the new knowledge and capacity should be taken into account at strategic level. It may also influence the vision.

V. SKILLS FOR THE FUTURE

Today educational system produces the traditional professionals. Many of them face the difficulty in finding job in their region or country. The most audacious travel for job; change country, language and continent. They have to adapt to new conditions and to new culture.

As mentioned before, our future depends on our capacity to adapt, to detect opportunities and to gather necessary skills and knowledge and to transform them into economic values, in balance with ecosystems. It also depends on the rapidity of our decision making, on our risk taking ability in a dynamic environment, and on our ability to use the computer, regardless of its form, as an intelligent assistant. The latter facilitates an innovation without boundaries between fields (out of the box thinking).

Facing the affluence of information and solicitations, a new skill is required – the innovation know-how. This is the art of finding and exploiting strategic information and of gathering momentum and developing the knowledge and skills essential to the success of this enterprise, which is innovation in its entirety.

These skills are numerous – from the management of ideas and people to the implementation and commercialization.
Although Europe has a long innovation tradition since the industrial era, globalization has changed the odds. The factors such as the slowing down, the obsolescence of some sectors and the emergence of others, as well as the relocation of activities, influence active knowledge and skills. The lack of interest expressed by youths in scientific studies will lead to a shortage of engineers. Some skills are disappearing with retirements, which are sometimes accelerated by the economic crisis. Knowledge capitalization approaches are saving a part of the strategic and "sensible" skills, but these initiatives are quite rare and are often initialized too late.

The European document Putting Knowledge into Practice (European Commission 2004) specifies that the lack of skills, notably in the fields of sciences, engineering and ICT, is a challenge for European education. Another publication, Innovate for a Competitive Europe (European Commission 2004) advises companies to learn how to transform the absorbed knowledge into action. Such an innovation dynamics combines the knowledge and skills in value creation. Kolding et al 2009 describe the skills we need to acquire to face the post crisis era in Europe. The authors are convinced that the ICT skills are the most important, but they did not mention what approach to ICT and to computer programming should be used.

Companies training departments need to focus on the transformation of today capacities good for industrial economy to those that are essential for Knowledge Economy. The progress can be measured using for example the trees of knowledge software, or other that may help.

VI. Conclusion

To build a sustainable future we need more than data base, reports and dashboards, we need a disruptive innovation in the way we build, evolve, maintain and manage the human capital.

We need a new educational system, having the ambitious task of changing mentalities and values, to educate a culture of knowledge cultivators and to increase imagination and creativity. Main challenge of education is to teach how to learn, the curiosity, adaptability, capacity of solving problem with limited resources and to undertake and succeed collectively. This education is based on exchanges, listening and respecting the others opinions; an education for all, to learn from nature, from the past and from differences, in which technology and means of communication have a significant role to play.

We need to use the power and "intelligence" of computers and other connected devices. When programmed using "knowledge thinking", they can bring a significant helps in storing, updating, displaying, matching and finding the relevant elements of human capital.

We need to create synergy between educational programs and local needs as well as a dynamics vision for the future.

New metrics could be: boldness, imagination, associations (links making), and capacity to find and use the appropriate knowledge, mental flexibility, knowledge and ecosystem thinking, capacity to transform ideas in value and to envision the future. The estimation of 5D impacts of resulting activities – economic, technologic, cultural, social and environmental, could be added to measure the progress. Such a wise management of intellectual capital, supported by electronic "intelligent" assistants and appropriate measure of progress is essential for the development of companies, regions and countries.

References

[1] T. Schultz Investment in Human Capital, The American Economic Review, 1961, Vol. 51, No 1, pp. 1-17.

[2] K. E. Sveiby The New Organisational Wealth - Managing and measuring Knowledge-Based Assets. Berrett-Koehler, San Francisco, 1997

[3] N. Geraud, P. Rincel., N. Vandois ARAMIS-GM Un système intelligent d'aide à la décision pour la gestion des effectifs de Gendarmerie Mobile, Systèmes Experts et leurs applications, Avignon 1990.

[4] V. Lacroix, Lacroix, Lieutenant Colonel Daville : RAMSES I in système d'aide à la décision pour la sécurité des Jeux Olympiques, Systèmes Experts et leurs applications, Avignon 1991

[5] OCDE 1996, Measuring What People Know. Human Capital Accounting for the Knowledge Economy

[6] G. S. Becker Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. Chicago, University of Chicago Press. ISBN 978-0-226-04120-9, 1964

[7] L. Edvinsson, M. S. Malone, Intellectual Capital: Realizing your Company's True Value by Finding Its Hidden Roots. New York: Harper Business, 1997

[8] OCDE, The Well-being of Nations. The Role of Human and Social Capital. Education and Skills, 2011, http://www.oecd.org/site/worldforum/33703702.pdf

[9] P. Dixon, J. Gorecki, Sustainagility. How Smart Innovation and Agile Companies will Help Protect our Future. Kogan Page Publishers, 2010, London. 232p, 20

[10] F. Berkes, J. Colding,, C. Folke (eds). Navigating Social-Ecological Systems, Cambridge University Press, 2003, UK, pp. 352-387

[11] E. Mercier-Laurent, Innovation Ecosystems, Wiley, 2011, 248p. ISBN 978-1-84821-352-8

[12] J. P. Aubin, Viability Theory, Birkhauser Boston, 1991

[13] C. Savage, 5th Generation Management: Integrating Enterprises through Human Networking, The Digital Press, Bedford, 1990

[14] D. Amidon, The Innovation Strategy for the Knowledge Economy, Heineman Butterworth, Boston, 1997

[15] A. Lenkowa, Oskalpowana ziemia, Omega, Wiedza Powszechna, Warsaw, Poland, 1969

[16] E. P. Eckholm, Losing Ground. Environmental Stress and World Food Prospects, W.W. Norton and Company, New York, 1976

[17] Y. Arthus-Bertrand, Home, 2009 https://www.youtube.com/watch?v=jqxENMKaeCU

[18] M. Autier, P. Lévy Pierre, Les arbres de connaissances, 1992, La Découverte, Paris

[19] J. Kolodner, Kolodner Jeannet, Case-Based Reasoning, Morgan Kaufman, 1993, 668p, ISBN 978-1558602373

[20] European Commission, Implementing the partnership for growth and jobs: Making Europe a pole of excellence on corporate social responsibility, 2006

[21] European Commission, 2004, Innovate for a Competitive Europe. A New Action Plan for Innovation, 2 April 2004

[22] M. Kolding, M. Ahorlu, C. Robinson C.., Post crisis: e-skills are needed to drive Europe's innovation society, IDC EMEA,2009, London, United Kingdom

[23] A. M Youriev History of human capital, 2014, http://www.yuriev.spb.ru/polit-chelovek/human-capital-resource

DOI: 10.15439/2014F444
ACSIS, Vol. 3

Music Information Retrieval as a Key Framework to Explore Legal Issues Linked to Personal Data Computation

Pierre Saurel
Université Paris-Sorbonne, SND
Email: pierre.saurel@paris-sorbonne.fr

Francis Rousseaux
Université Reims Champagne, CReSTIC
Email: francis.rousseaux@univ-reims.fr

Marc Danger
ADAMI
Email: mdanger@adami.fr

Abstract—**The forthcoming European regulation on data privacy penalizes violations by a fine of up to one hundred million euros: European Music Information Retrieval researchers must be compliant with any personal data processes. They are not allowed to transfer personal data to the rest of the world, excepted by using so-called "Safe Harbors".**

Detection of any personal data is mandatory, and "whether a person is identifiable, account should be taken of all the means reasonably likely to be used to identify or single out the individual directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification".

The paper proposes a roadmap for ISMIR involving:

- **Methodology (Is "Privacy by Design" a universal solution?);**
- **Epistemology (Are all authorship attribution algorithms separable into data and processes?);**
- **Science (What characterizes a maximal subset from the big data that could not ever be computed by any Turing machine to identify a natural person with any algorithm?);**

Politics (How can ISMIR influence data privacy policies? Is it possible through some metadata standardization activities?).

I. Introduction

THE Music Information Retrieval (MIR) community addresses a wide range of scientific, technical and social challenges, dealing with processing, searching, organizing and accessing music-related data and digital sounds through many aspects, considering real scale use-cases and designing innovative applications, exceeding its academic-only initiatory aims.

Recent Music Information Retrieval tools and algorithms aim to attribute authorship and to characterize the structure of style, to reproduce the user's style and to manipulate one's style as a content [1], [7]. They deal for instance with active listening, authoring or personalised reflexive feedback. These tools will allow identification of users in the big data: authors, listeners, performers [2], [10].

As the emerging MIR scientific community leads to industrial applications of interest to the international business (start-up, Majors, content providers, platforms) and to experimentations involving many users in living labs (for MIR teaching, for multicultural emotion comparisons, or for MIR user requirement purposes) the identification of legal issues becomes essential or strategic.

The MIR community already seized the technical challenge of Digital Right Management. This challenge was one of identified legal issue related to copyright and Intellectual Property. The MIR community seized the challenge related to Information Access. This challenge was connected to security, business models and right to access [11]. Privacy is another legal challenge. A classification of personal data and processes is necessary to address this challenge precisely. A naive classification appears when you quickly look at the kind of personal data MIR deals with:

User's evaluation, comments, annotation and music recommendations are obvious personal data as long as they are published under their name or pseudo;

Internet Protocol (IP) addresses, Media Access Control (MAC) addresses and addresses allowing identification of a device or an instrument, are linked to personal data;

Any information allowing identification of a natural person, as some MIR processes do, shall be qualified as personal data and processing of personal data.

But the legal professionals do not unanimously approve this classification. For instance the Court of Appeal in Paris judged in two decisions (2007/04/27 and 2007/05/15) that the IP address is not a personal data.

II. Regulation of Personal Data Processes

A specific classification of MIR personal data processes must consider the applicable law of personal data and take the diverging international regulations into account.

A. Taking the divergence between European and American legal approaches into account

Europe regulates data protection through one of the highest State Regulations in the world (two Directives and a Regulation of the European Parliament and of the Council to come) when the United States lets contractors organize data protection through agreements supported by consideration and entered into voluntarily by the parties. These two legal approaches are deeply divergent. United States lets companies specify their own rules with their consumers while Europe enforces a unique regulated framework on all companies providing services to European citizens. For instance any company in the United States can define how long they keep the personal data, when the regulations in Europe would specify a maximum length of time the personal data is to be stored. And this applies to any company offering the same service.

The European Commission's Directive on Data Protection (95/46/CE – The Directive) prohibits any transfer of personal data to non-European Union countries that do not meet the European Union adequacy standard for privacy protection is strictly forbidden. The divergent legal approaches and this prohibition alone would outlaw the proposal by American companies of many of their IT services to European citizens. In response the U.S. Department of Commerce and the European Commission developed the Safe Harbor Framework (SHF) [8]. Any non-European organization is free to self-certify with the SHF and join.

The European Parliament voted on 12 March 2014 a new Proposal for a Regulation on the protection of individuals with regard to the processing of personal data. The Directive allows adjustments from one European country to another and therefore diversity of implementation in Europe when the regulation is directly enforceable and should therefore be implemented directly and in the same way in all countries of the European Union. This regulation should apply in 2016. This regulation enhances data protection and sanctions to anyone who does not comply with the obligations laid down in the Regulation. For instance (Article 79) the supervisory authority will impose, as a possible sanction, a fine of up to 100 million euros or up to 5% of the annual worldwide turnover in case of an enterprise, whichever is higher.

B. Data protection applies to any information concerning an identifiable natural person

Under French Law were personal data only defined considering sets of data containing the name of a natural person. This State of the Law changed with the application in France of the 95/46/CE European Directive. The definition of personal data has been extended; the 95/46/CE European Directive (ED) defines 'personal data' (Article 2) as: "any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity".

For instance the identification of an author through the structure of his style as depending on his mental, cultural or social identity is a process that must comply with the European data privacy principles.

C. Safe Harbor is an identified Framework allowing to avoid to pay the financial fine up

Compliance with Safe Harbor is an issue for an organization using MIR processing to fulfill the high level European standard about personal data, to operate abroad and to be confident in avoiding prosecution regarding personal data. An American organization may decide to enter the US-EU SHF's requirement. This Company has to design a data privacy policy complying the seven Principles (SHP).

First of all organizations must identify personal data and personal data processes. Then they apply the SHP to these data and processes. By joining the SHF, organizations must implement procedures and modify their own information system whether paper or electronic.

Organizations must notify (P1) individuals about the purposes for which they collect and use information about them, to whom the information can be disclosed and the choices and means offered for limiting its disclosure. Organizations must explain how they can be contacted with any complaints. Individuals should have the choice (P2) (opt out) whether their personal information is disclosed or not to a third party. In case of sensitive information explicit choice (opt in) must be given. A transfer to a third party (P3) is only possible if the individual made a choice and if the third party subscribed to the SHP or was subject to any adequacy finding regarding to the ED. Individuals must have access (P4) to personal information about them and be able to correct, amend or delete this information. Organizations must take reasonable precautions (P5) to prevent loss, misuse, disclosure, alteration or destruction of the personal information. Personal information collected must be relevant (P6: data integrity) for the purpose for which it is to be used. Sanctions (P7 enforcement) ensure compliance by the organization. There must be a procedure for verifying the implementation of the SHP and the obligation to remedy problems arising out of a failure to comply with the SHP.

III. Classification for Music Information Retrieval Personal Data Processing

Considering the legal definition of personal data we can now propose a less naive classification of MIR processes and data into three sets: (i) nominative data, (ii) data leading to an easy identification of a natural person and (iii) data leading indirectly to the identification of a natural person through a complex process.

A. Nominative data and data leading easily to the identification of a natural person

We first consider two sets of processes. The first set aggregates the information systems directly containing the name of a natural person. The second set aggregates the cases allowing a direct or an indirect identification easily done for instance through devices.

In these two sets we find that the most obvious set of data concerns the "Personal Music Libraries" and "recommendations". As long as one recommends music to a user or analyze their personal library, he certainly deals with his privacy?

B. Data leading to the identification of a natural person through a complex process

The third set of personal data aggregates the information systems when a natural person is indirectly identifiable using a complex process, like some of the MIR processes.

Can one work on Machine Learning and especially on Categorization with no consideration about the taste or the style of the consumers or of the users? These processes belong for the most part to this third set. Looking directly at the data without any sophisticated tool does not allow any identification of the natural person. In contrast, MIR nonlinear algorithms or machine learning leads frequently to an indirect identification [7].

Usually do MIR algorithms use inputs to build new data which are outputs or data stored inside the algorithm, like weights for instance in a neural net [6].

C. The legal criteria of the costs and the amount of time required for identification

This third set of personal data is not as homogeneous as it seems to be at first glance. Can we compare sets of data that lead to an identification of a natural person through a complex process?

The European Proposal for a Regulation designs the concept of identifiability. It gives a legal criteria to determine if an identifiable set of data is or is not personal data. It regards the identification process (Recital 23) as a relative one that would change according to the effectiveness of the identification: *"To determine whether a person is identifia*ble, account should be taken of all the means reasonably likely to be used either by the controller or by any other person to identify or single out the individual directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the pro*cessing and technological development."*

How can we define, as MIR practitioners, when a set of data will result in an easy identification and should then be classified into the second set or, on the other hand, is especially uncertain so that it cannot be considered and categorized as personal data? New criteria are necessary to answer these new challenges and interrogations about MIR processes.

IV. What is the Third Set about

"Identifiability", in our classification, is a potentiality of a set of data. This set should be qualified as being personal data if the cost and the amount of time required for identification are reasonable. These new criteria are a step forward since the legal qualification is not an absolute one anymore and depends specifically on the state of the art [13].

A. Available technology and technological development to take into account at this present moment

The internet is one of the high level technologies that modifies the identifiability of a set of data. All over the world, people publishes data and personal data without heeding the potentialities of these data. They are usually not completely aware about the ways these data can be used to describe their personal habits. When a listener tags music or recommends an item, he publishes information allowing to paint a portrait of his personality. If a company exploits this user data without integrating strong privacy protections, he can encounter legal issues and extensive disaffection from his customers.

The volume of data have increased faster than "Moore's law": This is the "Big Data". New data is generally unstructured and traditional database systems such as Relational Database Management Systems cannot handle the volume of data produced by users and by machines & sensors. This challenge was the main drive for Google to define a new technology: the Apache Hadoop File System. Within this framework, data and computational activities are distributed on a very large number of servers. Data is not loaded to be computed, and the result stored. Here, the algorithm is close to the data. This situation leads to the epistemological problem of separability into the field of MIR personal data processing: are all MIR algorithms (and for instance the authorship attribution algorithms) separable into data and processes. An answer to this question is required for any algorithm to be able to identify the set of personal data it deals with.

Databases of personal data are no more clearly identified. We can identify new scientific challenges about MIR personal data processing. These are the result of five complementary sides of the situation.

Data Sources Profusion. Many new databases and datawarehouses are developed every day allowing to trace and recognize many kind of information. Software, as Spotify for example, become new kind of live and on-line data sources providing a flow of music consumption information from numerous users. These data sources will shortly integrate new devices under the Internet of Things. Not all of these data are of high-quality. These new data and datasources dot not always preserve the legal rights of the users. Taking this into account will be of great help to shape reliable and sustainable systems.

Crossing & Reconciling Data. Data sources are not separated and independent one of the other. The aggregation of the data will first allow a more precise identification through user id, cookies or emails and then make technically possible to bring closer, combine and blend data that were earlier incommensurate.

Temporal Aspects. The memory of the web, such as information systems in general, is beyond all what people can usually imagine. The Public Status of Data changes frequently. Most of the users do not trace their own personal data. A video posted during a party could reappear suddenly when one is seeking a job. All the traces collected one day with a given purpose could be technically exploited later with a different or an opposite purpose.

Permanent Changes. The general instability of the data sources, technical formats and flows, applications and use is another strong characteristic of the situation. The impact on personal data is very likely. If the architecture of the systems changes a lot and frequently, the social norms also change. Users today publicly share information that they would have considered totally private a few years earlier. And the opposite could be the case.

User Understandability and Control. These situations are less and less easy to understand for normal users. The complexity of the systems and of the interactions between humans and machines results in non-linear causality which may leads to confusing situations. The case of the private Facebook posts displayed all at once on the timeline (Sept. 2012) is full of meaning. Facebook announced that there was no bug. Those information were old personal posts which became more publically visible with the new timeline. This situation is due to two simultaneous factors: the misunderstanding of a human user combined with the rate of change of an information system.

Changes in the Information Technology result to a shift in the approach of data management: from computational to data exploration. The main question is "What to look for?" Many companies design new systems and processes to "make the data speak". Direct marketing is one of these players: dataflow of personal data are produced through the big data.

One of the solution could be a stabilization of the dataflow through a universal design of metadata. This could be a way to speed up a specific classification of MIR processing of personal data into identifying and non-identifying processes.

This situation results into a new scientific challenge: What could be an absolute criterion about the identifiability of personal data extracted from a set of data with a MIR process? How could we define into the big data, a maximal subset that could not ever be computed by any Turing machine to identify a natural person with any algorithm?

B. Personal data produced on the fly: the Gamelan Project as a case study

IRCAM supervised the Gamelan Project (2009-2013) in partnership with EMI, INA and UTC [3] [4].

Gamelan is a software environment, built upon the production ecosystem, to address the reconstitution issue of digital music production, by combining trace engineering, knowledge modelling and knowledge engineering. Reconstitution is usually relegated afterwards. Gamelan reconstructs the composer-system interactions that have led to the creation of a work of art that is about to leave the production studio. The purposes concerns long-term preservation, repurposing, versioning, evolution of the work of art and the disclosure of the contingencies of its initial outcome.

A creator finishing his or her work in a studio marks the end of the production process: the long-awaited object is finally there, thus the creator, the producer, the sound engineer and all the people involved are happy or at least relieved; the goal is reached and the story reaches its end. However, at this very moment, because the final object is there, no one wonders about its reconstitution.

But —say ten years later— when "back-catalog" teams of music companies want to edit some easy-to-sell Greatest Hits in up-to-date audio formats, delving into the musical archives is no longer easy. Returning to the reachable-recorded digital files, it may be difficult to figure out which one of the bunch of files left is the one needed. File dates and file names cannot be trusted.

Closer in time —say two months after the production— the simple task of collecting vital information on the contributors who actually worked on the project may turn into a real problem. There is a whole set of information on contributions (name, role, time spent, etc.) necessary to manage salaries, rights and royalties that regularly proves hard to collect afterwards. Evidently, this kind of information would be far easier to collect directly at production time.

On the surface this is nothing to do with privacy and personal data! But in fact, and it is typically the case as soon as a complex person-software device is involved, this type of project invites us to rethink classical approaches and qualifications of privacy issues. The Gamelan project exemplifies several of the many R&D emerging questions that are raised in the digital audio processing domain.

First of all reconstitution requires to collect traces during the production process itself. Automatically-collected software traces differ from human-entered traces. The former can be seamlessly collected through automatic watching components, with interfaces traces and logs as heuristic material, while the latter inevitably requires a human contributor for information that cannot be automatically retrieved or inferred from automatic traces. A full-production tracking environment would resemble Living Labs, towards a Living Studio.

Secondly, these traces call for an appropriate knowledge model. To remain as uninvasive as possible, such a model should provide means to determine which information is worth asking people during the production or not compared to the cost of disturbing the creativity. Without a knowledge model, it would not be possible to interpret the traces or to determine the kind of traces worth retrieving. To achieve this model, professional knowledge must be identified, listed and characterized with experts, defining a digital music production Knowledge Level.

Within Gamelan, traces from used operating system and from used professional applications are extracted. Semantic networks dealing with typical digital audio composition acts are involved towards some abstraction of those traces, but personal data is nowhere considered: some real time digital audio flow is involved, transformed on the fly by creative acts that assign the composers particular style and their artistic singularity.

The composers style, as part of built up personal data, often not even named, is computed to support the Gamelan reconstruction process: what is to be reconstituted has to do with the abstract truth of the piece of art and its stylistic genesis. To understand that "the composer is currently testing a sample within the whole piece framework" is more efficient than being aware of a succession of cut-paste-listen actions that has to be generalized.

Thus some personal data, like artistic style, is built up on the fly, relating to processing algorithms, knowledge bases and title repositories [5], evolving from the system experience itself, and only known by the system. The ultimate target is clearly the style-recognition [1] of the creators, viewed as the correlation between their practice and the character of their work of art.

C. What if the Gamelan Project is called in front of a law court?

The Gamelan Project is a case study where machines produce personal data on the fly. Under our classification it is a case study which produces a kind of personal data from the third set. In that kind of case, one must be reactive and cannot just apply the Safe Harbor Principles at the end of the project. How could we apply the Safe Harbor Framework from the beginning and avoid any prosecution for breaking the data privacy laws since we cannot decide at first which set of data is personal?

A special methodology must be applied taking into account the possibility that, during the project, new data and processes could be qualified as personal data.

V. Privacy by Design: a Methodology for MIR Project Management?

Privacy by Design (PbD) was developed in the 1990s. This methodology has become an international reference about privacy. It now evolves taking the big data into account.

A. Foundations Principals (FP) of Privacy by Design

PbD is grounded on on seven FP1: PbD *"*is an approach to protect privacy by embedding it into the design specifications of technologies, business practices, and physical infrastructures. That means building in privacy up front – right into the design specifications and architecture of new systems and processes. PbD is predicated on the idea that, at the outset, technology is inherently neutral. As much as it can be used to chip away at privacy, it can also be enlisted to protect privacy. The same is true of processes and physical infrastructure*"*:

- Proactive not Reactive (FP1): the PbD method is grounded not on reactive but on proactive parameters anticipating and preventing privacy invasive events before they occur;
- Privacy as the Default Setting (FP2): the default parameters attempt to fix the maximum degree of privacy;
- Privacy embedded into Design (FP3): the architecture of IT systems takes privacy into account;
- Full Functionality – Positive Sum, not Zero-Sum (FP4): PbD results into a "win-win" solution considering the different interests and objectives;
- End-to-End Security – Full Lifecycle Protection (FP5): the design of the solution regarding to privacy and security measures is defined in a complete Lifecycle;
- Visibility and Transparency — Keep it Open (FP6): PbD is an open process and the quality of the solution can be asserted by an external audit. Transparency is a key success factor;
- Respect for User Privacy — Keep it User-Centric (FP7): as they apply PbD will SI designers take the user's personal interest into account especially concerning his privacy and personal data.

PbD is a key-concept in legacy [9] when you have to design numerical and digital processes. The European Union2 affirms that *"PbD means that privacy and data protection* are embedded throughout the entire life cycle of technologies, from the early design stage to their deployment, use *and ultimate disposal".* Europe [12] took the Canadian experiments into account when it decided to use PbD as a key-concept in the heart of the legal data protection.

B. Prospects for a MIR Privacy by Design

PbD is a standard for designing systems and processing involving personal data. PbD was enforced by the new european proposal for a Regulation (Article 23). It becomes a method for these designs whereby it includes signal analysis methods and may interest MIR developers.

This proposal leads to new questions as for instance: Is PbD a universal methodological solution about personal data for all MIR projects? Most of ISMIR contributions are still research oriented, in the sense of Article 83 of the "Safeguarding Privacy in a Connected World". To say more about that intersection, we need to survey the ISMIR scientific production, throughout the main FP.

FP6 (transparency) and FP7 (user-centric) are usually respected among the MIR community as source code and processing are often (i) delivered under GNU like licensing allowing audit and traceability (ii) user-friendly. However, as long as PbD is not embedded, FP3 cannot be fulfilled and accordingly FP2 (default setting), FP5 (end-to-end), FP4 (full functionality) and FP1 (proactive) cannot be fulfilled even. Without any PbD embedded into Design, there are no default settings (FP2), you cannot follow an end-to-end approach (FP5), you cannot define full functionality regarding to personal data (FP4) nor be proactive. Principle of pro-activity (FP1) is the key. Fulfilling FP1 you define the default settings (FP2), be fully functional (FP4) and define an end-to-end process (FP5).

In brief is PbD useful to MIR developers even if PbD is not the definitive martingale!

VI. Structuring the Data Bases heeding Privacy by Design: The case of the Gamelan project

The three sets of personal data designed considering legal rules relative to data privacy should be tuned and man-

[1] http://www.ipc.on.ca/images/Resources/7foundationalprinciples.pdf

[2] "Safeguarding Privacy in a Connected World – A European Data Protection Framework for the 21st Century" COM(2012) 9 final.

aged to be able to be stored into different tables or servers.

The Gamelan project [3] is a used case allowing the design of structuration of the personal data. The Gamelan project was designed into three layers to:

- Track production process,
- Interpret collected traces according to a domain ontology,
- Help querying and visualizing to foster production understanding.

These levels are closed to the classical layers: the data level, the information level and the knowledge level. The classification of the three sets could deal with these three levels.

The data level is more or less specifically relative to the two first sets. The information and knowledge level are connected to the third set.

Concerning the Gamelan project, the learning processes at the information and knowledge level can be tuned depending on the goals and aims. Depending on this tuning and on the learning time will the processed data become personal data or not meaning that they have allowed the identification of a person in the data base.

VI. Conclusion and Future Work

Methodological Recommendations. This classification leads to methodological recommendations for MIR researchers. The first step is to audit the used algorithm and the data. Could the algorithm identify a natural person in the sense of the two first sets of our classification? In that case, the researcher should use the SHF. To use SHF is not as simple as it seems to be. In some cases it can lead to huge industrial challenge for instance regarding Cyber Security (P5).

In some cases the MIR community develops new personal data on the fly. It can be so when researchers use all the known data algorithms and data analysis especially relative to machine learning. The PbD methodology should then be applied. This methodology frames a design that preserves personal data and avoids any unintentional loss of data.

But the time when data (on the one hand) and processing (on the other hand) were functionally independent, formally and semantically separated, has ended. Nowadays, MIR researchers currently use algorithms that support effective decision, supervised or not, without introducing ‘pure’ data or ‘pure’ processing, but building up acceptable solutions together with machine learning or heuristic knowledge that cannot be reduced to data or processing: The third set of personal data may appear, and raise theoretical scientific problems.

Political Opportunities. The MIR community computes algorithms dealing with style description since a long time. The MIR community could join expert groups dealing the legal aspects of personal data. It could explain to the lawyers these algorithms and the ones relative to machine learning. This could be of great benefit for both parties.

Future Scientific Works. The three sets and the new definition of personal data leads to new pure scientific challenges. These challenges constitute our research program for future works. What are the conditions to specify a set of data resulting into an identification in the sense of the second set. When can we assure that an algorithm allows a too hazardous recognition? In that case we would say that the data are not personal in a legal sense? How can we design or carve a maximal subset from the big data that could not lead to the identification of a natural person by any Turing machine and with any known or forthcoming algorithm? These are some of the new scientific challenges we are now dealing with.

References

[1] S. Argamon, K. Burns, S. Dubnov (Eds): The Structure of Style, Springer-Verlag, 2010. DOI 10.1007/978-3-642-12337-5

[2] K. Barkati, A. Bonardi, A. Vincent, F. Rousseaux: “GAMELAN: A Knowledge Management Approach for Digital Audio Production Workflows”, Proceedings of the European Conference on Artificial Intelligence, Workshop “Artificial Intelligence for Knowledge Management”, 2012.

[3] C. Barlas: “Beating Babel - Identification, Metadata and Rights”, Invited Talk, Proceedings of the International Symposium on Music Information Retrieval, 2002.

[4] T. Bertin-Mahieux, D.P. Ellis, B. Whitman, P. Lamere: “The Million Song Dataset”, Proceedings of the International Symposium on Music Information Retrieval, 2011.

[5] J.S. Downie, J. Futrelle, D. Tcheng: “The International Music Information Retrieval Systems Evaluation Laboratory: Governance, Access and Security”, Proceedings of the International Symposium on Music Information Retrieval, 2004.

[6] A. Gkoulalas-Divanis, Y. Saygin, Vassilios S. Verykios: “Special Issue on Privacy and Security Issues in Data Mining and Machine Learning”, Transactions on Data Privacy, Vol. 4, Issue 3, pp. 127-187, December 2011.

[7] D. Greer: “Safe Harbor - A Framework that Works”, International Data Privacy Law, Vol.1, Issue 3, pp. 143-148, 2011.

[8] M. Levering: “Intellectual Property Rights in Musical Works: Overview, Digital Library Issues and Related Initiatives”, Invited Talk, Proceedings of the International Symposium on Music Information Retrieval, 2000.

[9] F. Pachet, P. Roy: “Hit Song Science is Not Yet a Science”, Proceedings of the International Symposium on Music Information Retrieval, 2008.

[10] V. Reding: “The European Data Protection Framework for the Twenty-first century”, International Data Privacy Law, volume 2, issue 3, pp.119-129, 2012. DOI 10.1093/idpl/ips015

[11] A. Seeger: “I Found It, How Can I Use It? - Dealing With the Ethical and Legal Constraints of Information Access”, Proceedings of the International Symposium on Music Information Retrieval, 2003.

[12] A.B. Slavkovic, A. Smith: “Special Issue on Statistical and Learning-Theoretic Challenges in Data Privacy”, Journal of Privacy and Confidentiality, Vol. 4, Issue 1, pp. 1-243, 2012.

[13] P. Symeonidis, M. Ruxanda, A. Nanopoulos, Y. Manolopoulos: “Ternary Semantic Analysis of Social Tags for Personalized Music Recommendation”, Proceedings of the International Symposium on Music Information Retrieval, 2008.

DOI: 10.15439/2014F484
ACSIS, Vol. 3

Tool dilemmas of innovation

Jolanta Sala
Powiślański College, ul.
11Listopada 13, 82-500 Kwidzyn,
Poland
Email: jolasala@interia.pl

Halina Tańska
University of Warmia and Mazury
in Olsztyn, ul. Słoneczna 54,
10-689 Olsztyn, Poland
Email: tanska@uwm.edu.pl

Abstract—The article has indicated the phenomenon of the low-level of ICT use in Polish enterprises as well as its consequences, in particular an unimpressive activity of innovative companies. The approach that has been proposed is to identify the multidimensional expanse of information management through the identification and analysis of information gaps. The article has exposed the importance of the seven vulnerable areas such as development, competency, motivation, emotion, generation, efficiency, technique and technology. Tools, individualized for each company in relation to the information and ICT, are an important determinant of innovation.

I. Introduction

ICT technologies depreciate a number of traditional skills, among other things they have eliminated the core competences of engineering-technical workers which used to be crucial (that is, the use of slide rule, drawing board, rapidograph). The phenomenon of the depreciation is in particular visible in the Polish industry as its negative effect was multiplied by improper realization of the system transformation. In Polish service companies the situation is not good as well. Jeffrey Sampler, a famous researcher of ICT technologies suggests, that the companies should make a list of competences they are currently lacking (incompetence, lack of skills, awkwardness) and which they will need in the future, instead of determining their key competences. This was the intention of the article's authors who present different information gaps referring to few areas, although they do not use all types identified both in professional literature and practice.

It is worth considering "what far-reaching effects are connected to the so-called digital disability syndrome, which can be understood as depriving some communities of the access to information and their sources (the Internet) or their inability to acquire it" [1].

The authors have been publishing articles on improper approach towards creating the base for informative society in our country [2], [3] as its effects are visible now. Poland is at the end of the ranking consisting of the EU countries. One can analyze in detail the reasons for this situation, but it is difficult to expect pointing out consequences with respect to the institutions and personal responsibility for the disappointing results. Therefore, it is more pragmatic to focus on the key aspects of repairing the current situation. Undoubtedly, one of the effects is the low percentage of innovatively active and innovative enterprises in Poland (18,1% of industrial enterprises and 13,5% of the ones providing services) [4], and one of the reasons are the mistakes in conducting the policy connected to ICT technologies.

II. Concealment or the lack of awareness

The authors have repeatedly indicated that in various studies, reports, and analysis on the use of ICT in socio-economic life their authors move on the border between the official image and the truth. This phenomenon is quite common and it occurs also in governmental institutions, the Central Statistical Office and others [5]. Summing up, reporting on the complacency as well as the optimism go much too far and the measurements are interpreted too loosely and imprecisely. This applies also to the enterprises and individuals (households).

The proper use of ICT in the economy has a direct impact on the competitiveness and innovativeness of Polish enterprises. Unfortunately, the use of ICT in enterprises has often an image character and this has an influence on reporting. Therefore, it is difficult to discern why it is so bad while it so good? There is no doubt the cause of this condition is an "easy" investment in modern "furnishings" of the offices in PCs and the lack of elementary skills of using them by the employees. The lack of training or improper selection of courses deepen the bad situation in the case of more experienced employees (usually those aged 30+, which did not experienced the mass ICT education from the level of primary school). In the conditions of unemployment the process of hiring young people is limited, and among this group the best ones decided to emigrate for financial reasons.

Therefore, it is difficult to give a verdict in this particular case, whether there is a concealment of or even distorted

This work was not supported by any organization

reporting or it is a genuine lack of awareness of the average low-level use of ICT (many times on the level of IT illiteracy). In such cases it is also difficult to speak about the usage of ICT to increase the innovative activeness of the enterprise.

Unfortunately, inadequate ICT tools and the lack of skills to use them both in private and professional lives often result in the unconscious and unvested ability to obtain information from various sources available in a digital form. This phenomenon is acute in the Polish SME sector. While analyzing information issues in the economy it is essential to appeal to the achievements of a new scientific discipline which is certainly the economics of information specified by professor J.Oleński. In this paper, it is worth recalling the importance of the meta-informational minimum and meta-informational and information gaps.

III. Meta-informational and informational minimum

The thesis which does not need to be verified is that social groups and economic subjects can perform their duties when the informational needs in this regard are established. Much more far-reaching is the belief that the deepening social, cultural and economic polarization of the communities, regions and countries is "a consequence of the ever-growing functional minimum of information, high-level and sustained growth of the informational requirements" [6]. It should be assumed that the reason for this popularization is the difference in the available information resources and opportunities to access the global informational systems. A situation in which there are disparities cited here leads to the creation of "information gap". It is universal and considered from many angles. According to the authors, it covers many areas (planes), thus it seems legitimate to discuss its dimensions.

The attempt of dimensional analysis should be started from the meta-level and from recalling the definition of meta-information (that is information about information) that can be entered into a perfect circular shape (Fig. 1) which allows avoiding the justification of the order and adding or removing individual dimensions (such as paradigm, infrastructure, human, economic subject, environment) depicted in Figure 1 in a form of a pie slice. The similar situation concerns identifying the gaps (that is the development, technology-technological, competence, educational, language, politic-legal and quality ones etc.), which were localized in direct environment of "information dimensions circle".

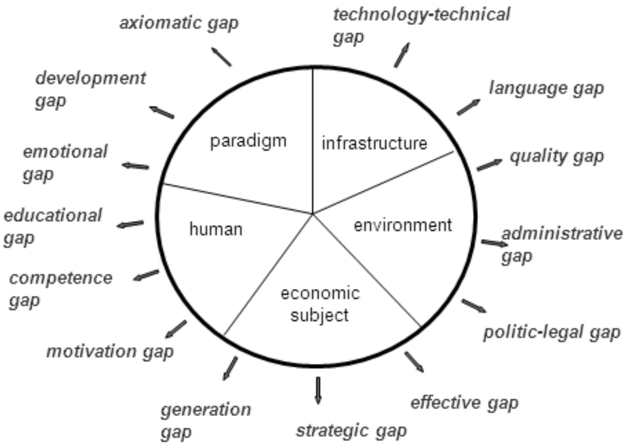


Fig. 1 Multidimensional information expanse

Undoubtedly, the graphic model of multidimensional information expanse presented in Figure 1 requires a deeper interpretation. It illustrates the complexity of the analyzed phenomena and inspires to define a more uniform and useful structure. A distinct perspective presented in Figure 2 meets these requirements.

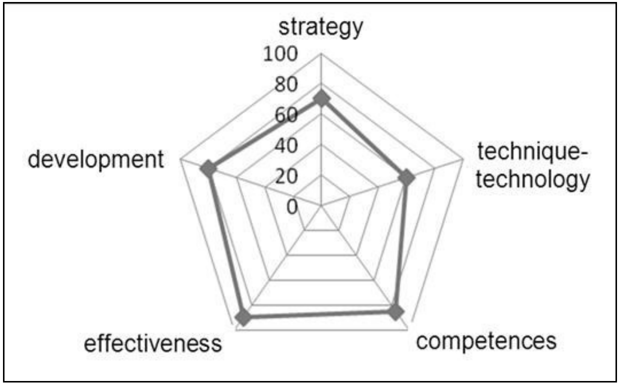


Fig. 2 Dimensions of meta-informational gap

The positioning of meta-informational gap on a radar chart enables significantly more precise analysis of the socio-economic reality and, above all, measuring the quality of events. Today one can notice a steady technical, organizational and civilization progress which results in enlarging the gap between available resources of information and information which is necessary for efficient, effective operation in specific situations. In highly developed countries the governments and socio-economic organizations do not allow the creation of such a situation because one of the important tasks of the national information infrastructure is to fill the information gap. This means the elimination of discrepancies between information that is available and information that is needed. A simplified model of information gap is presented on Figure 3, where the coordinates are the amount of information and the complexity of the problem. There is a close correlation (elation) between the size of the gap (reducing vulnerability) and incurred costs (expenses incurred). Not always it is worth aiming at minimizing the information gap due to the economic aspect.

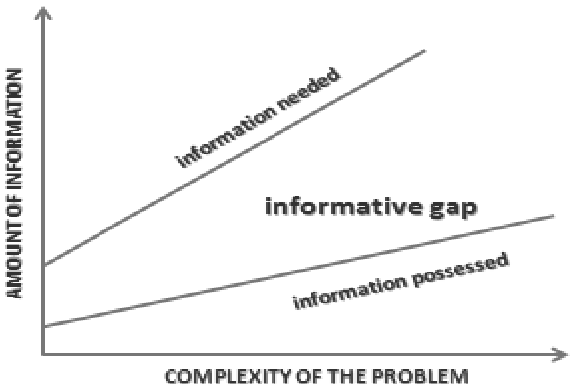


Fig. 3 Information gap

Precise determining information gap planes is associated with the definition of multidimensional expanse of information management (EIM). To concrete EIM one can assume the record in the form of Cartesian quotient according to the formula (1):

EIM = X(Z, P, TI, F, t), (1)

where:

EIM – information expanse management,

X – the sign of Cartesian quotient on elements presented in brackets,

Z – information resources,

P – information processes,

TI – information technologies,

F – functions of managing (planning, organizing, control etc.),

t – time horizon of the management (operative, tactical, strategic).

In such a way it is easy to create „the map of tasks and problems" where „each point on information expanse management is a specified task related to a specific resource in a particular time horizon, in the context of specified function of management" [7].

Therefore, one can return to the consideration of the informational needs using two complementary approaches cited in the academic textbook quoted above. The first proposed by prof. H. Egeman considers informational needs in the context of the scope (resources, processes and information technologies), time horizon of the management (operative, tactical and strategic management) as well as the function of managing information (planning, organizing, control etc). It is understandable that it involves "planning, organization, implementation and control of the implementation of the ongoing work in using informational organizations (including access to information) and current (daily) administration of information to ensure proper quality, consistency and safety (including confidential) information" [7]. According to the second approach whole information needed for the user U in his daily business or because of his or her non-economic interests conditions the type of problem to be solved Q as well as the knowledge and experience of an individual (user U). Prof. B. Stefanowicz divided informational needs into two complementary subsets. Iu symbol represents a subset of the information needed to solve Q, and already available to the user. And the symbol of L, is a subset of this information, which is needed and is not directly accessible (Fig. 4). The development of informational needs can be written down by using the formula (2):

$$< U, Q, M > \Rightarrow \boldsymbol{I} \Rightarrow \boldsymbol{Iu} \cup \boldsymbol{L} \qquad (2)$$

The most important features of the set L, which was called an information gap by B. Stefanowicz can be characterized in the following way:

1) The gap is always "somebody's" gap so it cannot be analyzed in separation to a specified user and the problem to be solved;

2) Volatility in time t;

3) Fuzziness of the set L borders L;

4) The set of needed but unknown information divided on the basis of content, mutually connected by multiple relationships.

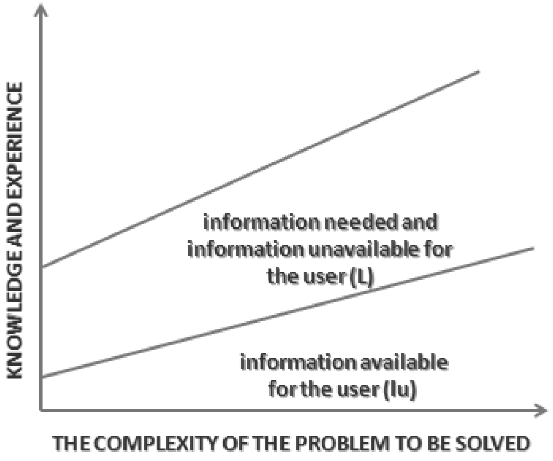


Fig. 4 Information gap – knowledge and experiences view

IV. PRECISE ANALYSIS OF GAPS AS THE KEY TO PERFECTNESS

The abundant literature in the field of information society and economic practice research conducted by the authors confirmed the need for a new approach [8], which will not only facilitate the survival of businesses, but will allow them to become more innovative and pro-developmental. In this regard, the term development gap should be mentioned. It means the difference between the potential actions (for example organizational culture, management skills, logistics skills, available resources, that is, passive force) and the potential effects (for example efficiency and cultural aspirations, the structure of power, features strategic leadership). In other words, it is a difference between the desired and real state of the enterprise. So the development gap is the difference between production abilities of the enterprise and actual achievements. In many cases the creation of development gap is connected to the lack of ability of the management staff to guarantee and maintain a proper pace of enterprise's development as well as the lack of proper tools.

The pace of development of enterprises is highly dependent on the employees and their competences as well as bilateral (creative) relationship between the employee and the employer. For companies operating in area of a constant change it is reasonable to question whether the employees have sufficient competences to perform the work tasks according to new requirements. Their failure means the arrival of the competence gap and the need to diagnose the areas. Competence gap in the organization is a collection of knowledge, skills and attitudes which the company hasn't obtained yet and which are a precondition for the proper functioning of the organization after the implementation of the changes. Triangle of competence with basic dependencies is presented in Figure 5.

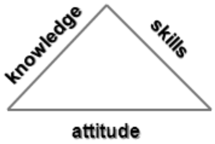


Fig. 5 The competence triangle – the relationship between the factors

Although fig.5 is more suggestive, in practice its usage is limited. Nevertheless, there are techniques and instruments supporting the ones willing to control the competence gap. However, one must accept the standards of competences created in a form of European and national competence frameworks (ERK, PRK) and also keep up with changes enforced by the fast development of ICT.

At the same time one should develop and deepen the pro-active and pro-innovative involvement of most employees otherwise the phenomenon of motivation gap (MG) will appear, which according to the concept of D.A. Nadler and E. F. Lawler is the difference between the two values, that is the level of employees' expectations about a particular stimulus (A) and degree of meeting these expectations by an incentive system (B). Formula (3) presents motivational gap in a more formal way:

$MG = B - A$. (3)

Fig. 6 presents the motivational gap in a very unequivocal way.

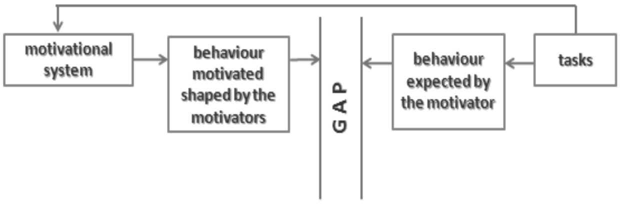

motivational system
behaviour motivated shaped by the motivators
GAP
behaviour expected by the motivator
tasks

Fig. 6 Motivational gap as the difference between the expected and actual behavior of the employee [9]

It is also worth looking at the company as a diverse set of individuals (employees) teaming two worlds within its borders: the people who live in an active way and people who work for a few hours to live. Recently passion and entrepreneurial attitudes are promoted as a desirable behavior. They break down and turn around the traditional image of the division of work invested in the life of man and reverse the proportions. The change in the area is clearly presented below (Fig. 7). Life begins to be consumed by the work/activity. In professional literature, this phenomenon is identified as emotional gap [9].

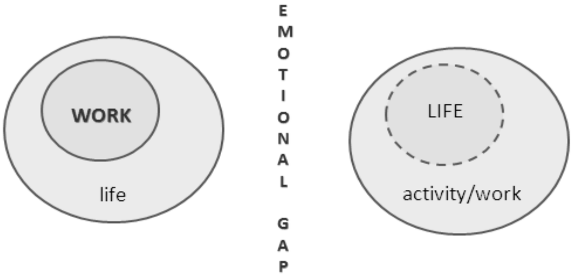


Fig. 7 Two sides of supposedly the same world [10]

It should also be noted that only a devoted employee is aware of the needs of the business and working with others, with the aim of increasing productivity for the sake of the economic subject. At the same time in the last decades of the twentieth century and two decades of the twenty-first century in the workplace a new generation of so called "virtual reality generation" and "green generation" appeared, which is a major challenge for the managers. In the professional literature, "the difference in the set of values i.e. shared by different age groups is often called generation gap" [11].

In the context of team management the gap in efficiency seems particularly interesting. Stoner and Wankel perceive it as a determinant of quality. They note that it is "the difference between the tasks set out in the formulation of objectives and results, which is likely to be reached in the case of the continuation of the current strategy. Efficiency gap may be the result of determining more difficult tasks or result from the designation of insufficient effectiveness in the past due to the effective prevention of competition, changes in the environment or loss of resources, or because of the failure to rethink the strategy. The higher the efficiency gap is, the greater should be the change in strategy" [12].

Unfortunately, due to exponential growth of information, the processes of globalization and competition, the issue of information management becomes complex and requires appropriate tools, techniques and technologies. It is difficult not to agree with the authors Grudzewski and Hajduk, who define the technical and technological gap as relatively permanent difference of the technology potential in the various national economies, determined by measuring the difference in the levels of the creation of products and their production. The change of situation in which technical and technological gap is eliminated will require additional resources for B+R and reorientation of the existing approach toward the economy to a modern approach.

VI Conclusion

Issues related to the information gap constitute a multifaceted and complex research problem, which can be analyzed from different perspectives. Multidimensionality of the concept of an information gap and the meta-informational gap raises difficulties both in its explicit definition and interpretation as well as in the process of measuring. Understanding the information gap helps in determining the actions involving its restriction, and such actions are necessary even in the context of macroeconomic.

Identification of information gap is definitely a dynamic process. While comparing fourteen gaps defined in multidimensional information expanse on Figure 1, the existence of only seven gaps was signalized, that is development, competence, motivational, emotional, generation, effective and technical-technological ones. In economic practice exposing their utility requires, in the case of all economic subjects, an individual precision. It may even turn out that other dimensions will be needed. Only after considering necessary modifications one can gather the ICT tools, competences and methodology of development keeping in mind the philosophical aspect The concept of development with the element of evaluation is superior in relation to the idea of development.

The authors have collaborated in formulating such an approach toward design and production for businesses from Pomerania. The effectiveness of this approach is not yet tested and focuses on the tools of integrated computer systems such as CAD/CAM/CAE for employees working on engineering and technical positions.

References

[1] W. M. Grudzewski, I. K. Hejduk, Zarządzanie technologiami. Zaawansowane technologie i wyzwanie ich komercjalizacji. Difin, Warszawa, 2008, p.29.

[2] J. Sala, H. Tańska, „Kwalifikacje społeczeństwa informacyjnego",. In: Problemy społeczeństwa informacyjnego, A. Szewczyk, ed. vol I, Uniwersytet Szczeciński Wydział Nauk Ekonomicznych I Zarządzania Instytut Informatyki w Zarządzaniu, Wydawnictwo Printshop, Szczecin, 2007, p.108-115.

[3] J. Sala, H. Tańska, „Wykluczenia w globalnym społeczeństwie informacyjnym w kontekście niedorozwoju rynku usług telekomunikacyjnych i pocztowych", in: Funkcjonowanie rynku telekomunikacyjnego i pocztowego w warunkach postępującej elektronizacji gospodarki, H. Babis, ed., Uniwersytet Szczeciński Wydział Zarządzania i Ekonomiki Usług Katedra Ekonomiki I Organizacji Telekomunikacji Katedra Polityki Gospodarczej Łączności, Zeszyty naukowe, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin, 2007.

[4] Innovative activity of enterprises in 2008 – 2010. 2011. CSO (GUS) SO in Szczecin, Warszawa, p.149.

[5] J. Sala, H. Tańska, „Rozwiązania instytucjonalne na rzecz transferu wiedzy i kompetencji", in Zeszyty Naukowe Politechniki Łódzkiej, vol. 53, Łódź, 2013.

[6] J. Oleński, Elementy ekonomiki informacji. Katedra informatyki Gospodarczej I Analiz Ekonomicznych, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski, Warszawa, 2000, pp.497-498.

[7] A. Rokicka-Broniatowska, Wstęp do informatyki gospodarczej. (red.), Szkoła Główna Handlowa, Warszawa, 2006, p.148.

[8] J. Sala, H. Tańska, „Kształcenie kadr dla potrzeb gospodarki elektronicznej", in: Współczesne aspekty informacji, Monografie I Opracowania, vol. 551, ‚J. Goliński, K. Krauze, ed., Szkoła Główna Handlowa w Warszawie, Warszawa, 2008, pp.299-307.

[9] R. .Rutka, M. Czerska, Metoda identyfikacji zdolności systemu motywacyjnego do kreowania zachowań oczekiwanych przez pracodawcę, in Zeszyty Naukowe Politechniki Łódzkiej, vol. 51, Łódź, 2013, p.77.

[10] J. Strużyzna, „Puzle „bycia zatrudnionym" – wyzwania nowego HRM dla menedżerów", in Zeszyty Naukowe Politechniki Łódzkiej, vol. 51, Łódź, 2013, p. 43.

[11] J. A. F. Stoner, Ch. Wankel, Kierowanie. PWE, Warszawa, 2011, p.447.

[12] J. A. F. Stoner, Ch. Wankel, Kierowanie. PWE, Warszawa, 1992, p.112.

DOI: 10.15439/2014F462
ACSIS, Vol. 3

Information Models and Methods of the University's Scientific Knowledge Life Cycle Support

Zhomartkyzy Gulnaz
D. Serikbayev East Kazakhstan
State Technical University,
69 Protozanov A.K.
Ust-Kamenogorsk, Kazakhstan
Email: zhomartkyzyg@gmail.com

Balova Tatiana
D. Serikbayev East Kazakhstan,
State Technical University,
69 Protozanov A.K.
Ust-Kamenogorsk, Kazakhstan
Email: tbalova@ektu.kz

Milosz Marek
Lublin University of
Technology,
36b Nadbystrzycka,
Lublin, Poland
Email: m.milosz@pollub.pl

***Abstract*—The main aim of this work is to develop methods and technologies of the university's scientific knowledge management. The paper examines the concept of knowledge management and life-cycle processes of the university's scientific knowledge. Text Mining and Semantic Web technologies are used to develop the ontological information model and for information resources processing. The paper describes the developed information model of the university's scientific knowledge, the methods of forming scientific profiles, and the concept of the university's scientific knowledge semantic portal.**

I. Introduction

KNOWLEDGE , intellectual property, and intellectual resources have been understood in recent decades as a major driving force of the economy of the "third wave" (the new economy), the economy based on knowledge of expanded reproduction [1]. Knowledge as intellectual capital is gradually becoming one of the most important factors in the development of economy and the society.

For a modern institution of higher education as an open social and economic self-organizing system the processes of creation, accumulation, and dissemination of knowledge is becoming a key factor for training competitive specialists.

Paraphrasing [2], we can determine that the university's scientific knowledge as the combination of data and information with added opinions, skills and experience of the university's scientists and faculty members is a valuable asset that can provide an advantage in the market of educational services in project activities, scientific and practical work, and innovative activities.

The intellectual capital or intangible assets of the university are the source of new scientific knowledge. Edvinsson developed a hierarchical structure of intellectual capital of a higher education institution - a "Skandia Value Scheme" model [3]. The main components of this model are human assets (the capital) and innovating capital which are tacit knowledge.

Tacit knowledge form the human capital which is embodied in the university staff as a body of knowledge, qualifications, each employee's innovation, as the system of values, culture and philosophy of the institution. It is believed that practical tacit knowledge is the key to decision making and management.

There is a continuous exchange between explicit and implicit knowledge and their transformation. [4]. Nonaka I., Takeuchi H. suggested a cyclical process of knowledge transformation: socialization, externalization, combination, and internalization.

Kantner defines the concept of knowledge management as a strategy for the organization and the process of knowledge transformation [5], [6]. Currently, there is the increasing number of research papers devoted to the issue of the community of practitioners' knowledge management and the implementation of individual processes of knowledge transformation. [7], [8].

The European concept of knowledge management [2] identifies five processes of knowledge life cycle: knowledge identification, creation, storage, distribution and use. Knowledge life cycle actually reflects the methods and technologies of knowledge management at each technological step.

The purpose of the university's knowledge management is to improve human and innovative capital, to accelerate its development and competitiveness in the market of educational services in research, practical and innovative activities.

The effective use of the university's intellectual capital directly depends on the support of possibilities for knowledge creation, storage, dissemination and use. [9], [10].

The importance of developing knowledge management systems (KMS) is due to the fact that the knowledge which is disseminated, acquired and exchanged generates new knowledge [11].

In this regard, it is important to determine what knowledge management system must be created and what transformations must be implemented to use the existing intellectual capital successfully. There arises a need in processes, infrastructure and organizational procedures at a higher education institution that would allow its employees to use its corporate knowledge base. This paper examines some models and methods of the university's scientific knowledge life cycle support.

II. THE UNIVERSITY'S SCIENTIFIC KNOWLEDGE MANAGEMENT SYSTEM

Knowledge management in an enterprise is the systematic process of identification, use, and transfer of information and knowledge which people can create, improve and apply [12].

It is the process by which the enterprise generates knowledge, accumulates and uses it to gain a competitive advantage [13, 14].

In this paper the university's scientific knowledge management system (SKMS) is considered as an aggregate of information, software, technical means, and organizational solutions aimed at efficient management of the university's available intellectual resources and training specialists who meet the modern requirements. The purpose of SKMS at the university is the formation of a unique ontology-based integrated intellectual environment to improve the competitiveness of the university's science and education. The university's SKMS is the technological component of the university's SKM, which provides the creation, organization and dissemination of scientific knowledge among the university's staff.

The main functions of scientific knowledge management at the university (SKM) are consistent with the university's functions defined in [15]. They are divided into analytical, integration and new knowledge generation.

The analytic functions of SKM include:

- the search of knowledge in the information flow, content filtering;
- identification and classification of existing knowledge according to certain criteria, the formation of the staff's scientific profile, and the monitoring of the university's of scientific schools development.

The SKM integrative function provides:

- the introduction of classified knowledge in the corporate memory and evaluation of its integration with educational programs implementation;
- the extraction of knowledge from the corporate memory by sharing the knowledge between departments, different levels of management, as well as the exchange of expertise and experience of the staff;
- ensuring the accessibility of knowledge for management decisions making, search for ideas, generating ideas, and training.

The function of new knowledge creation provides the fixation of explicit and tacit knowledge in the university's scientific knowledge base.

The life cycle of scientific knowledge is shown in Figure 1.

The implementation of SKMS or its components at the university will allow users:

- to receive the right information at the right place at the right format and in a timely manner with minimal effort;
- to reduce the number of human errors and to increase the quality of decisions;
- to improve communication, reduce the information loss and distortion;
- to stimulate the sharing of knowledge and best practices.

The process-oriented On-To-Knowledge methodology [9] is used as the basis for the university's knowledge management.

There are following approaches to knowledge management: organizational, technological and ecological [16]. The technological approach puts the application of IT-technologies in line with the organizational measures. The model of technological approach to knowledge management is shown in Figure 2.

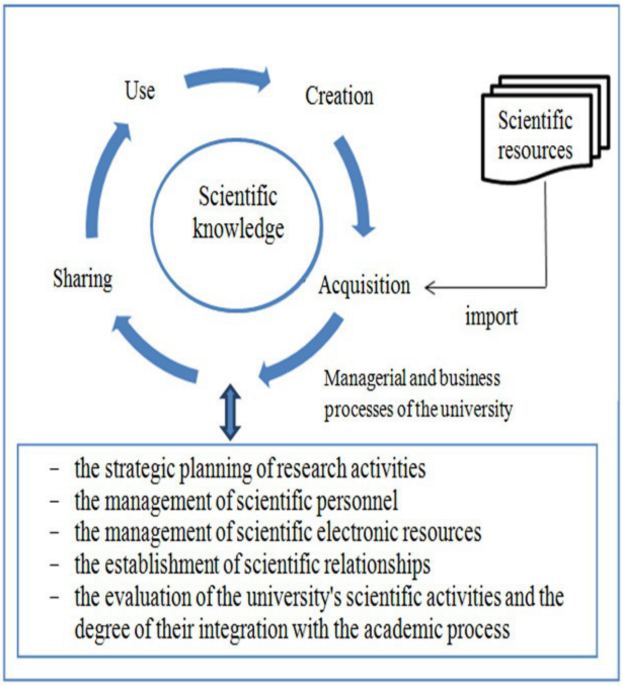

Use
Creation
Scientific resources
Scientific knowledge
Sharing
Acquisition
import
Managerial and business processes of the university
- the strategic planning of research activities
- the management of scientific personnel
- the management of scientific electronic resources
- the establishment of scientific relationships
- the evaluation of the university's scientific activities and the degree of their integration with the academic process

Fig. 1. The life cycle of scientific knowledge

The introduction of knowledge management and its related processes of developing and maintaining the SKM at an enterprise usually involve working with unstructured information resources.

The accumulation of knowledge is a complex process involving different work in the cycle of knowledge transformation [5]. The stage of knowledge acquisition includes:

- the knowledge acquisition by analyzing documents (Text Mining) and databases (Data Mining);
- metadata annotating / creation;
- the extraction of the employees' tacit knowledge;
- structuring / classification;
- the formation of organizational memory, knowledge integration and storing.

The use of knowledge suggests that the available knowledge is used by the university staff to perform their jobs more efficiently, and newly created knowledge affects both scientific and educational activities.

The university's SKMS integrates intellectual resources, knowledge management tools and processes of knowledge transformation. The general structure of the university's SKMS is shown in Figure 3.

"The university's intellectual resources" component in the overall structure of SKM determines human and structural resources with their related formalized human capital and innovative capital of the university. The sources of scientific knowledge resources are the electronic version of the university's scientific journal, the electronic version of the conference

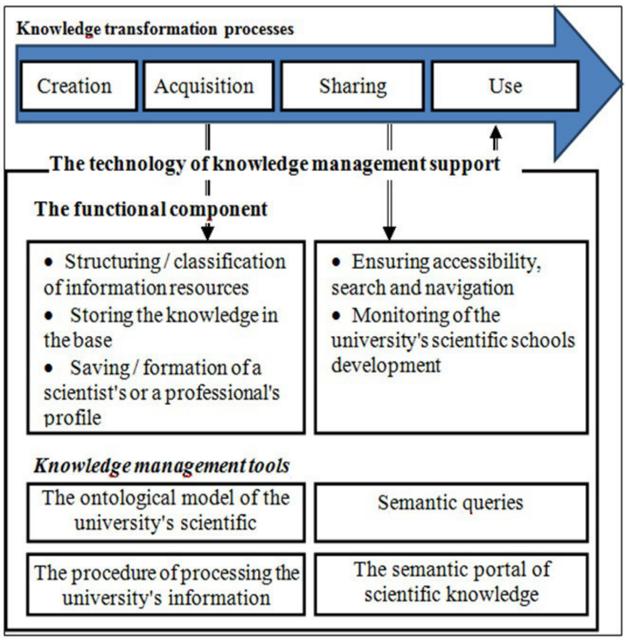

Knowledge transformation processes
Creation
Acquisition
Sharing
Use
The technology of knowledge management support
The functional component
• Structuring / classification of information resources
• Storing the knowledge in the base
• Saving / formation of a scientist's or a professional's profile
• Ensuring accessibility, search and navigation
• Monitoring of the university's scientific schools development
Knowledge management tools
The ontological model of the university's scientific
Semantic queries
The procedure of processing the university's information
The semantic portal of scientific knowledge

Fig. 2. The model of technological approach to knowledge management

proceedings, a bibliographic database "Irbis".

Knowledge processing and transformation consists of the following processes:

- creation: defining of tacit knowledge (which includes the university staff's knowledge) and of explicit knowledge (in the form of paper or electronic documents or records);
- accumulation, which includes: the ontology as a conceptual framework for describing knowledge resources and a set of methods for the formation of the university's scientific knowledge base;

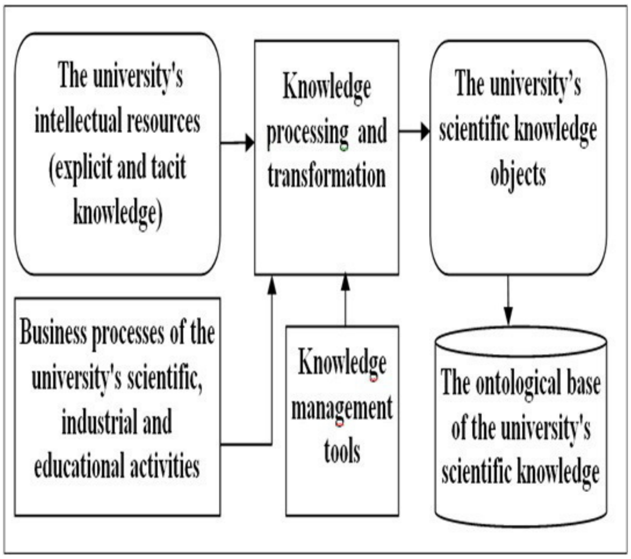

The university's intellectual resources (explicit and tacit knowledge)
Knowledge processing and transformation
The university's scientific knowledge objects
Business processes of the university's scientific, industrial and educational activities
Knowledge management tools
The ontological base of the university's scientific knowledge

Fig. 3. The general structure of the university's SKMS

- sharing: providing users with opportunities for semantic search and navigation;
- tacit knowledge sharing which is intended for communication in scientific community and new knowledge formation. The participants of tacit knowledge sharing can be individual employees and individual research groups;
- using knowledge to improve the efficiency of research activities at the university.

Business processes at the university include:

- business processes of development: the development of innovative learning technologies; participation in research grants; improving academic qualifications of employees, establishing scientific relationships with companies and enterprises;
- the processes of postgraduate education: library resources management, electronic information resources management;
- the university management processes: personnel management processes and tools for personnel development, strategic planning of the university's research activities organizing knowledge sharing in knowledge networks;
- management of research activities and scientific and production activities;

Knowledge management tools to support the processes of transformation of scientific knowledge are listed below:

- information technology;
- organizational and administrative mechanisms;
- corporate culture;
- technical infrastructure;
- legal aspects.

IT knowledge management tools consist of a set of information technology, providing targeted development and effective functioning of the processes of transformation and networks of scientific knowledge.

III. THE INFORMATION MODEL OF THE UNIVERSITY'S SCIENTIFIC ACTIVITIES MANAGEMENT

Ontology, as a common language in knowledge management, is a conceptual domain model as a system of concepts, their properties and relations [17].

The information model of the university's knowledge can be described as the ontology which includes the basic concepts of the university's scientific activities, such as organizational structure, subjects, the objects of scientific schools and research, information resources, other subdisciplines, etc. [18].

In the ontology of scientific activities in the "Research directions" class the subclasses correspond to the major research areas of the university.

The subclasses correspond to major research areas in the ontology of scientific activity in the "Research directions" class.

The ontology, as a common language in knowledge management, is a conceptual domain model as a system of concepts, their properties and relations. The use of ontology in knowledge management system makes it possible:

- to integrate the information distributed in various document repositories, databases and knowledge;
- to generalize and systematize the available information, acting as a metamodel;
- to use the automated logical conclusion for better search results, acquiring new knowledge and analyzing information;
- to use more effective mechanisms to receive, visualize and search for knowledge.

The ontological information model of knowledge database supports semantic queries in SPARQL and SPARQL-DL.

Example: semantic query for researchers and information resources in scientific directions can be written as follows:
Example 1:
Person and (peopleHasPublicationIR some (PublHasDivis some TopicsSolidStatePhysics))
Example 2:
Article and (PublHasDivis some TopicsSolidStatePhysics)

Navigation in scientific knowledge and information resources is done by the use of semantic links between the classes of the ontology.

IV. THE PROCEDURE OF THE UNIVERSITY'S INFORMATION RESOURCES PROCESSING WITH THE PURPOSE TO FORM SCIENTIFIC PROFILES

A. The Stages of the University's Information Resources Processing

The main stages of the information resources processing are given below:

1) Extraction of terminological collocations. Pearson criterion is used to detect collocations [19];
2) Feature selection. Mmutual Information method is used as a method for evaluating the importance of terms (Mmutual Information) [20];
3) Classification of texts according to scientific areas. The method of k nearest neighbour (kNN) is used for text classification [20], [21];

Working with the text files in the corpus with the purpose to perform statistical calculations requires the following preliminary steps:

1) To pre-translate to .txt format the files of different formats (pdf, doc,. docx) in the corpus;
2) To delete all hyphenation beforehand;
3) To perform lemmatization of all the text files in the corpus, to delete all punctuation marks, to change all uppercase letters to lowercase letters.

The corpus of documents for processing has been compiled from articles in various fields published in the "Solid State Physics" journal. The founders of the journal are the Russian Academy of Sciences, the Department of General Physics and Astronomy of RAS, Ioffe Physical-Technical Institute of the Russian Academy of Sciences [22].

A detailed description of processing stages is given in the following sections.

B. Collocation Extraction and Feature Selection for the Classification of Scientific Texts

A collocation is regarded as a non-random combination of two or more lexical items common to most scientific texts in a particular scientific field. The set of terminological collocations generated by the specified collection of scientific texts describes a narrow subject area (topics and subtopics) of this collection.

For automatic extraction of terminology collocations from scientific texts a freely distributable Java-library LingPipe interface is used [19]. The array of obtained collocations is ranked in order of importance, where the sequence of lexical tokens is dependent. The significance of the collocations is calculated based on the collocation the Pearson independence statistics. The higher the value of the significance of the collocation, the less the likelihood that the sequence of tokens is independent.

The general scheme of the formation of a software dictionary is shown in Figure 4.

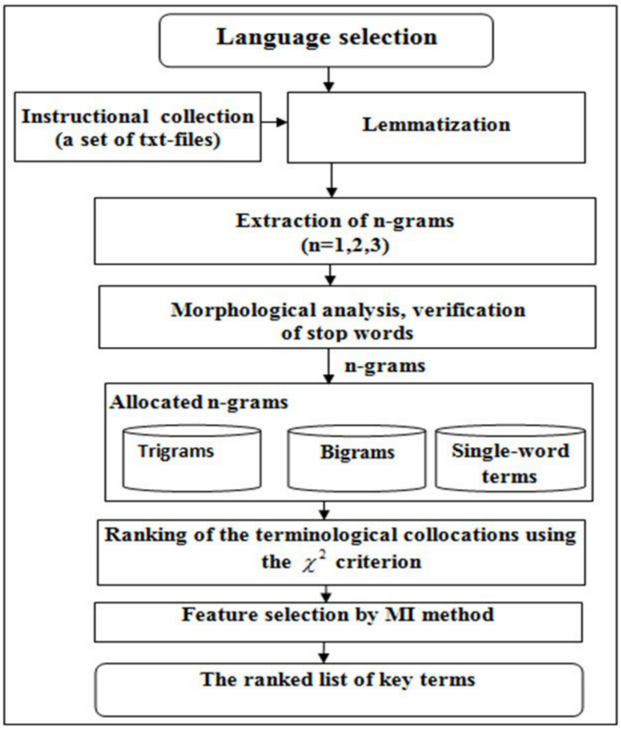

Language selection
Instructional collection (a set of txt-files)
Lemmatization
Extraction of n-grams (n=1,2,3)
Morphological analysis, verification of stop words
n-grams
Allocated n-grams
Trigrams
Bigrams
Single-word terms
Ranking of the terminological collocations using the χ^2 criterion
Feature selection by MI method
The ranked list of key terms

Fig. 4. The extraction and selection of domain terms

The main modification of the method based on the static approach includes the preliminary use of morphological templates of filters [23], [24].

To obtain the list of dominant terms using the χ^2 it is necessary to solve the following tasks:

- the extraction of collocations with the calculated coefficient of significance;
- the determination of the morphological characteristics of each word in the n-gram;

- the removal of stop words and the selection of phrases that match the templates;
- the saving of collocations in a database table.

The following restrictions were set for a bigram and a trigram: the minimum frequency bigram equal to 10, the minimum frequency of trigrams equal to 15.

The thus obtained term-candidates form a list of n-grams (bigrams, trigrams).

Single-word terms are extracted based on a combination of frequency and the inverse document frequency of the term. The weight of a single-word term is calculated by the formula [20]:

$$Tf - Idf_{t,c} = tf_{t,c} \times log\frac{N}{df_t} \tag{1}$$

where $tf_{t,c}$ is term frequency in the collection of the c class; df_t is the number of documents in the collection of the c class which contain the term; N is the number of documents in the collection.

The generated list of terms with weights $Tf - Idf_{t,c}$ is ranked by a certain threshold value, a number of terms are selected which are further recorded in the database table.

Table 1 presents the results of the developed module for extraction and separation of terms according to domains.

Table 1 shows some uninformative words with high critical value of χ^2 (such as the final edition, viewpoints, etc.), as well as some informative words with a lower value χ^2 (a superconducting property, a superconducting parameter).

The further stage of the vocabulary formation is the selection of features to eliminate noise-terms. Feature selection enhances the effectiveness of training the classifier by reducing the size of the vocabulary and the classification accuracy. The measure of utility $A(t,c)$ of each term in the lexicon is calculated for each class, and N terms with the largest value of $A(t,c)$ are selected. All other terms are discarded and are not involved in the classification.

To remove non-informative terms the method of mutual information was chosen [20]. The measure of mutual information estimates how much information about a class in information-theoretic sense the term includes. The measure the usefulness $MI(t_k, c)$ is calculated, and k terms with the highest values of this measure are selected. To select k terms $t_1, ..., t_k$ for a given class, the following formula is used:

$$MI(t_k, c) = log_2\frac{A \times Q}{(A+C) \times (A+B)} \tag{2}$$

where: A is the number of documents which belong to category c and contain term t; B is the number of documents, which do not belong to category c and contain term t; C is the number of documents which belong to category c and do not contain term t; Q is the instructional the training set of documents.

The results of applying the mutual information method for the selection of features obtained in the previous step are shown in Table 1.

TABLE I
THE COMPARISON OF MUTUAL INFORMATION VALUES AND χ^2 OF TERMS FOR THE "SOLID STATE PHYSICS" DOMAIN

Terms	**Critical value χ^2**	**Value MI**
final version	40462,68	0,093
point of view	23093,13	0,100
superconducting granule	15534,01	1,000
intercrystalline boundary	14328,13	1,000
the first turn	13659,12	0,212
high-temperature superconductor	11518,91	1,000
superconducting transition	6566,12	1,000
phase transformation	4703,25	1,000
the object of study	4413,52	0,415
volume ratio	3584,50	0,263
the discussion of results	3196,66	0,553
at present	3175,12	0,263
ternary alloy	2434,56	1,000
metallic conductivity	2288,77	1,000
mentioned above	2243,87	0,652
amorphous film	1910,57	1,000
maximum value	1832,25	0,049
the system of equations	744,95	0,000
superconducting state	665,31	1,000
electronic spectrum	460,36	1,000
superconducting property	256,72	1,000
superconducting parameter	144,78	1,000

As illustrated in Table 1, some terms with low rates of χ^2 (superconducting parameter and superconducting properties) have a high value of MI. At this stage it is necessary to perform a selection of informative terms weighted by their mutual information MI, which are selected in the domain vocabulary and then used for the text classification.

C. The Formation of Scientific Profiles Based on Classifications of Information Resources in Scientific Domains

For the classification of scientific resources kNN- classification is used. The classification task in machine learning is a task to assign an object to one of the predefined classes based on its formal characteristics. kNN method (k method of nearest neighbor) is a vector classification model. kNN classifier assigns the document to the prevailing class of nearest neighbors (Figure 5), where k is the method parameter. The k parameter in kNN method is often selected on the basis of experience or knowledge about the classification task at hand.

As a result of the university's scientific resources processing the documents' profiles are formed [25]. A document profile is defined as the vector of all relevant topics of its ontology:

$$PD(d) = (R_1^d, ..., R_c^d) \tag{3}$$

where: R_c^d are relevant topics c of document d.

Accordingly, the academic profile of a staff member is defined as the profile of all his publications:

$$PD(a) = (R_1^{da}, ..., R_i^{da}) \tag{4}$$

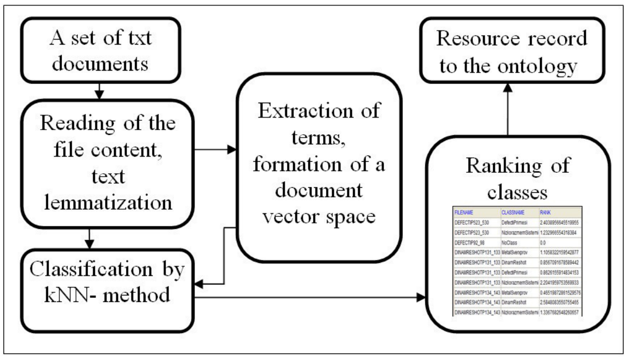

A set of txt documents
Reading of the file content, text lemmatization
Classification by kNN- method
Extraction of terms, formation of a document vector space
Resource record to the ontology
Ranking of classes

Fig. 5. The scheme of classification algorithm

where: R_i^{da} - are all the documents of the author. The final step of the text classification is the formation of the document's semantic profile by creating the individuals of "Information resources" class in the ontology of scientific research activity.

The classifier is written in Java, such tool sets as LingPipe, Apache Lucene (free Java library for text processing and high-speed full-text search) are used for further text processing. The classification results are $k-$nearest neighbors ranked classes, the parameter k is equal to 5.

V. The Concept of the Semantic Portal of the University's Scientific Knowledge

Previously considered models and methods formed the basis for the semantic portal of the university's scientific knowledge.

Operational intelligent query processing and adaptability to a user's needs is the basic idea of the functioning of the portal which is being developed. Therefore, there is a need to automate a time-consuming search and analysis of data in the process of creating and maintaining the university's scientific knowledge base.

The semantic portal platform architecture is shown in Figure 6.

To ensure the systematization of scientific knowledge and information resources the university's semantic portal supports the following functions:

- the software for navigation through the ontology of the university's scientific knowledge;
- the organization of search queries on the ontology concepts and relations;
- the classification of information resources to determine the development of the university's scientific schools and directions.

For faster data it was decided to load the ontology schema into an intermediate RDF-store of TDB built into Jena solution which supports all the features of work with JenaAPI, including preparation of SPARQL queries. The RDF-store contains the description of scientific knowledge ontology in the form of RDF-triples. This description corresponds to the graph which nodes are the subjects and objects of RDF-predicates, and the ribs are the RDF-predicate itself.

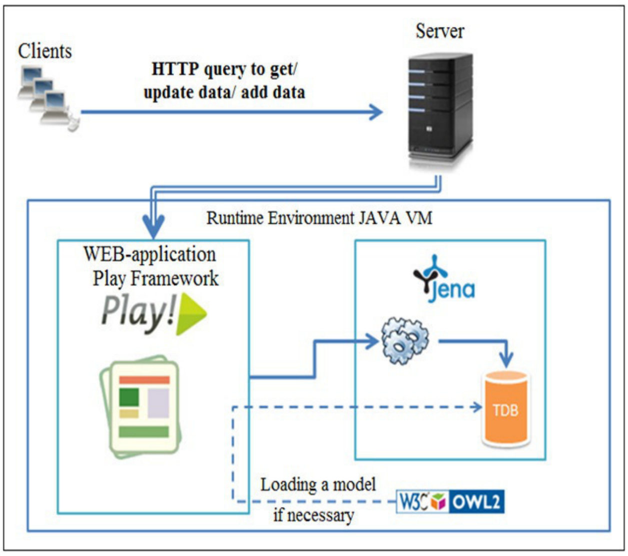

Clients
Server
HTTP query to get/ update data/ add data
Runtime Environment JAVA VM
WEB-application Play Framework
Play!
Jena
TDB
Loading a model if necessary
W3C OWL2

Fig. 6. The architecture of the semantic portal software platform of e-University's scientific knowledge

The main components of the semantic portal of the e-university's scientific knowledge are: the ontology of scientific knowledge, the ontology editor, the module of information resources classification and indexing, the module of navigation and search through the portal content, the database of ontological information.

The portal's architectural components provide the user with a transparent semantic access to the necessary data. User queries are handled by the server applications which are associated with the semantic components. Remote clients work with the portal in all modern browsers using HTTP protocol. Queries are sent to the web application server. In the JavaVM (Virtual Machine) run-time environment stream starts query processing.

When data is obtained using SPARQL or Jena API abstract model of data access, Jena TDB store (RDF-triples store) is accessed; the insert operations (delete, update) are performed, and the answer to the user is generated.

VI. Conclusion

This paper examines some models, methods, and technologies to support life cycle processes of the university's scientific knowledge. The ontology of the university's research activities is used as the information model of SKMS. The paper describes the procedure of the university's information resources processing. The thematic classification of documents based on the developed procedures for handling information resources allows forming of employees' scientific profiles and realizing a personalized search engine for the university's scientific knowledge semantic portal. The semantic concept of the university's scientific knowledge semantic portal is described. The software implementation of the semantic portal allows searching for any ontology object according to the following classes: researchers, research areas of the university, events, key terms, organizations, departments, subdepartments, university publications. The pilot project of the university's scientific activity semantic portal has allowed us to generate a fragment of the university's scientific knowledge base.

The work was performed under grant "The development of an e-university's ontological knowledge base", state registration number 0213RK00305.

References

[1] A. Toffler, *The Third Wave.* Moscow, 2002.

[2] European Guide to good Practice in Knowledge Management, Part 1:Knowledge Management Framework, 2004, http://enil.ceris.cnr.it/Basili/EnIL/gateway/europe/CEN_KM.htm

[3] L. Edvinsson, "Developing intellectual capital at Skandia," *Long Range Planning. J.,* vol. 30(1), 1997, pp. 366 - 373, http://dx.doi.org/10.1016/S0024-6301(97)90248-X

[4] I. Nonaka and H.Takeuchi, *Company - creator of knowledge. Origin and development of innovation in Japanese firms.* Moscow: Olimp-Business, 2003.

[5] H. Zaim, "Performance of Knowledge Management Practices: a causal analysis," *Knowledge Management. J.,* vol. 11, No.6, 2007, pp. 54-67, http://dx.doi.org/10.1108/13673270710832163

[6] J. Kantner, "Knowledge Management, Practically Speaking," *Information System Management. J.,* vol. 16(4), 1999, pp. 7-15, http://dx.doi.org/10.1201/1078/43189.16.4.19990901/31198.2

[7] Yuh-Jen Chen, Yuh-Min Chen, Meng-Sheng Wub, "An empirical knowledge management framework for professional virtual community in knowledge-intensive service industries," *Expert Systems with Applications. J.,* vol. 39, 2012, pp. 13135 - 13147, http://dx.doi.org/10.1016/j.eswa.2012.05.088

[8] K. Stock, T. Stojanovic, F. Reitsma, Yang Oue, M.Bishr, J.Ortmann, A. Robertson, "To ontologise or not to ontologise: An information model for a geospatial knowledge infrastructure," *Computers-Geosciences. J.,* vol. 45, 2012, pp. 98 - 108, http://dx.doi.org/10.1016/j.cageo.2011.10.021

[9] S. Staab, H-P. Schunurr, R. Studer, Y. Sure. "Knowledge processes and ontologies," *IEEE Intelligent Systems. J.,* vol. 16(1), 2001, pp. 26 - 34, http://dx.doi.org/10.1109/5254.912382

[10] D.V. Kudryavtsev, *Knowledge management systems and the use of ontologies.* St. Petersburg. Univ Polytechnic. University Press, 2010.

[11] R. Maier, *Knowledge Management Systems.* Information and Communication Technologies for Knowledge Management, Springer, 2007.

[12] K. Hafeez and H. Abdelmegid, "Dynamics of human resource and knowledge management," *Operational Research Society. J.,* vol. 54, 2003, pp. 153-164, http://www.jstor.org/stable/4101606

[13] T.A. Gavrilova, *Knowledge Engineering. In Innovative development: economy, intellectual resources, knowledge management.* Moscow: INFRA-M, 2009.

[14] M. Alavi and D. E. Leidner, "Knowledge management and knowledge management systems: Conceptual foundations and research issues," *MIS Quarterly. J.,* vol. 25(1), 2001, pp. 107 - 136, http://dx.doi.org/10.2307/3250961

[15] L.A. Trofimova and V.V. Trofimov, *Knowledge Management.* St. Petersburg, 2012.

[16] A.F. Tuzovskii. "Developing knowledge management systems based on a single ontological knowledge base," *In Bulletin of the Tomsk Polytechnic University. J.,* vol. 310(2), 2007, pp. 182 - 185.

[17] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologist.* Burlington, USA, 2011.

[18] Y. Zagorulko. "Information model of scientific knowledge portal," *Information Technology. J.,* vol. 12, 2009, pp. 2 - 7.

[19] Alias LingPipe, http://alias-i.com/lingpipe.

[20] Ch.D. Manning, P. Raghavan and H. Schutze, *Introduction to Information Retrieval.* Cambridge University Press, 2009, http://nlp.stanford.edu/IR-book/

[21] H. Altncay and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," *In Proceedings of the Pattern Recognition Letters,* 2010, pp. 1310 - 1323, http://dx.doi.org/10.1016/j.patrec.2010.03.012

[22] Science journal "Solid State Physics", http://journals.ioffe.ru/ftt/.

[23] Multiword Recognition and Extraction, http://www.ilc.cnr.it/EAGLES96/rep2/node38.html.

[24] D.S. Novikov. Automatic allocation of the terms of the texts subject areas and linkages between them. In Information and telecommunication technologies and mathematical modeling of high-tech systems in 2012. RUDN; Russia, http://conf.sci.pfu.edu.ru/index.php/ittmm/2012/paper/view/245l.

[25] K.V. Kryukov, O.P. Kuznetsov and V.S. Suhoverov, "On the notion of formal competence researchers," *In Proceedings of the III International Scientific and Technical Conference - OSTIS-2013,* Minsk, 2013, pp. 143-146, http://www.conf.ostis.net/index.php?title=OSTIS-2013

Software Systems Development & Applications

SSD&A is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the discipline of software engineering. The SSD&A area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This area investigates both established traditional approaches and modern emerging approaches to large software production and evolution.

Events that constitute SSD&A are:

- ATSE'14 - 5th International Workshop Automating Test Case Design, Selection and Evaluation
- MDASD'14 - 3rd Workshop on Model Driven Approaches in System Development

3rd Workshop on Model Driven Approaches in System Development

FOR many years, various approaches in system design and implementation differentiate between the specification of the system and its implementation on a particular platform. People in software industry have been using models for a precise description of systems at the appropriate abstraction level without unnecessary details. Model-Driven (MD) approaches to the system development increase the importance and power of models by shifting the focus from programming to modeling activities. Models may be used as primary artifacts in constructing software, which means that software components are generated from models. Software development tools need to automate as many as possible tasks of model construction and transformation requiring the smallest amount of human interaction.

A goal of the proposed workshop is to bring together people working on MD languages, techniques and tools, as well as Domain Specific Languages (DSL) and applying them in information system and application development, databases, and related areas, so that they can exchange their experience, create new ideas, evaluate and improve MD approaches and spread its use. The intention is to target an interdisciplinary nature of MD approaches in software engineering, as well as research topics expressed by but not limited to acronyms such as Model Driven Software Engineering (MDSE), Model Driven Software Development (MDSD), and OMG's Model Driven Architecture (MDA).

1st Workshop on MDASD was organized in the scope of ADBIS 2010 Conference, held in Novi Sad, Serbia. From this year, MDASD becomes a regular bi-annual FedCSIS event.

Topics

Submissions are expected from, but not limited to the following topics:

- MD Approaches in System Design and Implementation – Problems and Issues
- MD Approaches in Software Process Models
- MD Approaches in Databases and Information Systems
- MD Approaches in Software Quality and Standards
- Metamodeling, Modeling and Specification Languages
- Model Transformation Languages
- Model-to-Model, Model-to-Text, and Model-to-Code Transformations in Software Process
- Transformation Techniques and Tools
- Domain Specific Languages (DSL) and Domain Specific Modeling (DSM) in System Specification and Development
- Design of Metamodeling and Modeling Languages and Tools
- MD Approaches in Requirements Engineering and Business Process Modeling
- MD Approaches in System Reengineering and Reverse Engineering
- MD Approaches in HCI development
- MD Approaches in GIS development
- MD Approaches in Document Engineering
- Model Based Software Verification
- Theoretical and Mathematical Foundations of MD Approaches
- Organizational and Human Factors, Skills, and Qualifications for MD Approaches
- Teaching MD Approaches in Academic and Industrial Environments
- MD Applications and Industry Experience

Event Chair

Luković, Ivan, University of Novi Sad, Serbia

Steering Committee

France, Robert, Colorado State University, USA, United States

Mernik, Marjan, University of Maribor, Slovenia

Ristić, Sonja, University of Novi Sad, Serbia

Tolvanen, Juha-Pekka, MetaCase, Finland

Program Committee

Amaral, Vasco, The New University of Lisbon, Portugal

Bryant, Barrett, University of North Texas, United States

Budimac, Zoran, Faculty of Sciences, Univ. of Novi Sad, Serbia

Chen, Haiming, Chinese Academy of Sciences, China

Erradi, Mohammed, ENSIAS, Mohammed-V Souissi University, Morocco

Fertalj, Krešimir, University of Zagreb, Croatia

France, Robert, Colorado State University, USA, United States

Gray, Jeff, University of Alabama, United States

Ivanović, Mirjana, University of Novi Sad, Serbia

Janousek, Jan, Czech Technical University, Czech Republic

João Varanda Pereira, Maria, Instituto Politecnico de Braganca, Portugal

Karagiannis, Dimitris, University of Vienna, Austria

Kardaş, Geylani, Ege University International Computer Institute, Turkey

Kollár, Ján, Technical University of Kosice, Slovakia

Kosar, Tomaž, University of Maribor, Slovenia

Krdzavac, Nenad, Michigan State University, United States

Kühne, Stefan, Universität Leipzig, Germany

Liu, Shih-Hsi Alex, California State University, United States

Maćoš, Dragan, University of Applied Sciences, Germany

Melo de Sousa, Simão, University of Beira Interior, Portugal

Mernik, Marjan, University of Maribor, Slovenia

Milosavljević, Gordana, University of Novi Sad, Serbia

Nešković, Siniša, University of Belgrade, Serbia

Porubän, Jaroslav, Technical University of Kosice, Slovakia

Rangel Henriques, Pedro, Universidade do Minho, Portugal

Ristić, Sonja, University of Novi Sad, Serbia

Seidl, Martina, Johannes Kepler University, Austria

Selic, Bran, Malina Software Co., Canada

Sierra Rodríguez, José Luis, Universidad Complutense de Madrid, Spain

Slivnik, Boštjan, University of Ljubljana, Slovenia

Suvajdžin-Rakić, Zorica, University of Novi Sad, Serbia

Tolvanen, Juha-Pekka, MetaCase, Finland

Wimmer, Manuel, Vienna University of Technology, Austria

DOI: 10.15439/2014F36
ACSIS, Vol. 3

Function-Oriented Inclusive Design

Markus Modzelewski
Universität Bremen, 28359
Bremen, Enrique-Schmidt-Straße
5, Germany
Email: modze@tzi.de

Michael Lawo
Universität Bremen, 28359
Bremen, Am Fallturm 1, Germany
Email: mlawo@tzi.de

Abstract—Technological advancements regarding functional capabilities of products affect product development processes. We observe the following: More functionality can be included in smaller devices. New devices are invented as hybrids between existing devices. Products can be individually adapted for end customers. Although the current product development processes already include contextual information about target customer groups and scenario of use, this information is strongly linked to single devices. We present a hierarchical superstructure above devices regarding functional capabilities able to categorize devices by functionality but also recommend devices for a set of functionalities.

I. Introduction

RECENT technological advancements emerged new opportunities in product design. Devices e.g. "phablets" include functional aspects related to mobile phone but also tablets in a single unit (Segan 2012). Automotive entertainment systems can be used to make a call or browse the internet (Zeller et al. 2001; BMW AG n.d.). With these new concepts, new usability issues appeared creating a burden especially for elderly customers regarding the need to learn how to use such devices (Clarkson et al. 2003; Langdon & Thimbleby 2010).

On the other hand, design guidelines and pre-experience of designers regarding usability issues are included in recommendation-driven product design as presented by the VICON project (Lawo et al. 2011; Modzelewski et al. 2012; Kirisci et al. 2012). The output of the VICON software solution ConVic[1] is extracted by a set of target user models, typical environments and typical tasks of a product (Modzelewski 2014). Design recommendations are presented defining qualitative and quantitative issues that should be included into product design increasing the awareness of designers about product customers. The ontology hierarchy used so far for storage and inference focuses on single devices (i.e. mobile phones). In this paper we extend the functional aspects including a selection of functionalities and inference of devices which can represent each service.

II. Existing Modifications in Product Development

According to VDI 2221 the product development process can be seen as a phase based sequence of states and outputs (VDI-Gesellschaft Entwicklung Konstruktion Vertrieb 1993). In the beginning the specification of requirements consists of a detailed analysis of user needs and functional aspects which are aimed for realisation by a product design. An analysis of a requirement list produces a functional structure for product decision. Product design consists of a draft phase in which first sketches (not function oriented) are constructed. In the subsequent CAD phase first virtual prototypes in virtual environments are produced which lead to prototypes. This enables an iterative process until the final product.

Customer involvement methods as seen by Strang and Linnhoff-Popien present different approaches how to include end customers into product design (Strang & Linnhoff-Popien 2004). The most frequently used Quality Function Deployment (QFD) method represents an analytical approach for first design phases with the involvement of end users (Akao 2004). A conversion of consumer demands into quality characteristics is utilized for an iterative deployment of a design quality function describing a "relation" between consumer and product. End customers are only involved during specification and concept phases but not for prototype testing, which leads to the categorization of QFD as a design for type. The types design with and design by have the pro of involvement of end customers in more phases with the con of being more cost-intensive (in addition to other pros and cons).

If customer involvement does not focus on device structures but functions as presented by this paper, new possibilities appear regarding design creativity but also the handling and integration of user needs aiming (a more) inclusive design (Coleman & Lebbon 2005; Dong et al. 2004; Newell & Gregor 2000).

[1] Available at SourceForge (http://sourceforge.net/projects/convic)

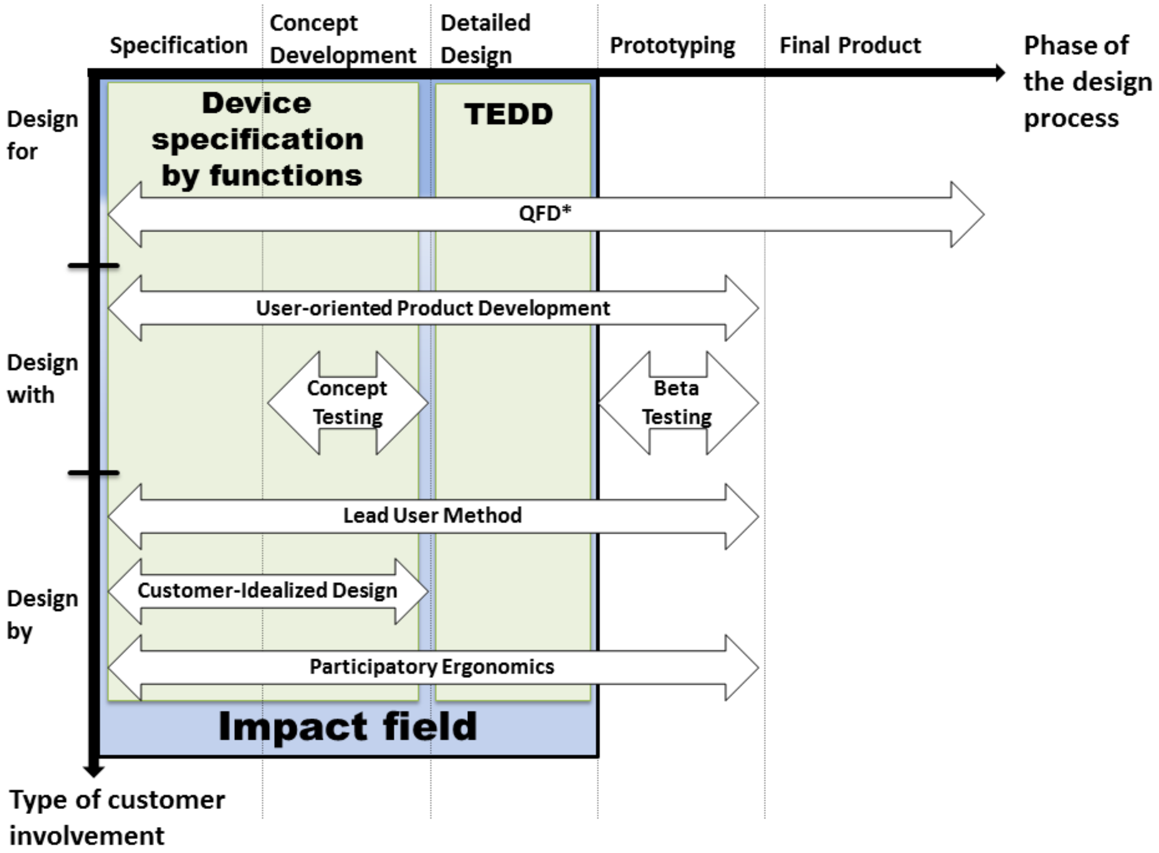


Figure 1: Impact field of function oriented approach on existing customer involvement methods.

- Quality function deployment

In QFD end users of a product are only involved before specification by forming relevant parameters and requirements of the product. If the design is not limited by a product but rather on functions, individual needs and capabilities of users can be integrated for more suitable products. As a result, inventions of new products based on end user requirements can be specified.

- User-oriented Product Development

This method focuses upon involvement of end customers after first concepts by direct involvement (design with users). Users are able to give feedback iteratively. If design is not design but function oriented, a use-analysis phase regarding functional possibilities is possible.

- Concept testing

After specification of products, this method aims an evaluation by customers regarding sketch drafts and first concepts. It should be supplemented with later evaluations (e.g. beta testing). If function oriented, concepts of solutions regarding functional parameters are tested.

- Beta testing

Beta testing focuses on product testing after first prototypes with end users. As products are already created in this phase, it should be supplemented by prior customer involvement strategies. From customer involvement perspective, this method starts when prototypes are already created, so the impact if a function oriented approach is applied is minimal.

- Lead User Method

“Lead Users” represent end customers facing needs that will be general in a future market. Design specifications until first prototypes are created by these users (supported by designers) to find appropriate solutions if functions cannot be executed accordingly. By orienting on functions instead of devices in this method, lead users are able to access a more sophisticated spectrum of solutions.

- Consumer idealized design

This method aims the generation of product designs by consumers with support of a facilitator as a group exercise. Participants of the exercise select representatives of target markets who create: A product design, a list of requirements of the product with focus to the target market and a record of reasons regarding design choices. If a design is not limited to a specific device, end user needs can be more directly integrated into a product.

- Participatory Ergonomics

The involvement of employees who participate in the production into the design process as designers but also as end users regarding product testing. This method especially aims the integration of experience from different groups knowing about limits regarding the production of a product. As production of a product mainly depends on hardware capabilities, a function-oriented approach would infer new production possibilities un-limiting device structures and enhancing (new) device inventions.

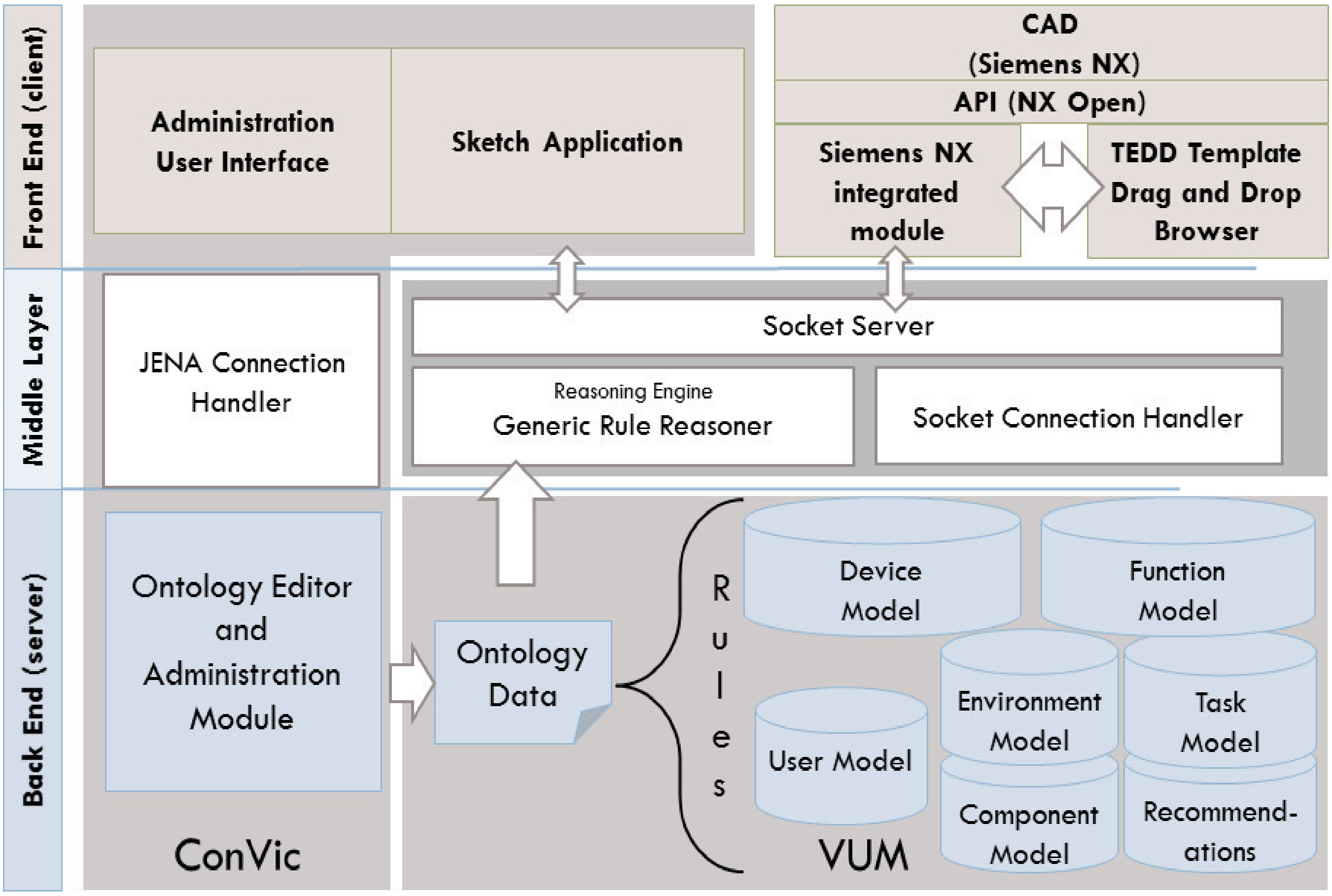


Figure 2: Device Model, Function Model and TEDD integrated into the VICON system architecture

The approach presented by this paper aims to integrate into two parts of product development (see Figure 1). In specification and concept development phase device specifications are described. By utilizing a model driven function superstructure designers (and end customers) are not limited by device structures but can orient on functional parameters. Other projects utilizing ontology approaches e.g. VERITAS are presented in (Modzelewski 2014; Poirson & Delangle 2013).

III. Model Driven Function Superstructure

The ontology approach of ConVic utilizes Virtual User Models representing the scenario specification of the target product. It contains a User, Environment, Task, Component and Recommendation Model describing the target end customer group including impairments, typical environments with focus on e.g. background lighting, typical tasks for the target product, product components and qualitative and quantitative recommendations (Matiouk et al. 2013; Kirisci et al. 2011).

Device related models are Task and Recommendation Models. For each device, an ontology including both models is added to the initial ontology containing models of users, environments and components. To represent a variety of devices, the ontology containing Task and Recommendation Models is separated and can be replaced.

The pro of this approach is changing only the device related models for new products. Without this separation all task and recommendation instances must be recreated if producing a new device. Here we see redundancies. A less redundant approach would rely on functions instead of devices.

Figure 2 presents the approach integrated into the VICON system architecture. On ontology server side, device model and function model aim to implement the context of functional capabilities of devices into the existing approach. Existing models (User, Environment, Task, Component and Recommendation Model) can be re-used here as an initial ontology set. Already existing rules of can be reused but new rules must be added for the integration of function and device model merging all different models together (Modzelewski 2014). As only forward chain rules are implemented, the superstructure of function and device model on top of existing models can be implemented in addition to existing rules.

A function oriented approach does not focus on devices as the main target but on a set of related functions (see Table 1). It is possible for designers to choose functions, which the target device is able to perform, inferring a list of devices for realization. To create a new device, only a new instance as a member of the device model has to be created (see Table 2).

An additional benefit of this separation is the representation of functional context information and the connection to the Component Model necessary for the dynamic presentation of CAD object templates (see section VI).

Table 1: Function Model predicates for ontology definition

Function Model	
Predicate	**Description**
Name	Name of a single function
ID	Unified ID for identification
Description	Brief description of the function
TaskSequence	Sequence of Task Model Instance IDs describing the function

The aim of the Function Model is to categorize functions individually with the purpose of building a relation between functional aspects of a design and task specific descriptions. The relation between functions and end user capabilities is made by the Task Model in which each instance categorizes problematic User Model Profiles. For example if a task can hardly be performed by users with manual dexterity impairments, a task instance consists of a value of "MD1, MD2" for the "Impairment" predicate (MD1 means mild, MD2 moderate manual dexterity impaired profile groups, see D2.2 of the VICON project).

Table 2: Device Model predicates for ontology definition

Device Model	
Predicate	**Description**
Name	Name of a single device
ID	Unified ID for identification
Description	Brief description of this device
Functions	Sequence of Function Model Instance IDs the device is able to perform
Components	Sequence of Component Model IDs, describing recommended components for CAD

The approach of the Device Model is similar. Devices can be described by the definition of each predicate for single instances. By creating this connection of function and device, designers are able to select a function and choose which device would be most suitable to represent this function. On the other side there is a need to create new devices, if a set of function is selected and no device is able to cover these.

IV. Device Selection Approach

The possibility to infer devices from functions represents a central unit in the modification of the current ontology approach. With addition of Function and Device Model new devices can be emerged if existing devices are not able to represent all selected functions.

Figure 3 presents an exemplary bonding between functions and devices regarding functionality. Users of the system are able to select a set of functions which can internally generate a set of tasks to integrate into the final ontology used. The selection defines instances of the general model are used for one specific target product. In the example there is no device available which is able to integrate a function set of "Make a call", "Read E-Mails" and "Watch TV". The closest device able to perform two of three functions is a mobile phone. Derived from this scenario, two issues can be stated: (1) It is necessary to create/invent a new device which is able to perform all three functions and (2) the new device should orient on mobile phones as they already are able to cover most of the functions.

Another question raised by this hypothetic scenario is why mobile phones are not able to realize the function to "Watch TV" inferring to include functions to devices if necessary.

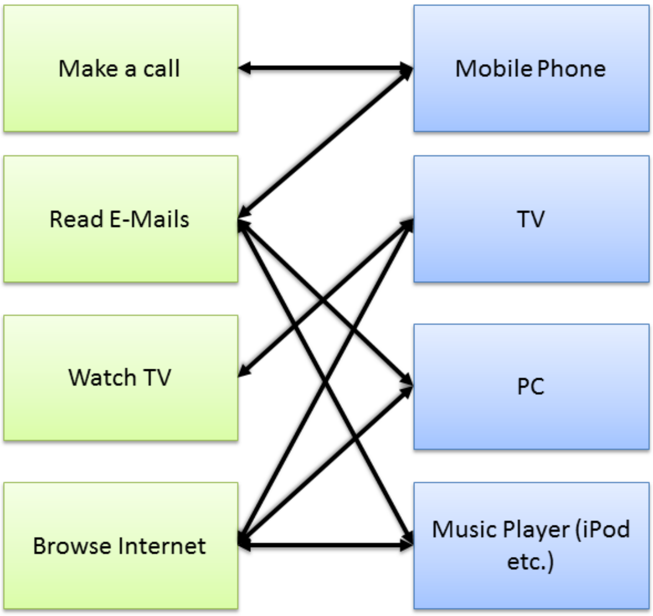


Figure 3: Exemplary bonds between functions and devices

The inference of the ontology created in the VICON project includes additional contextual information for bonding of each model regarding the output of recommendations based on a specific scenario (Lawo et al. 2011; Kirisci et al. 2012; Vicon Consortium 2012). If both Function and Device Model are included, predefined templates regarding a CAD model can be presented. An extension of the Component Model is necessary to include CAD objects including contextual data as brightness of a screen, noise of a speaker (templates). Designers will be able to choose from a set of template components to add or replace their current product design for products that can be used by an as wide group of population as possible.

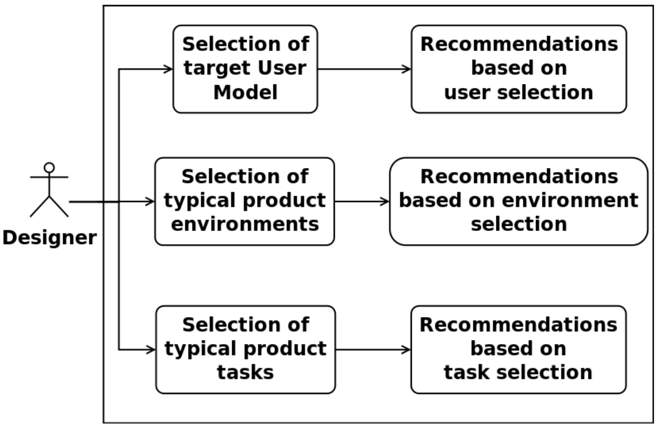

Designer
Selection of target User Model
Recommendations based on user selection
Selection of typical product environments
Recommendations based on environment selection
Selection of typical product tasks
Recommendations based on task selection

Figure 4: Abstract use case definition of designers using the support system ConVic

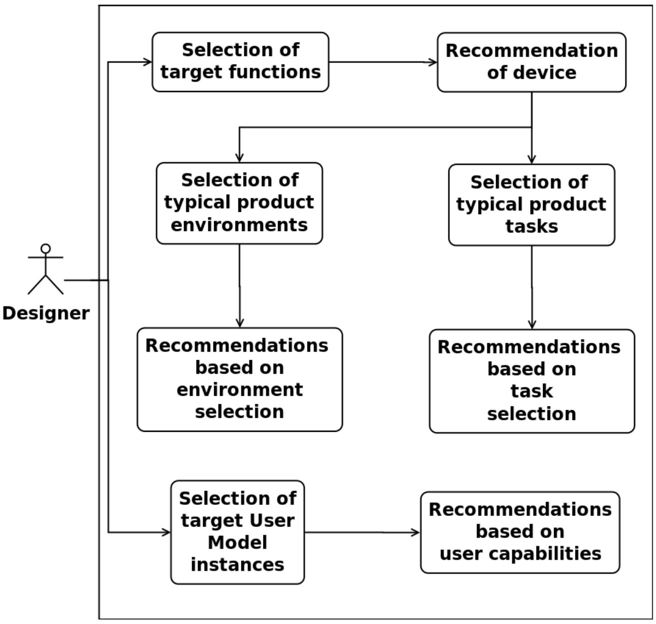

Designer
Selection of target functions
Recommendation of device
Selection of typical product environments
Selection of typical product tasks
Recommendations based on environment selection
Recommendations based on task selection
Selection of target User Model instances
Recommendations based on user capabilities

Figure 5: Abstract use case definition of designers related to functional aspects

The basic idea is to support designers through the complete product development process. Figure 4 presents a very abstract view of general input and output during the sketch design phase defining the selection of target users, typical product environment and typical tasks that can be performed by the target product. For the second phase the input of product components is added inferring the output of quantitative recommendations (Kirisci et al. 2011). The extension of a Function and Device Model enables this procedure to infer recommendations based on functional aspects for a more general view. Instead of the selection of environment and task specifications, designers are able to select functions. Derived from these functions, scenario-related context can be applied including (one or more) environments and tasks.

In addition to the recommendation output as seen in Figure 5, it is also possible for designers to get CAD related templates of product components. An extension to the Component Model of CAD templates enables the possibility to generate from function selection a set of inclusively designed components.

V. Extension of Component Model

The component model used in ConVic does not include CAD related context information but component features (e.g. how many states does a component have) necessary for the presentation of qualitative and quantitative recommendations (Matiouk et al. 2013; Modzelewski 2014). An extension is needed regarding the inference of templates as defined in the previous chapter. The additional predicate contains the file path of a CAD file is utilized for the next step of template presentation. Instances of the component model are also inferred from recommendations directly, CAD templates can be produced by two designer selections. The current selection of target User Model identifying end user capabilities but also the selection of functions the device should be able to perform.

These templates include parametrical data as attributes to identify additional aspects regarding functionality of a component. Initial CAD objects do not include any contextual data. For example if a press button is created, inside of the CAD environment it is represented as a cube shape object. The representation of contextual data regarding functionality but also component related aspects set a basis for the improvements of product development processes.

Another additional pro is the simplification measure to create individual products for customers, if preferences of end customers regarding component functionalities can already be included during CAD phase.

VI. Inclusive Design Templates for CAD

The utilization of the selection of functionalities is able to produce a set of inclusively designed templates for the CAD environment. Based on the target functionality selection (Figure 5) the system is able to produce a set of components predefined in the Component Model for implementation in the current CAD environment Siemens NX[2]. The templates already include contextual information necessary for the adaptation during the CAD phase as seen in (Modzelewski 2014).

The conceptual tool **TEDD** (**TE**mplate **D**rag and **D**rop browser for inclusive design) presents a set of inclusively designed objects.

[2] http://www.plm.automation.siemens.com/de_de/products/nx/

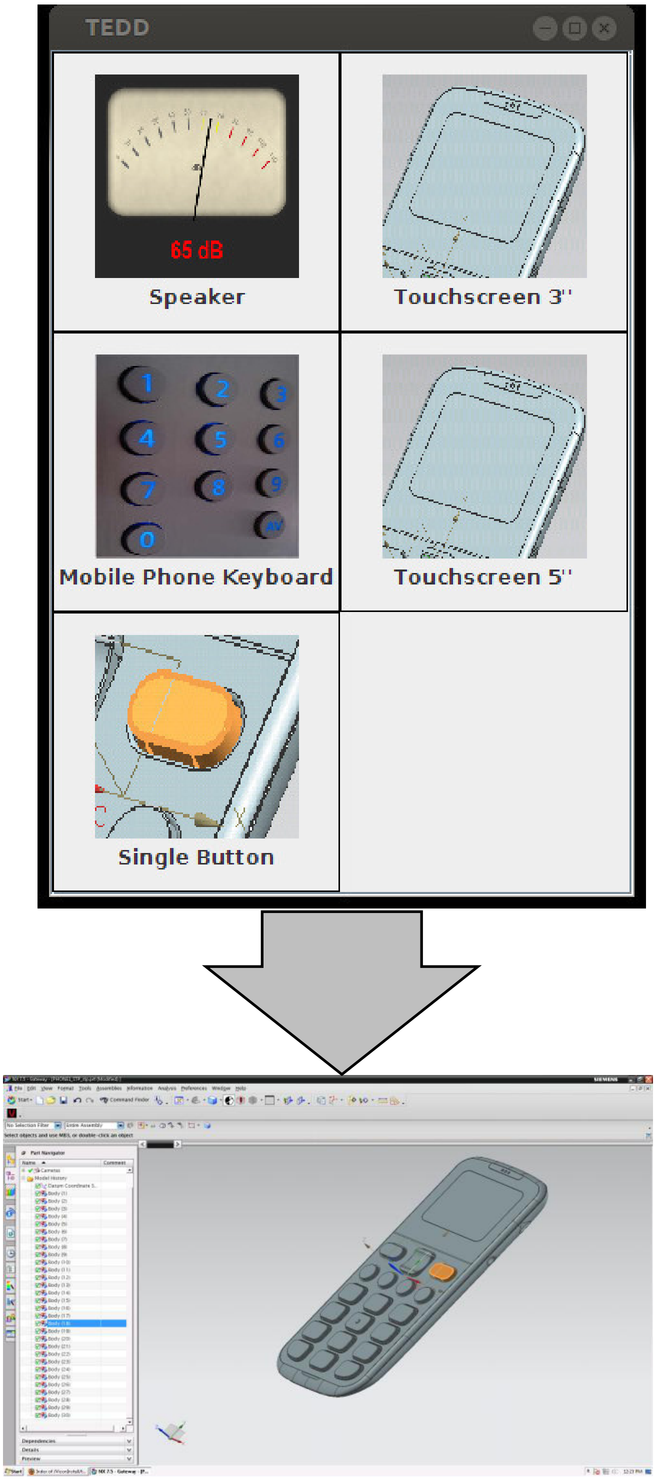


Figure 6: TEDD - Template browser for Drag & Drop of inclusively designed components in CAD environment

Figure 6 presents an exemplary setup based on components necessary for the functionality to make a call. Using the selection of the end user (designer) of functions, a set of possible devices for each function is generated using the Functions predicate as seen in Table 2. In addition to existing qualitative and quantitative recommendations of the VICON system, Users are able to include predefined templates and adjust them.

The following rule describes a simple rule based solution for connecting functions with devices by predicates as seen in (The Apache Software Foundation 2013). First two variables "?x" and "?y" are created including variables for their ID predicate. The predicate Functions is stored in the variable *"?fun"*. By concatenating the variable *"?funid"* between *"(.*)"*, a regular expression is utilized which is always true, if the function ID is found within the *"?fun"* variable. The second concatenation is used for definition of the recommendation class containing all recommended devices for a single function.

(?x rdf:type Vicon:Function), (?x Vicon:FunctionID ?funid),
(?y rdf:type Vicon:Device), (?y Vicon:DeviceID ?devid), ,
(?y Vicon:DeviceFunctions ?fun),
strConcat("(.)", ?funid,"(.*"), ?reg), regex(?fun,?reg),*
strConcat("Vicon:FunctionReco_", ?funid, ?recoClass)
-> (?y rdf:type ?recoClass).

By adding redundant components as several versions of touchscreens, the impact on creative process of design is as small as possible.

In the next step integration into Siemens NX CAD environment is aimed, utilizing the integrated knowledge-reuse module[3].

VII. Conclusion

This paper presented an approach and a tool for the generation of a set of inclusive design CAD templates based on the target functions of a product design. Furthermore a function and device model was introduced representing functional and device related aspects utilized for the generation of an optimal device configuration. The developed software is open source and accessible via the VICON project homepage (Vicon Consortium 2013). The usefulness of the approach was evaluated with eleven physical product designers. The results of this investigation with these designers as well as the outcome of an evaluation with forty-eight beneficiaries of products designed using the presented approach are published in (Modzelewski 2014).

VIII. Future Work

Regarding a more detailed functional design Stone and Wood (Stone & Wood 2000) presented a functional basis regarding material, energy and signal. It would be interesting how and if it is necessary to represent each single aspect of functional design but also if this would be suitable to even ignore devices as an instance focussing only on functional aspects. An extension and population of the Component

[3] Siemens NX knowledge-reuse module: https://www.plm.automation.siemens.com/de_de/products/nx/for-design/knowledge-re-use/library.shtml#lightview-close

Model regarding the integration of CAD templates into the model is advantageous for the generation of recommended components. For this implementation a variety of CAD templates including contextual data is needed.

IX. References

Akao, Y., 2004. Quality function deployment: integrating customer requirements into product design, Productivity Press.

BMW AG, BMW Website - BMW Techniklexikon : Controller.

Clarkson, J. et al., 2003. A designer-centred approach. Inclusive design: Design for the whole population.

Coleman, R. & Lebbon, C., 2005. Inclusive design. Helen Hamlyn Research Centre, Royal College of Art.

Dong, H., Keates, S. & Clarkson, P., 2004. Inclusive design in industry: barriers, drivers and the business case. User-Centered Interaction Paradigms for Universal Access in the Information Society, pp.305–319.

Kirisci, P. et al., 2012. Supporting Inclusive Design of Mobile Devices with a Context Model. InTech Open Science.

Kirisci, P. et al., 2011. Supporting inclusive design of user interfaces with a virtual user model. Universal Access in Human-Computer Interaction. Users Diversity, pp.69–78.

Langdon, P. & Thimbleby, H., 2010. Inclusion and interaction: Designing interaction for inclusive populations. Interacting with Computers, 22(6), pp.439–448.

Lawo, M. et al., 2011. Virtual User Models - Approach and first results of the VICON project. In P. Cunningham & M. C. (Eds), eds. eChallenges e-2011 Conference Proceedings. IIMC International Information Management Corporation Ltd.

Matiouk, S. et al., 2013. Prototype of a Virtual User Modeling Software Framework for Inclusive Design of Consumer Products and User Interfaces. In Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion. Springer, pp. 59–66.

Modzelewski, M., 2014. An ontology-based approach to achieve inclusive design support in the early phases of the product development process. Bremen: University of Bremen, Dissertation. Available at: http://nbn-resolving.de/urn:nbn:de:gbv:46-00103662-13 [Accessed March 11, 2014].

Modzelewski, M. et al., 2012. Creative Design for Inclusion Using Virtual User Models. Computers Helping People with Special Needs, pp.288–294.

Newell, A.F. & Gregor, P., 2000. User sensitive inclusive design - in search of a new paradigm. In Proceedings on the 2000 conference on Universal Usability. CUU ’00. New York, NY, USA: ACM, pp. 39–44. Available at: http://doi.acm.org/10.1145/355460.355470.

Poirson, E. & Delangle, M., 2013. Comparative analysis of human modeling tools.

Segan, S., 2012. Enter the Phablet: A History of Phone-Tablet Hybrids.

Stone, R.B. & Wood, K.L., 2000. Development of a Functional Basis for Design. Journal of Mechanical Design, 122(4), p.359. Available at: http://mechanicaldesign.asmedigitalcollection.asme.org/article.aspx?articleid=1446060 [Accessed March 18, 2014].

Strang, T. & Linnhoff-Popien, C., 2004. A context modeling survey. In Workshop Proceedings.

The Apache Software Foundation, 2013. Reasoners and rule engines: Jena inference support.

VDI-Gesellschaft Entwicklung Konstruktion Vertrieb, 1993. VDI 2221, Methodik zum Entwickeln und Konstruieren Technischer Systeme und Produkte. Verein Deutscher Ingenieure (Hrsg.): VDI-Handbuch Methodisches Konstruieren, Berlin.

Vicon Consortium, 2013. ConVic | Free Graphics software downloads at SourceForge.net. Available at: http://sourceforge.net/projects/convic/ [Accessed March 21, 2014].

Vicon Consortium, 2012. Project Deliverable 2.2 - Virtual User Model (Final release). Available at: http://www.vicon-project.eu.

Zeller, A., Wagner, A. & Spreng, M., 2001. iDrive-Zentrale Bedienung im neuen 7er von BMW/iDrive-centralised operation in the new 7series of BMW. VDI-Berichte, (1646).

Author Index

Achkoski, Jugoslav 211
Ahfock, Daniel 61
Atique, Mohammad 177

Balova, Tatiana 271
Bellalouna, Monia 55
Bialas, Andrzej 201
Bielak, Halina 93
Bouchekara, Houssem 157
Boudjefdjouf, Hamza 157

Cabaj, Krzysztof 143
Cackowski, Adam 171
Chen, Ray-Ming 3
Chudzikiewicz, Jan 183
Czapliński, Paweł 171

Danger, Marc 259
Dinh, Khoa Truong 143
Dolińska, Iwona 149
Domagała, Piotr 21
Drager, Steven 131

Falowo, Olabisi E. 165
Feitosa, Eduardo 101
Finaskins, Dmitrijs 71
Fliszkiewicz, Mateusz 45
Franczyk, Bogdan 237, 245
Furtak, Janusz 183

Geiger, Gebhard 119
Gladden, Matthew 219
Gomula, Jerzy 37
Grottker, Ulrich 123

Henchiri, Abir 55

Ivanov, Vladimir 13

Kawala-Janik, Aleksandra 29
Khaznaji, Walid 55
Kokosiński, Zbigniew 21
Korbaś, Grzegorz Paweł 29
Kornecki, Andrew J. 131
Krasuski, Adam 45
Krenski, Karol 45
Kreutz, Diego 101
Kujawa, Dominik 79
Kukliński, Sławomir 143
Kuniszewski, Damian 171

Lauks, Gunars 71
Lawo, Michael 283
Levashenko, Vitaly 37

Majer, Marcin 85
Masiukiewicz, Antoni 149
McGrory, Clare 61
McKeever, William 131
Mehasni, Rabia 157
Melki, Antoine 111
Mercier-Laurent, Eunika 253
Miezitis, Gundars 191
Milosz, Marek 271
Modzelewski, Markus 283

Nayse, Sham P 177

Oberländer, Jan 237
Orlandi, Antonio 157

Palyukh, Boris 13
Pancerz, Krzysztof 37
Pańczyk, Michał 93
Paulis, Francesco De 157
Peter, Michael 123
Peters, Daniel 123
Podpora, Michal 29, 85

Radziszewski, Paweł 143
Robak, Marcin 245
Robak, Silva 245
Rousseaux, Francis 259
Ruzic-Dimitrijevic, Ljijana 229
Rządkowski, Grzegorz 149

Sala, Jolanta 265
Seifert, Jean-Pierre 123
Sibanda, Clifford C. L. 165
Sorel, Pierre 259
Sotnikov, Alexander 13
Stevic, Milorad 229
Szymański, Karol 171

Tańska, Halina 265
Taranovs, Romans 191
Thiel, Florian 123
Trajkovik, Vladimir 211

W
Wawrzyniak, Piotr 171
Wronkowski, Łukasz 171
Wytrębowicz, Jacek 143

Z
Zaitseva, Elena 37
Zalewski, Janusz 131
Zhomartkyzy, Gulnaz 271