DOI: 10.15439/2014F177
ACSIS, Vol. 2

Anonymization of data sets from Service Delivery Platforms

Radosław Naumiuk (1,2)
(1) Orange Labs CBR
ul. Obrzeżna 7
02-691 Warsaw, Poland
(2) Military University of
Technology
ul. Gen. Sylwestra Kaliskiego 2
00-908 Warsaw, Poland
Email: radek92n@gmail.com

Jarosław Legierski
Orange Labs
CBR
ul. Obrzeżna 7,
02-691 Warsaw, Poland
Email: jaroslaw.legierski@orange.com

***Abstract*—The paper presents an anonymization of telecommunication data sets collected through Service Delivery Platforms (SDP), and describes an example tool SDPAnonymizer to make such operation. Information from SDP are processed in form of log files, consisting data sets, which show activity of users of APIs (Application Programming Interfaces). Data sets which should be anonymized contain sensitive data, for example: Names, MSISDN numbers (Mobile Station International Subscriber Directory Numbers) or IP addresses processed by Service Delivery Platforms..**

I. Introduction

IN ICT infrastructure of telecommunication operators, there is processed a large number of data related to subscribers activity and data from flow of information between different systems (call and messaging flows, billing, payment etc.). One of the systems taking part in processing of those information flow, is a Service Delivery Platform (SDP), which is e.g. used for exposure telecommunication services in Internet in form of API (ang. Application Programming Interface), for external programmers. In SDP, a large number of information is being processed, which is related to activity of users and used APIs. Among this data, there are both: information which can be public accessible (for example, the global number of API calls), and as well sensitive data, which publishing is impossible without the anonymization data sets for example the number of MSISDN, ID of localized people, content of SMS, MMS or USSD (Unstructured Supplementary Service Data) messages, Names etc.). This publication focuses on information processed by SDP and proposing a way of their anonymization, based on processing single data sets and presenting an example of application for processing them.

The article is organized as follows: Part I is a short introduction. Part II includes a review of some of the different works related to anonymization of telecommunication operator's data. Part III describes concept of Service Delivery Plaforms. Part IV concentrates on anonymization methods used in the application prototype. Part V presents examples of data sets undergoing the anonymization. Part VI describes the architecture of a solution and part VII shows plans of its further development. The last VIII part contains a short summary.

II. Telecommunication Data Anonymization

In literature, there can be found some examples of anonymization of information used in telecommunication systems. In position [1], the authors give examples of anonymization of IP addresses, made in real time, used during passive monitoring of networks based on TCP/IP protocol. Many works concentrates on issues of anonymization of information related to localization of mobile terminal (Location Based Services - LBS) [2], [3] made in order to provide privacy and security of information about subscriber's location (for example in case of unauthorized use of LBS). Many of data repositories on telecommunication operator's side has the information stored in form of records in databases. Due to that, the most safe way of anonymization is data processing on level of whole repositories, for both single records as well as their sets (data sets) according to k-anonymity methods [4], proposed in following literature positions [5], [6], and widely discussed in part IV of this paper.

III. Service Delivery Platforms

SDP (ang. Service Delivery Platform) is a system present in architecture of telecommunication operator systems, that manages network enablers and open API to allow third users to use these enablers. Such services can be for example:

978-83-60810-58-3/$25.00 © 2014, IEEE

functionality prepaid or exposing of functions of the operator in form of API e.g. in order to build services (in the sense of applicationsproviding specific functions) outside the operator - [7], [8], [9], [10]. The role of the SDP is creating and control of sessions and protocols for chosen service. Versatility of SDP usage makes, that it is present in almost every interaction of user with telecommunication service. This fact makes the service platform the source of large amount of data, which can be used in business, statistic or maintenance purposes, however they must undergo anonymization similarly often, before they can be provided.

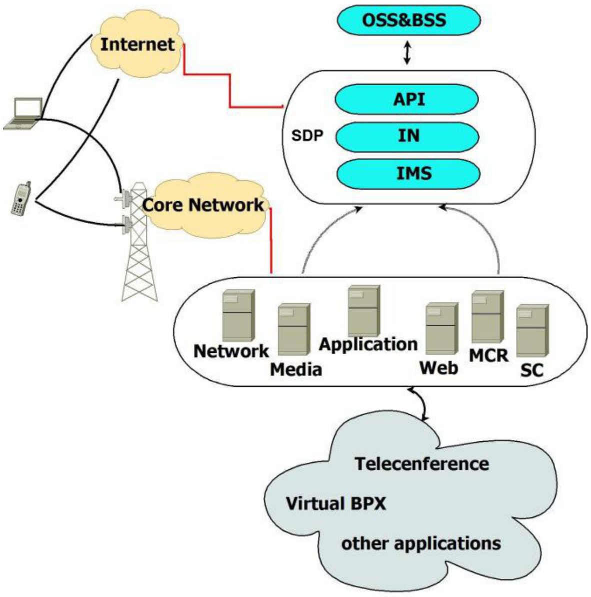


Fig. 1.Concept of Service Delivery Platform

IV. ANONIMIZATION METHODS

Anonymization of data is an operation of data processing, in way to disallow identification of individuals and sensitive data, existing in the processed data, in a way so that the data set remained as readable as possible, and so that it doesn't lose any of the content of general nature, and especially statistical data. Because in the information received from service platform, the basic unit of data is a single record, which represents - from the point of view of project-oriented programming - an event, in the following publication, the authors concentrated on processing of data in form of single records in data set, saved in logs of the SDP platform, provided in form of text files.

The operation of data processing focuses on methods based on data deletion and methods based on data disturbance [5], [6].

Among the most often used methods based on data deletion there are:

1) the deletion of variables - deletion of a record of direct identifiers, which allow identification (e.g. MSISDN number, name, surname, IP addresses etc.)

2) the deletion of records - deletion of individual records in data set. This method is advised only, when identification of sensitive data is possible despite the usage of different anonymization methods.

3) global recording - is based on aggregation of data in data set, according to predefined criteria (e.g. the number of terminals in given area defined by Local Area Code - LAC in various time intervals).

The methods based on data disturbance, in case of application presented in the following article, concentrate on methods from post-randomization group.In case of data from SDP, was implemented algorithms such as MD5 or SHA2 and replacement of sensitive data present in record, with character string generated by used hash function.

1) the method based on MD5 algorithm (Message-digest algorithm version 5). MD5 is an algorithm from cryptography field. It is a popular cryptographic shortcut function, which generates, from any data string its 128-bit hash .

2) the method based on SHA algorithm (Secure Hash Algorithm) - SHA is a family of connected with each other, cryptographic shortcut functions, designed by NSA (National Security Agency) and published by National Institute of Standards and Technology. It has several versions: SHA-1,SHA-256,SHA-384,SHA-512. In SDPAnonymizer application, was used the SHA algorithm, in SHA-256 version).

V. TEST DATA SET

As test data, was used a set of records from test Service Platforms exposing API to Orange network. Data sets were in form of text files in two versions:

Data partially aggregated - set of files with API users activities sorted on the basis of the MSISDN number in form of several files:

getlocation.csv - terminal location api function calls

getop.csv – which operator function calls

sendsms.csv –Send SMS function calls

sendussd.csv –Send USSD function calls

Every of presented above of files had following structure:

DATA HOUR|LOGIN=MSISDN

Where: DATA HOUR – date and time of the event, LOGIN number MSISDN of the API user e.g.:

```
2011.11.29 18:17:17|LOGIN=48500163047
```

The second type of data were raw data in txt format in form:

```
recordID   Data Time Event   GivenNameName  MSISDN
IP address deviceidhttpsessionID

2691    237633      2011-03-09 14:43:57  Authentication
attempt (ussd pass sent)
JaroslawLegierski48500163047192.168.20.124 null
1299678218892
2693    265971      2011-03-09 14:44:41  index.jsp          -
Authentication                                   successfull
     JaroslawLegierski48500163047   192.168.20.124
     null 1299678218892
2694    265971      2011-03-
0914:45:20WebServ:http://10.255.240.50
:2006/tp/orangelabs/jslee/oc/webservices/sendRestUssdN
otify?number=48500163047&text=Default+JSLEE+US
SD+WebService+text&webSerName=sendRestUssdNotif
y    Jaroslaw Legierski48500163047  192.168.20.124
     null 1299678218892
2695    265971      2011-03-09                    14:46:17
WebServ:http://10.255.240.50:2006/tp/orangelabs/jslee/o
c/webservices/sendRestMMS?number=48508367971&su
bject=test+123&priority=Low&text=Default+MMS+tex
t+sent+from+JSLEE+WebService+1233&webSerName
=sendRestMMS
     JaroslawLegierski48500163047192.168.20.124 null
     1299678218892
2696          265971          2011-03-09          14:47:20
WebServ:http://10.255.240.50:2006/tp/orangelabs/jslee/o
c/webservices/getRestTS?number=48500163047&webSe
rName=getRestTS
     JaroslawLegierski48500163047192.168.20.124 null
     1299678218892
```

All of the example presented above records, included activity related to API usage, provided through SDP platform by the users (developers) in form of sending SMS, MMS, USSD message, logging in to the system or usage of mobile terminal location function.

VI. SYSTEM ARCHITECTURE

This section of the publication contains the architecture and functionality of the SDPAnonymizer application. The SDPAnonymizer application. The SDPAnonymizer is a simple application operating on SDP platform files in form introduced in chapter V. The application process the anonymization of data sets included in the files.

Input file containing database, which is the data set to anonymize, is loaded to application, and then modified by one of the user-chosen methods. The final result of the program activity is the output file containing anonymized data. The functionality of application is based among other things on algorithms already existing in Java libraries, such as MD5 [11], [12] or SHA2, and on mathematical operations (e.g. summation, deletion etc.).

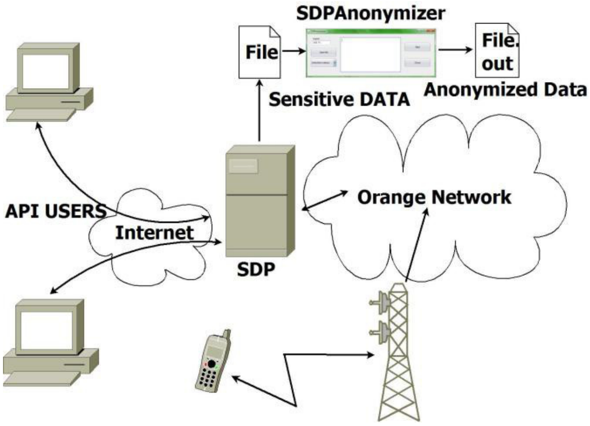


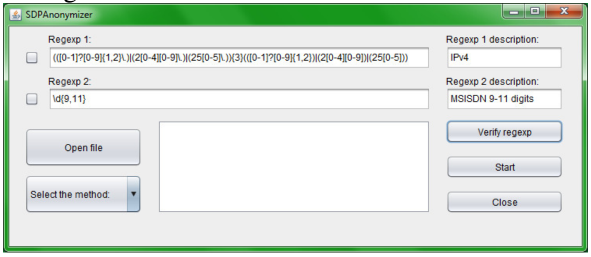
Fig.2. SDPAnonymizer– application concept

A) Programing environment

The application for data anonymization was developed and tested for Java environment JRE in 1.6 and 1.7version. In creation of the SDPAnonymizer application, was used Eclipse environment in "Juno Service Release 2" version. Application user's graphical interface was created with the usage of standard SWING libraries included in JDK.

B) Application

SDPAnonymizeris an application, dedicated for anonymization of data sets from operator service platforms.Tool was developed as easy to use and simple stand-alone solution not integrated with any ETL (Extract, Transform, and Toad)data management framework.



□ Regexp 1: (([0-1]?[0-9]{1,2}\.)|(2[0-4][0-9]\.)|(25[0-5]\.)){3}(([0-1]?[0-9]{1,2})|(2[0-4][0-9])|(25[0-5]))

Regexp 1 description: IPv4

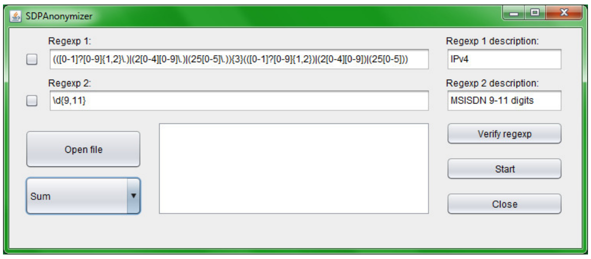
□ Regexp 2: \d{9,11}

Regexp 2 description: MSISDN 9-11 digits

Fig.3. Application window.

In the program, there were implemented methods such as: SUM, DELETE RECORD, CATCH RECORD, DELETE NUMBER, MD5, SHA-256. Program identifies variables for anonymization using regular expressions defined in GUI in two textboxes.

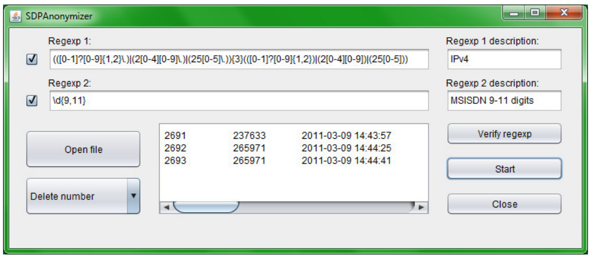
The SUM method is a method from a group of methods based on deletion of data as a part of global data set recording. In the version implemented in SDPAnonymizer, summation is based on summing of numbers of API usage by one user identified by MSISDN number.


SDPAnonymizer

Regexp 1: □ (([0-1]?[0-9]{1,2}\.)|(2[0-4][0-9]\.)|(25[0-5]\.)){3}(([0-1]?[0-9]{1,2})|(2[0-4][0-9])|(25[0-5]))
Regexp 1 description: IPv4
Regexp 2: □ \d{9,11}
Regexp 2 description: MSISDN 9-11 digits
Method: Sum

Fig. 4. Application window –SUM method

DELELE NUMBER – is a method from group of methods based on deletion of variables. The MSISDN, IP address or Name is identified in the processed file using regular expressions, and it's deletion, through replacement by XXXX string.

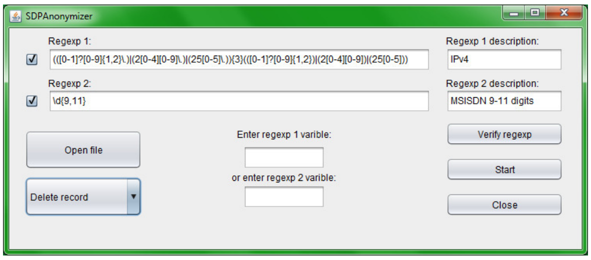

SDPAnonymizer

Regexp 1: ☑ (([0-1]?[0-9]{1,2}\.)|(2[0-4][0-9]\.)|(25[0-5]\.)){3}(([0-1]?[0-9]{1,2})|(2[0-4][0-9])|(25[0-5]))
Regexp 1 description: IPv4
Regexp 2: ☑ \d{9,11}
Regexp 2 description: MSISDN 9-11 digits
Method: Delete number

```
2691    237633    2011-03-09 14:43:57
2692    265971    2011-03-09 14:44:25
2693    265971    2011-03-09 14:44:41
```

Fig. 5. Application window – DELETE NUMBER

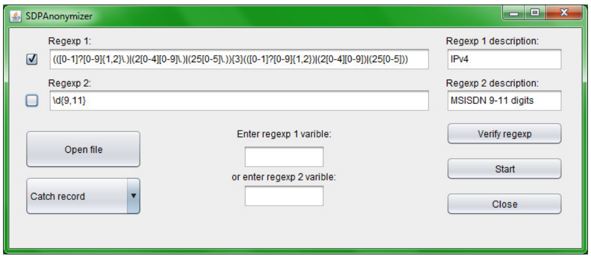
DELETE RECORD method is based on deletion of records in data set. It works by deleting all records containing defined variables (MSISDN, IP address or Name etc.).


SDPAnonymizer

Regexp 1: ☑ (([0-1]?[0-9]{1,2}\.)|(2[0-4][0-9]\.)|(25[0-5]\.)){3}(([0-1]?[0-9]{1,2})|(2[0-4][0-9])|(25[0-5]))
Regexp 1 description: IPv4
Regexp 2: ☑ \d{9,11}
Regexp 2 description: MSISDN 9-11 digits
Method: Delete record
Enter regexp 1 varible:
or enter regexp 2 varible:

Fig. 6. Application window – DELETE RECORD

Complementary action to the method above is the CATCH RECORD method, which is based on deletion of all records in data set, which do not have defined variable.


SDPAnonymizer

Regexp 1: ☑ (([0-1]?[0-9]{1,2}\.)|(2[0-4][0-9]\.)|(25[0-5]\.)){3}(([0-1]?[0-9]{1,2})|(2[0-4][0-9])|(25[0-5]))
Regexp 1 description: IPv4
Regexp 2: □ \d{9,11}
Regexp 2 description: MSISDN 9-11 digits
Method: Catch record
Enter regexp 1 varible:
or enter regexp 2 varible:

Fig. 7. Application window– CATCH RECORD

As it was already mentioned before, the application also implements two methods from Post-randomization group implemented. MD5 and SHA2 are shortcut functions, which transform a variable defined by regular expression in a hash string. Examples of records including data, for which hash functions were calculated are listed below (for MSISDN numbers anonymization):

Original record:
2691 237633 2011-03-09
14:43:57Authentication attempt (ussd pass
sent)JaroslawLegierski48500163047192.168.20.12
4 null 1299678218892

MD5:
2691 237633 2011-03-09
14:43:57Authentication attempt (ussd pass
sent)JaroslawLegierskiaf4f2db11d6a477aeed011a0
2aa9d549192.168.20.124 null 1299678218892

SHA2:
2691 237633 2011-03-09
14:43:57Authentication attempt (ussd pass
sent)JaroslawLegierskid9e88a466dd7cc252751458
1b2c73b0bffe34173ce60dd0ace42aa5751e79e19
192.168.20.124 null 1299678218892

Data marked in grey is the effect of encrypting, made with use of MD5 and SHA2 shortcut function.

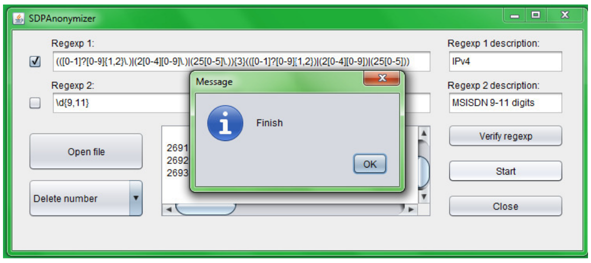
The program also has the option to define another data than MSISDN, IP address or Name which can be anonymized, through the use of regular expressions(Table1) defined in SDPAnonimizer GUI, which gives us the possibility to anonymize data in different bases (e.g. consisting of telephone numbers of different length and stored in different formats, IPV6 addresses, or terminal location coordinates).

Table 1. Example regular expressionsused in SDPAnonymizer

Regular expressions	Identified variable
\d(9,11)	MSISDN (9-11 digits)
^(25[0-5]|2[0-4]\d|[0-	IPV4 Network Address

1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$	
^(?:[0-9a-fA-F]{1,4}:){7}[0-9a-fA-F]{1,4}$	IPV6 Network Address
[\s-]([A-Z][a-z]*)+[\s-]([A-Z][a-z]*)	GivenName Name

The handling of events and errors, from the application's end user point of view was implemented through display of a dialog windows (pop-ups).



☑ Regexp 1: (([0-1]?[0-9]{1,2}\.)|(2[0-4][0-9]\.)|(25[0-5]\.)){3}(([0-1]?[0-9]{1,2})|(2[0-4][0-9])|(25[0-5]))

Regexp 1 description: IPv4

□ Regexp 2: \d{9,11}

Regexp 2 description: MSISDN 9-11 digits

Finish

2691
2692
2693

Fig. 8. Application window – dialog window, finish of file processing

VII. FUTURE WORK

In the future, the SDPAnonzmiyer program will be upgraded by functions, doing the process of microaggregation. The Micro-aggregation replaces the anonymized value - with medium value calculated for small group of records (micro-aggregate), which will for example allow to present the usage of communication functions, shown by API for example in hour intervals. Also additional method of encryption, which is the SHA3 method, will be added. In the next step the performance, and load test of application SDPAnonymizerare planned.

Target task is to transform the application into web service form, and exposure of anonymization methods in form of RESTlikeAPI, in order to increase the availability of anonymization methods for users - programmers.

V. SUMMARY

The SDPAnonymizer application is dedicated for Service Delivery Platform administrators, which, for example, share data from SDP system logs with external parties (e.g. the suppliers, which are responsible of the Level 3 maintenance of service platforms). Anonymization of information processed by service platforms in form of dedicated application, working on files solves following problems:

Legal (sensitive data such as MSISDN aren't shared) and data are anonymized as close as possible to the data source (anonymization is done by SDP administrator).

Computational - the SDP platform does not process data, so it's not overloaded with additional, computationally expensive tasks (data processing is done by additional component outside the SDP environment)

Mathematical - system administrator receives dedicated tool, with proper algorithms and mathematical functions already implemented (e.g. MD5, SHA).

Prototype of application SDPAnonymizer was made as part of the Open Middleware 2.0 Community by Orange Labs program[13]

REFERENCES

[1] Ubik, S.; Zejdl, P.; Halák, J., "Real-time anonymization in passive network monitoring," Networking and Services, 2007. ICNS. Third International Conference on , vol., no., pp.100,100, 19-25 June 2007

[2] Heechang Shin; Atluri, V.; Vaidya, J., "A Profile Anonymization Model for Privacy in a Personalized Location Based Service Environment," Mobile Data Management, 2008. MDM '08. 9th International Conference on , vol., no., pp.73,80, 27-30 April 2008

[3] Gedik, B.; Ling Liu, "Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms," Mobile Computing, IEEE Transactions on , vol.7, no.1, pp.1,18, Jan. 2008

[4] Latanya Sweeney K-Anonymity: A Model For Protecting Privacy, Journal International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems archive Volume 10 Issue 5, October 2002,pp 557 - 570

[5] Alpa K. Shah,State-of-art in Statistical Anonymization Techniques for Privacy Preserving Data Mining International Journal of Computer Science & Engineering Technology (IJCSET) Vol. 3 No. 7 July 2012

[6] L. Jaganraj, S. Balamurugan, Empirical Investigation on Certain Anonymization Strategies for Preserving Privacy of Social NetworkInternational Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 10, October 2013

[7] Podziewski, A.; Litwiniuk, K.; Legierski, J., "Emergency button — A Telco 2.0 application in the e-health environment," Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on , vol., no., pp.663,677, 9-12 Sept. 2012

[8] 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

[9] 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

[10] Korbel, P.; Skulimowski, P.; Wasilewski, P.; Wawrzyniak, P., "Mobile applications aiding the visually impaired in travelling

with public transport," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.825,828, 8-11 Sept. 2013

[11] Junwei Zhang, Jing Yang, Jianpei Zhang,Yongbin Yuan KIDS:K-anonymization data stream base on sliding window

[12] http://md5calculator.chromefans.org/

[13] Open Middleware 2.0 Community portal – http://www.openmiddleware.pl [20.05.2013]