# Vickrey-Clarke-Groves for privacy-preserving collaborative classification

Anastasia Panoui
School of Electronic, Electrical and
Systems Engineering
Loughborough University, UK
Email: a.panoui@lboro.ac.uk

Sangarapillai Lambotharan
School of Electronic, Electrical and
Systems Engineering
Loughborough University, UK
Email: S.Lambotharan@lboro.ac.uk

Raphael C.-W. Phan
Faculty of Engineering
Multimedia University, Malaysia
Email: raphael@mmu.edu.my

***Abstract*—The combination of game theory and data mining opens new directions and opportunities for developing novel methods for extraction of knowledge among multiple collaborative agents. This paper extends on this combination, and motivated by the work of Nix and Kantarcioglu employs the Vickrey-Clarke-Groves (VCG) mechanism to achieve privacy-preserving collaborative classification. Specifically, in addition to encouraging multiple agents to share data truthfully, we facilitate preservation of privacy. In our model, privacy is accomplished by allowing the parties to supply a controlled amount of perturbed data, instead of randomised data, so long as this perturbation does not harm the overall result of classification. The critical point which determines when this perturbation is harmful is given by the VCG mechanism. Our experiment on real data confirms the potential of the theoretical model, in the sense that VCG mechanism can balance the tradeoff between privacy preservation and good data mining results.**

## I. Introduction

DATA mining provides a range of useful tools for data manipulation and extraction of meaningful information from large data sets, that can improve our lives. For example, collaboration among hospitals and other healthcare institutions by providing the medical record sets, and thus creating a large database, can lead to better and more reliable research results. In a different scenario, markets can share their data related to the customers' shopping preferences, in order to make future product deals and offers that will increase the income. Furthermore, cooperation in international level among governments, by merging intelligence data sets, might result in strengthening the security against terrorism. However, in all cases it is important to ensure that sensitive information must remain hidden and not be disclosed.

This paper addresses the problem of privacy preserving collaborative data mining, motivated by a paper by Nix and Kantarcioglu [1]. A brief description of the setting is as follows: a number of participants, also called agents, jointly supply their individual data sets in order to perform a data mining task and extract information from the large database that is formed. As the trustworthiness of the agents is not guaranteed, it is necessary to add incentives for good behaviour. One approach is to have penalising strategies that will prevent inappropriate behaviour. However, game theory offers a solution with positive incentives. Our work, as in [1], employs a method from a branch of game theory, called mechanism design. More specifically, we use the Vickrey-Clarke-Groves (VCG) mechanism, in which the payoff of each agent contains the agent's contribution to the 'community'. Thus, if an agent's contribution harms the overall result, this agent will be charged and hence receive low payoff. Following the setting of [1] we also choose the data mining task to be classification. However, in contrast to [1], for simulation of an agent who supplies falsified data we modify the complete data set of the agent through a controlled amount of perturbation, rather than random perturbation of certain percentage of the data. Furthermore, apart from complete randomization of the data, which corresponds to the action of an agent who lies or an agent who aims for the maximum possible privacy, we also include small deviation from the true data. The latter action models an agent who wishes to preserve the privacy of his data without damaging the overall result. We show that this strategy results in information gain while keeping the agent's data private.

## II. Related work

The combination of data mining and game theory in a collaborative environment has opened a new direction for research. Halpern and Teague [2] address the problem of secret sharing and multi-party computation, under the assumption that the agents are rational, rather that being good or bad. They show that there exists a randomised secret sharing scheme in which the agents reach a Nash equilibrium that overcomes the iterated deletion of weakly-dominated strategies. In [3] Abraham at al. extend the work of [2], by introducing the notion of $k$-resilient equilibrium, which is similar to the Nash equilibrium, but instead of tolerating deviation from one player, it tolerates deviations by coalitions with at most $k$ members. Examination and analysis of the multi-party computation, and specifically of the secure sum computation problem under a game theoretic framework can be found in [4]. In many scenarios, in order to simulate real world situations the involved parties are divided into good or bad.

This work has been supported by the Engineering and Physical Science Research Council (EPSRC) of UK, under the grant EP/J020389

978-1-4673-4471-5/$25.00 © 2013, IEEE

However, under a game theoretic framework this approach is often replaced by settings where the participants are assumed to be rational, whose aim is to maximise their gain. In this context, the authors of [5] introduce the notion of rational secure computation and show that the ballot-box can be used to securely compute any function. Although security is an important issue to be addressed, the behaviour of the participants must also be examined. Thus, in order to discourage improper behaviour, [4], [6], [7] introduce penalising methods. In particular, assuming semi-honest players, [6] is concerned with the problems that arise in a sovereign information sharing setting. The goal is to ensure that the participants learn the result from the task on the shared information, without gaining any knowledge about the shared data. This is achieved by using an auditing device that will repeatedly check the players' actions, penalising inappropriate behaviour. Punishing strategies against malicious players is also examined in [7], in a setting which includes verification of the results, in addition to the information sharing. A different approach to punishing policies in order to achieve good behaviour is the use of VCG mechanism [1], [8]. In [8] this particular mechanism is employed for regression learning and in [1] for classification.

## III. Mechanism Design

Mechanism design is a branch of game theory concerned with the problem of social welfare [9]–[11]. The setting involves a set of $I$ agents, each one having their own private preferences on a set of alternatives, and a principal, whose role is to ensure that the rules of the mechanism will be followed. The aim of the mechanism is to help the agents make a collective choice that is beneficial for all. Formally, a mechanism is a collection of strategy sets $S_1, \ldots, S_I$ and an outcome function $g : S_1 \times \ldots \times S_I \to X$, where $X$ is a set of possible alternatives. Each alternative is associated with a utility function $u_i(x)$ (known also as payoff), which denotes the gain of agent $i$ when alternative $x$ is chosen. As different alternatives lead to different payoffs, clearly each agent has a different preference on the alternatives. In order to model the distinctiveness of the agent's preferences, we associate each agent with a type $\theta_i$, $i = 1, \ldots, I$. An important point is that the preference, and hence the type of each agent is private information and hence $\theta_i$ is known only to agent $i$. For this reason, in the game theoretic context, we are in an environment of incomplete information.

Once the agents have decided upon the preferences and their type has been determined, they report types $\hat{\theta}_i$, which might or might not coincide with $\theta_i$ (direct revelation mechanism). After $\hat{\theta}_i$ has been announced from all agents, the mechanism selects the collective choice to be

$$k^*(\hat{\theta}) = \arg\max_{k \in K} \sum_i v_i(k, \hat{\theta}_i),$$

where $K$ is the set of possible choices, $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_I)$ and $v_i(k, \hat{\theta}_i)$ is the valuation of agent $i$ on the choice $k$, when his reported type is $\hat{\theta}_i$.

### A. The Vickrey-Clarke-Groves Mechanism

The Vickrey-Clarke-Groves mechanism (denoted by VCG) is a mechanism where the utility function has the following quasi-linear form:

$$u_i(x, \theta_i) = v_i\big(k^*(\hat{\theta}), \theta_i\big) + t_i,$$

where $v_i\big(k^*(\hat{\theta}), \theta_i\big)$ is agent $i$'s valuation on the choice $k^*(\hat{\theta})$ when his type is $\theta_i$. The term $t_i$ denotes the payment rule and in this particular mechanism has the form:

$$t_i = \sum_{j \neq i} v_j\big(k^*(\hat{\theta}), \hat{\theta}_j\big) + h_i(\hat{\theta}_{-i}),$$

where $\hat{\theta}_{-i} = (\hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_I)$. In general, $h_i$ is an arbitrary function, but in the case of VCG mechanism is equal to the following:

$$h_i(\hat{\theta}_{-i}) = -\sum_{j \neq i} v_j\big(k^*_{-i}(\hat{\theta}_{-i}), \hat{\theta}_j\big),$$

where $k^*_{-i}(\hat{\theta}_{-i})$ is the social choice which has resulted from a mechanism with all agents excluding agent $i$. This particular formula for the function $h_i$ is called the pivotal or Clarke mechanism and reflects the contribution of agent $i$ to the community. The utility function has the final form:

$$\begin{aligned} u_i(x, \theta_i) = v_i\big(k^*(\hat{\theta}), \theta_i\big) + \\ + \Big(\sum_{j \neq i} v_j\big(k^*(\hat{\theta}), \hat{\theta}_j\big) - \sum_{j \neq i} v_j\big(k^*_{-i}(\hat{\theta}_{-i}), \hat{\theta}_j\big)\Big) \end{aligned} \quad (1)$$

If $k^*(\hat{\theta}) = k^*_{-i}(\hat{\theta}_{-i})$, which means that the reported type of agent $i$ does not change the social choice, then $t_i = 0$ and hence, $i$ is not charged. If $k^*(\hat{\theta}) \neq k^*_{-i}(\hat{\theta}_{-i})$, which means that agent $i$'s type changes the social choice (agent $i$ is pivotal), then $t_i < 0$. By allowing the payment rule $t_i$ to be negative, it is possible to have a mechanism with the following properties:

1. *ex post* efficient: the social welfare is maximised
2. incentive compatible: for all agents, true revelation of their type, i.e. $\hat{\theta}_i = \theta_i$, $\forall i \in I$ is a dominant strategy.

## IV. Our Scheme

Motivated by the work of Nix and Kantarcioglu in [1] we advance the potential of applying VCG mechanism in order to achieve privacy preserving collaborative classification. To comply with the game theoretic scenario, we assume a set of $I$ agents, each one possessing a data set $d_i$, under the assumption that the pairwise intersection of these sets is empty. All agents share the same strategy set:

$$S_1 = \ldots = S_I = \{\texttt{true}, \texttt{perturbed}, \texttt{randomised}\},$$

where `true`, `perturbed` and `randomised` correspond to an agent providing true, perturbed and randomised data, accordingly. The set $X$ of alternatives consists of the classification results. As explained in a previous section, the outcome of the mechanism, or in other words the collective choice, is that particular alternative which maximises the social welfare. In our scenario this is translated to achieving good classification results. As classification is a supervised mining task, this alternative corresponds to the accuracy of the classification, which measures the performance of the classifier. Following the notation of [1] we denote the classification accuracy on a data set $d$ by $acc(d)$. However, the lack of trust among the agents requires the introduction of privacy notions.

In our model, privacy is preserved by adding noise to the data values (perturbation). Although there are techniques to determine the distribution [12] and even to recover the true data from the noise [13], our experiment makes use of real data sets that do not have any particular trend, and thus those suggested methods for data recovery lead to poor results. For a clearer understanding of why this game theoretic approach succeeds, apart from the data perturbation, we also include complete randomization of the data, by replacing the true value with a random one. This random value is chosen from the interval formed by the minimum and maximum values of the attribute to be randomised. More formally, if $x_i$ is the true value then the randomised value is $\tilde{x_i} = t_i$, where $t_i \in [\min \texttt{attribute\_value}, \max \texttt{attribute\_value}]$. Regarding the perturbation, the method we use depends on the type of the data. For numeric attributes we have $x'_i = x_i + r_i$, where $r_i$ is chosen randomly from $[-a, a]$. If the attribute is of nominal type, then we use the AddNoise filter of the data mining toolset WEKA [14].

After the agents have decided on their preferences, their type is determined. The different types that we consider are: `per`, `rand`, `true`, where `per` describes an agent who provides perturbed data, `rand` corresponds to an agent who randomises the data and `true` represents an agent who is truthful. As all agents ideally prefer the extraction of information from true data, we regard their true type to be `true`, which corresponds to the accuracy of the classification on the union of the data sets $\bigcup_{i\epsilon I} d_i$ when all data is true. However, when an agent reports his type, the reported type $\hat{\theta}_i$ might not be the same as the true type $\theta_i$.

An important feature of the mechanism design concept is a trusted third party who acts as the authority that imposes the rules. This is the role of the mediator, who will perform the mining task and distribute the payoffs to each agent. If the mediator knew the true (private) type of the agents, then he could decide the outcome of the mechanism and distribute payoffs to the agents according to their types. However, as the types are private the particular form of the Clarke mechanism serves as an incentive for the agents to reveal the true type, and thus lead to a fair payoff distribution by the mediator. Rewriting the payoff function (1) using the accuracy, agent $i$ obtains the payoff:

$$u_i = acc(d) + \big(acc(\hat{d}) - acc(\hat{d}_{-i})\big), \tag{2}$$

with $d = \bigcup_{i\epsilon I} d_i$ being the union of the true data sets $d_i$, $\hat{d} = \bigcup_{i\epsilon I} \hat{d}_i$ is the union of the reported data sets $\hat{d}_i$ supplied by the agents and finally $\hat{d}_{-i} = \bigcup_{j\neq i} \hat{d}_j$, $i, j \in I$ corresponds to the data set formed from all data sets apart from the data of agent $i$. The expression

$$acc(\hat{d}) - acc(\hat{d}_{-i}) \tag{3}$$

calculates the loss or gain that agent $i$ poses to the overall outcome, in other words his contribution. Using the result of (3) as a reference point, we can determine whether the agent wishes to mask his data in order to keep it private, or his aim is to harm the ‘community’ by providing falsified data. More specifically, if the modification of his data results in classification accuracy that leads to (3) having a negative value, then his behaviour is considered harmful. However, if from the modification we obtain accuracy that keeps (3) non negative, then we infer that agent $i$’s intention is to preserve the privacy of his data without harming the overall outcome of the classification. Clearly, in an ideal situation agents would provide the true data and thus obtain high classification accuracy. However, as privacy is also required, the experimental results in the next section demonstrate that a controlled amount of perturbation results in both high accuracy levels and hiding of the data.

### A. Measuring Privacy

Since the preservation of privacy is equally significant to the extraction of information, truth telling is not a necessarily desired strategy. On the other hand, complete falsification results in poor information gain. Perturbation of the data is a reasonable compromise, but what is the limit of the perturbation range before reaches complete randomisation, and subsequently diminishes the information gain? The answer lies in the term $acc(\hat{d}_{-i})$ of (2) which indicates the accuracy that can be achieved using data sets from all agents except agent $i$. As long as the expression (3) remains non negative, the perturbation of agent $i$’s data causes insignificant reduction to the accuracy. If (3) becomes negative, then this is an indication that the perturbed data of agent $i$ harms the overall accuracy and hence, agent $i$ must obtain low payoff.

We suggest the following three different ways to measure privacy:

With respect to the distance from the true values:

$$\text{privacy} = \frac{|\text{perturbed value} - \text{true value}|}{|\text{randomised value} - \text{true value}|}$$

With respect to the range of the attribute values:

$$\text{privacy} = \frac{|\text{perturbed value} - \text{true value}|}{|\text{max value} - \text{min value}|}$$

With respect to the accuracy:

$$\text{privacy} = \frac{|\text{accuracy(perturbed data)} - \text{accuracy(true data)}|}{|\text{accuracy(randomised data)} - \text{accuracy(true data)}|}$$

Although in all cases the highest privacy is desirable, expression (3) poses a bound in the privacy that can be achieved without decreasing the agent’s payoff.

## V. Experimental Results

In support of the aforementioned model, this section presents our experimental results. The data set we used relates to the Civil War events in Africa, and was obtained from the Armed Conflict Location & Event Dataset [15]. For all data mining operations we used the toolset WEKA [14]. In particular, for the classification we applied the LibSVM to perform classification using the support vector machine method. Without loss of generality, we assumed that there are three agents with the following attributes:

Agent 1: $\{\texttt{year}, \texttt{source}\}$
Agent 2: $\{\texttt{actor1}, \texttt{actor2}\}$
Agent 3: $\{\texttt{latitude}, \texttt{longitude}\}$

All agents supply modified data, which can be either perturbed or randomised. We consider a small amount of perturbation for the attributes held by agents 1 and 2. In order to understand the sensitivity of the overall classification performance in terms of the amount of perturbation, we perform simulation study for a wide range of perturbation values while keeping the amount of perturbation on the data supplied by agents 1 and 2 fixed. We also study the classification performance when agent 3 completely randomises his attributes in order to understand the tradeoff between the privacy and the performance. The reason we choose agent 3 for greater modification of the data is due to his attributes `latitude`, `longitude` consisting of a wide range of real numbers. Moreover, the attributes of agent 3 form convex sets consisting of real numbers in the range of $[\min \texttt{latitude}, \max \texttt{latitude}]$ for the `latitude`, and $[\min \texttt{latitude}, \max \texttt{latitude}]$ for the `longitude`. Hence, the perturbed values also fall within these convex sets.

Let $x_i$ denote the true value of an attribute and $x_i'$ be the corresponding modified value. For perturbation of the numeric attributes, and particularly for the `year` attribute, we have that $x_i' = x_i + r_i$, with $r_i \in \{-1, 0, 1\}$. We perturbed both `latitude` and `longitude` as $x_i' = x_i + r_i$, with $r_i \in [-a, a]$. In our experiments we considered a range of perturbation as characterised by $a = 0.5, 1, 1.5, 2, 2.5, 3$. The nominal attributes (i.e. `source`, `actor1`, `actor2`) are perturbed using the AddNoise WEKA filter, with the noise parameter being 10%. For the randomisation of `latitude` and `longitude` $x_i' = t_i$, where $t_i$ is drawn uniformly at random from $[\min \texttt{latitude}, \max \texttt{latitude}]$ and $[\min \texttt{longitude}, \max \texttt{longitude}]$, respectively. In order to prevent overfitting we applied the 0.632 Bootstrap method [16] with 200 bootstrap samples, each one having the same size as the training set.

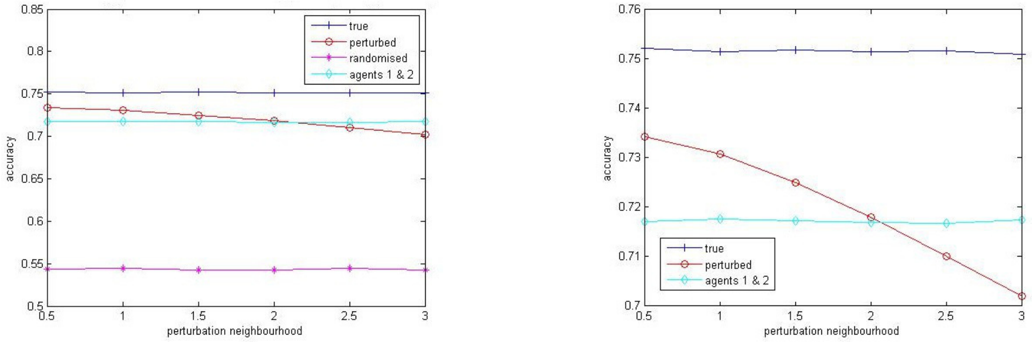
Figure 1(a) depicts the overall accuracy for four cases when the perturbation parameter $a$ takes the aforementioned values (a) all agents provide true data (legend —+—), (b) all agents perturb the data (legend –o–), (c) agent 1 and 2 supply perturbed data and agent 3 provides randomised data (legend –∗–) and (d) the classification is performed on the perturbed data of agents 1 and 2 only (legend –◇–). A closer look at these accuracies (Figure 1(b)) shows that between the ideal accuracy (which is achieved when all agents provide the true data) and a higher level of privacy (achieved when the data is perturbed), there is an interval where these two desired but contradictory properties are in balance. This interval lies between the accuracy of the classification on the true data and the output of (3), which is the accuracy that is achieved without the contribution of agent 3. Clearly, when agent 3 randomises, the resulting accuracy is significantly diminished.

Figures 2(a) and 2(b) depict the contribution (corresponding to the outcome of (3)) and payoff of agent 3. For a better understanding of these results, both figures present the charges and payoff, respectively, that result from the supply of perturbed data from agents 1 and 2, and true data from agent 3 (legend –◇–). As this situation offers the maximum payoff to agent 3, when he introduces perturbation in his data the charges increase and his payoff decreases. Perturbation of up to $a = 2$ (i.e., $2^0$ of perturbation) results in high payoff, and at the same time the data is concealed, as $2^0$ latitude is equal to 222km. Furthermore, both Figure 2(a) and 2(b) show that randomisation is not a beneficial approach due to very low payoff.

Regarding the privacy, Figure 3 shows the three different ways of measuring it. Clearly, the maximum privacy is achieved when the data is completely randomised. However, as randomisation results in poor classification accuracy, the actual maximum privacy that can be attained is represented by the line denoted by 'max-privacy-'. In all three subfigures, this line marks the critical point which separates the privacy with positive classification results from the privacy with undesirable classification results. Finally, Figure 4 presents a 3D overview of the relation among the perturbation, the accuracy and the privacy, for the three different privacy measures. The square on the figures denotes the critical point (as can also be seen in the intersection of those two curves in Figure 1(b)) where we have the maximum privacy while the accuracy of the classification is high and the perturbation of agent 3 is not harmful.

## VI. Conclusions

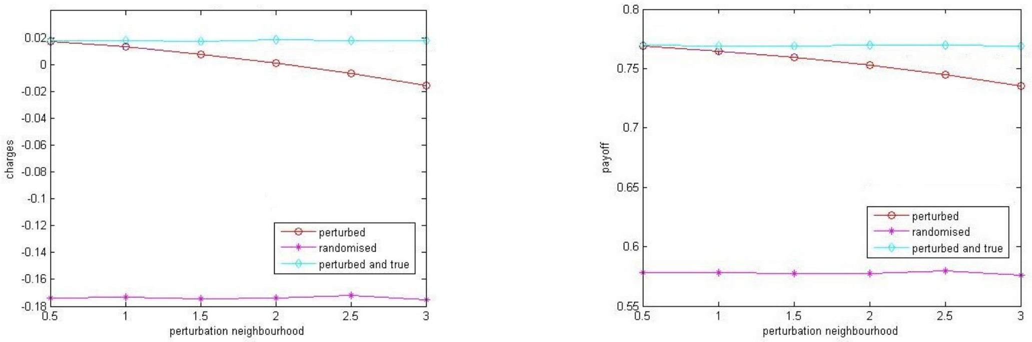
This work examined the problem of collaborative data mining using tools from game theory, while being able to offer data privacy to individual agents. In particular, motivated by [1] we used the Vickrey-Clarke-Groves mechanism in order to offer incentives that will prompt the agents to follow the rules. The behaviours that we considered are `true`, `per`, `rand`, corresponding to agent providing true, perturbed and randomised data. Our experiment showed that indeed the use of the VCG mechanism leads to high accuracy of the data mining task, while preserving the privacy of the data by allowing the agents to supply perturbed data. The key point of the VCG mechanism is that the gain of each agent includes the agent's contribution. Hence, the agent can perturb the data, as long as his contribution does not harm the overall result.



(a) The overall classification accuracy

(b) Magnified part of the overall accuracy, showing the interval where accuracy is high and the contribution of agent 3 is not harmful.

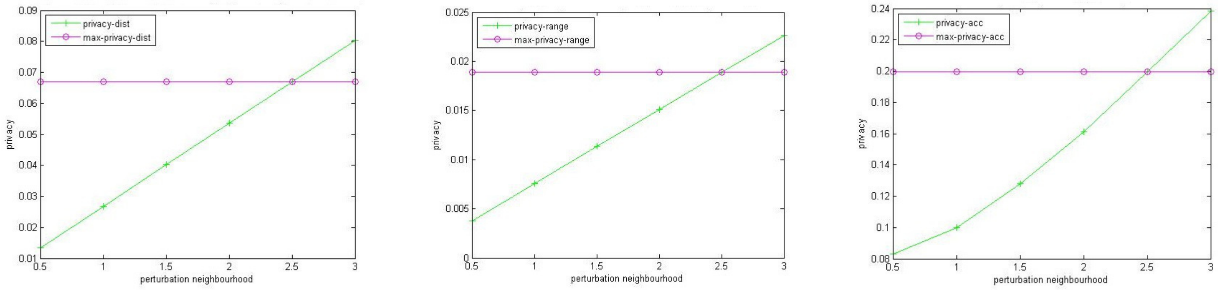
Fig. 1.



(a) The contribution of agent 3.

(b) The payoff of agent 3.

Fig. 2.



(a) With respect to the distance from the true values.

(b) With respect of the attribute values.

(c) With respect to the accuracy.

Fig. 3. Privacy

## REFERENCES

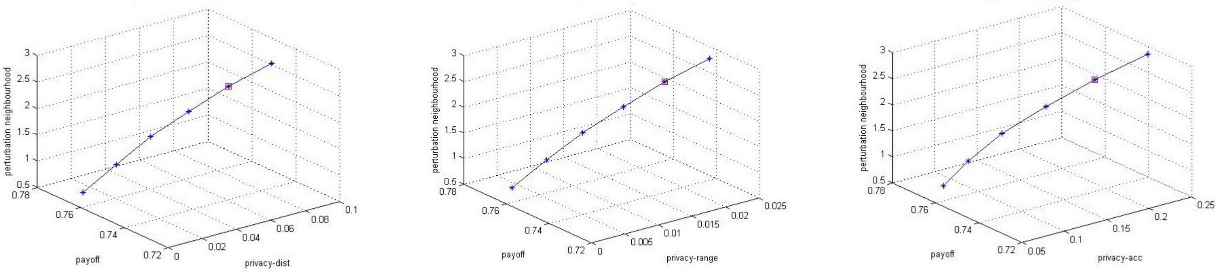
[1] R. Nix and M. Kantarcioglu, “Incentive compatible privacy-preserving distributed classification,” *IEEE Trans. Dependable Sec. Comput.*, vol. 9, no. 4, pp. 451–462, 2012.

[2] J. Halpern and V. Teague, “Rational secret sharing and multiparty computation: extended abstract,” in *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, 2004*. ACM, 2004, pp. 623–632.

[3] I. Abraham, D. Dolev, R. Gonen, and J. Halpern, “Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation,” in *Proceedings of the 25th annual ACM symposium on Principles of distributed computing*, ser. PODC ’06. New York, NY, USA: ACM, 2006, pp. 53–62.

[4] H. Kargupta, K. Das, and K. Liu, “Multi-party, privacy-preserving distributed data mining using a game theoretic framework,” in *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, ser. PKDD 2007. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 523–531.

[5] S. Izmalkov, S. Micali, and M. Lepinski, “Rational secure computation and ideal mechanism design,” in *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS ’05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 585–595.



(a) (b) (c)

Fig. 4. A 3D overview of the relation among the perturbation, the accuracy and the privacy, for the three different privacy measures.

[6] R. Agrawal and E. Terzi, "On honesty in sovereign information sharing," in *Proceedings of the 10th international conference on Advances in Database Technology*, ser. EDBT'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 240–256.

[7] R. Layfield, M. Kantarcioglu, and B. Thuraisingham, "Incentive and trust issues in assured information sharing," in *CollaborateCom*, 2008, pp. 113–125.

[8] O. Dekel, F. Fischer, and A. Procaccia, "Incentive compatible regression learning," *Journal of Computer and System Sciences*, vol. 76, no. 8, pp. 759–777, 2010.

[9] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. New York: Oxford University Press, 1995.

[10] D. Parkes, "Iterative combinatorial auctions: Achieving economic and computational efficiency," Ph.D. dissertation, Univesity of Pennsylvania, 2001.

[11] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*, 1st ed., ser. MIT Press Books. The MIT Press, 1994, vol. 1.

[12] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '00, vol. 29, no. 2. New York,USA: ACM, 2000, pp. 439–450.

[13] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proceedings of the Third IEEE International Conference on Data Mining*, ser. ICDM '03, Washington, DC, USA, 2003, pp. 99–.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[15] Raleigh, Clionadh, A. Linke, H. Hegre, and J. Karlsen, "Introducing acled-armed conflict location and event data," *Journal of Peace Research*, vol. 47, no. 5, pp. 1–10, 2010.

[16] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.