Privacy Consensus in Anonymization Systems via Game Theory

Rosa Karimi Adl, Mina Askari, Ken Barker, and Reihaneh Safavi-Naini

Department of Computer Science, University of Calgary, Calgary, AB, Canada
{rkarimia,maskari,kbarker,rei}@ucalgary.ca

Abstract. Privacy protection appears as a fundamental concern when personal data is collected, stored, and published. Several anonymization methods have been proposed to address privacy issues in private datasets. Every anonymization method has at least one parameter to adjust the level of privacy protection considering some utility for the collected data. Choosing a desirable level of privacy protection is a crucial decision and so far no systematic mechanism exists to provide directions on how to set the privacy parameter. In this paper, we model this challenge in a game theoretic framework to find *consensual* privacy protection levels and recognize the characteristics of each anonymization method. Our model can potentially be used to compare different anonymization methods and distinguish the settings that make one anonymization method more appealing than the others. We describe the general approach to solve such games and elaborate the procedure using k-anonymity as a sample anonymization method. Our simulations of the game results in the case of k-anonymity reveals how the equilibrium values of k depend on the number of quasi-identifiers, maximum number of repetitive records, anonymization cost, and public's privacy behaviour.

Keywords: Privacy Protection, Data Anonymization, Privacy/Utility Trade-off, Privacy Parameter Setting, Game Theory, k-Anonymity.

1 Introduction

Massive data collection about individuals on the Web raises the fundamental issue of privacy protection. A common approach to address privacy concerns is to use data anonymization methods [1–5]. During data anonymization identifiers are removed and data perturbation, generalization, and/or suppression methods are applied to data records.

Data anonymization promises privacy up to a certain level specified by some privacy parameter(s). In setting the privacy parameter, usually the amount of the expected data utility is considered and hence the level of privacy offered by an anonymization method is never set to the maximum. Since the risk to privacy is not completely removed, we postulate that data providers must be informed about the amount of privacy risk involved (represented as the privacy parameter's value) before deciding to provide their personal data to data collectors.

To bring data providers' privacy opinion into the cycle of data anonymization, we propose a game theoretic model that finds *consensual* privacy and utility levels by considering preferences of data providers as well as data collectors and data users. More

N. Cuppens-Boulahia et al. (Eds.): DBSec 2012, LNCS 7371, pp. 74–89, 2012.

specifically, we analyze the privacy/utility trade-off from the perspective of three different parties: a *data user* who wants to perform data analysis on a dataset and is willing to pay for it; a *data collector* who collects and provides privacy protected data to the data user; and *data providers* who can choose to participate in data collection if they see it as *worthwhile*. As these parties try to maximize their "profit" (payoff), the collective outcome of the game produces the equilibria [6] in our trade-off system. In an equilibrium state, no single player can achieve higher profits by changing their actions. Therefore, equilibria represent shared agreements (hence the term *consensus*) in which none of the players would attempt to behave differently. Using these equilibria, we are able to examine privacy trade-offs and analyze different characteristics of an anonymization technique such as the expected amount of privacy, precision, database size, and each party's profit. We believe that features of an anonymization technique must be inspected at equilibrium stages to provide more reliable evaluation results. The proposed model can be used as an evaluation framework to compare various anonymization methods from different perspectives. This is the first attempt to use game theory to analyze trade-offs in a private data collection system by considering preferences of data providers.

Paper Organization: The remainder of this paper is organized as follows: Section 2 discusses the related work. Section 3 describes basic definitions in game theory. Section 4 explains our game model and its ingredients. Section 5 provides a general solution to the game. Section 6 demonstrates a sample application of our model for the case of k-anonymity. Section 7 provides conclusions and suggests future directions.

2 Related Work

The issue of protecting individual's privacy while collecting personal information has motivated several research projects in literature. Our work mostly relates to anonymization techniques such as k-anonymity [1, 2], l-diversity [3], t-closeness [4], and differential privacy [5]. Anonymization techniques provide data privacy at the cost of losing some information. Several methods [7–11] have been proposed to evaluate the trade-off of privacy/utility. When data usage is unspecified, similarity between the original data and the privacy protected data is considered as information loss. The average size of equivalence classes [7] and discernibility [8] in k-anonymity are two examples of such generic metrics. However, most scholars have noticed that more reliable utility measures must be defined in the context of data application (*e.g.*,data mining and queries). Various measures of utility such as information-gain-privacy-loss ratio [9] and clustering and partitioning based measure [12] have been proposed to determine the next generalization step within anonymization algorithms. Sramka *et al.* [10] developed a data mining framework that examins the privacy/utility trade-off after the anonymization has been done using a mining utility. Machanavajjhala *et al.* [11] defines an accuracy metric for differential privacy in the context of social recommendation systems and analyzes the trade-off between accuracy and privacy. The existing privacy/utility trade-off methods all assume that a dataset already exists before choosing the privacy protection level for it. These methods do not consider the effect of privacy protection level on data providers' decision and hence the volume of the collected information.

In this work we use game theory to investigate steady levels of privacy protection by adopting a broader view of affecting parameters. Game theory has been successfully applied to analyze privacy issues from legal [13] and economic perspectives [14–17]. Kleinberg *et al.* [15] describe three scenarios modeled as coalition games [6] and use core and shapely values to find a "fair" reward allocation method in exchange for private information. The underlying assumption in these scenarios is that *any* amount of reward compensates for the loss of privacy protection. We believe this assumption oversimplifies the nature of privacy concerns and is not compatible with our perception of privacy. Calzolari and Pavan [16] use game theory to explore the optimum flow of customers' private information between two interested firms. The perspective of their work is possibly closest to ours but their model is substantially different from our work since they define a privacy policy as probability of revealing detailed customers' information to another party. Game theory has also been used as a means to address more technical aspects of privacy such as attacks on private location data [18], implementation of dynamic privacy [17], and questioning the assumption of honest behavior in multiparty privacy preserving data mining [19]. Our work builds on a commonly accepted definition of privacy among computer and social science scholars and adopts a game theoretical approach to find steady privacy levels. The novelty of our research lies on bringing the economic perspective to data anonymization issues and utilizing game theory for the first time to address privacy/utility trade-offs in a more realistic setting.

3 Preliminaries and Assumptions

In this paper we propose a game-theoretic framework to find steady level(s) of privacy protection for any arbitrary anonymization technique. We assume that the data providers are informed about having their personal information collected and the data collector is trustworthy in the sense that he fulfills his promises. Every instance of the game is modeled according to a chosen anonymization technique. A common factor between these techniques is a privacy parameter such as k in k-anonymity, l in l-diversity, and $1/\epsilon$ in differential privacy that indicates the level of privacy protection guaranteed by the corresponding privacy mechanism. To provide a generic game model, we use the letter δ to denote the privacy parameter. For any chosen anonymization technique, larger values for δ lead to higher privacy protection and lower data utility. The exact meaning of δ has to be interpreted according to the privacy definition chosen for the game. In this section we provide a brief overview of the game theoretic definitions used in this paper.

3.1 Sequential Game Model

Game theory is a mathematical approach to study interdependencies between individual's decisions in strategic situations (games). A game is explained by a set of *players* (decision makers), their *strategies* (available moves), and *payoffs* to each player for every possible strategy combination of all players (*strategy profile*). A strategy profile is a *Nash equilibrium* if none of the players can do better by changing their strategy assuming that other players adhere to theirs. Nash equilibrium is commonly used to predict stable outcomes of games and since it represents a steady state of a game [6], we use the term “stable” through the rest of the paper to denote the strategies found in the equilibrium. To capture a pre-specified order for players’ turn to move, a *game tree* is used to represent a *sequential game*. In this tree each node is a point of choice for a player and the branches correspond to possible actions. A sequence of actions from the root to any intermediate node or to a leaf node is called a *history* or a *terminal history*, respectively [6]. Payoff functions define Preferences of players over each terminal history. A player’s strategy explains his decision at any point in the game that he has to move.

Since the sequential structure of extensive form games is not considered in the concept of Nash equilibrium, the notion of “subgame-perfect Nash equilibrium” [6] is normally used to determine the robust steady states of such games. Every sub-tree of the original game tree represents a subgame. A strategy profile is a subgame perfect equilibrium if it induces a Nash equilibrium in every subgame [6]. The principle of *Backward induction* is a common method to deduce subgame-perfect equilibria of sequential games. Backward induction simply states that when a player has to move, he first deduces the consequences of every available action (how the subsequent player rationally reacts to his decision) and chooses the action that results in the most preferred terminal history.

The challenge of setting a desirable value for privacy parameter δ defines strategic situation with some ordering on players’ turn to move. As a result, we model the problem as a sequential game.

4 Game Description

To define a game-theoretic model for the challenge of finding a balanced value of δ, we must specify the decision makers (players), their preferences, and the rules of the sequential game. The following sections explain the details of our model.

4.1 Players

Players of the game are the following three parties:

Data Providers. Data providers are individuals that decide whether to provide their personal information at a specific privacy level δ and use the service offered by the data collector or to reject the offer. For example the service could be a discount on some online purchase activity or a software application offered for free. Since privacy preferences of each data provider is affected by several demographic and socioeconomic factors [20–22], it is practically infeasible to determine how much utility is gained by each data provider for each combination of δ and incentive. In an alternative approach, we rely on the assumption that data providers’ behavior is captured by a model based on some observation rather than a game theoretic analysis. Our assumed model is a regression model which captures how the number of data providers increases as the values of δ and incentive increase. Although this specific model has not been developed yet, similar studies have been conducted to explore the effects of other parameters (such as knowledge of privacy risks, trust, age, income level, *etc.*) on public’s privacy behavior [20, 22, 23]. A regression model that explains the effects of δ and incentive seems to be a natural extension to those studies. The assumed model generally considers data providers who are interested in both privacy and incentive and is defined as:

$$N = n(\delta, I) = \beta_0 + \beta_1\, h_1(\delta) + \beta_2\, h_2(I) \tag{1}$$

where N represents total number of individuals who accept the offer as a function of δ and incentive I (in terms of a monetary value). h_1 and h_2 are functions of δ and I. Parameters β_0, β_1, and β_2 are the intercept and marginal effects of $h_1(\delta)$ and $h_2(I)$ on individual's decision to participate in the data collection procedure. The functions h_1 and h_2 can be any non-decreasing functions of δ and I. This regression model does not assume accurate knowledge about privacy risks for data providers and as this knowledge increases, we expect to have larger β_1 to reflect a higher level of privacy concerns.

By assuming a regression model, we mostly *observe* data providers' behavior rather than directly *analyzing* it. This assumption *trims* the game tree by removing the data providers from the analysis of the game. Nevertheless, the effect of data providers' decisions is reflected in other players' payoff functions and paying specific attention to their impact on the final level of privacy is one of the distinctive strengths of our work.

Data Collector. A data collector is the entity who collects a dataset of personal data and provides it to some data users. The data collector receives offers from the data users, and based on their needs and the expected cardinality of the collected dataset announces a privacy level and some incentive to collect data from individuals. Once a data collector collects a dataset of personal information, he protects the privacy of the data providers at the consented level δ and provides the private dataset to the data user.

The data collector generally prefers to receive more money from the data user and spend less money on the amount of incentive he pays the data providers. Consequently, cardinality of the dataset (number of data providers) affects the payoff to the data collector. A detailed formulation of data collector's payoff is provided in Sect. 4.3.

Data User. A data user is an entity interested in accessing personal information for some data analysis purposes. A data user prefers a dataset with higher quality (more accurate query results) and higher cardinality (results with higher statistical significance). Privacy parameter δ affects these requirements in positive and negative ways. Therefore a data user chooses a value δ that balances the needs and initiates the game by offering some value for parameter δ and some price, p, for each data record. We give the detailed analysis for games with a single data user. The approach to model multiple data users and data reuse is explained elsewhere [24].

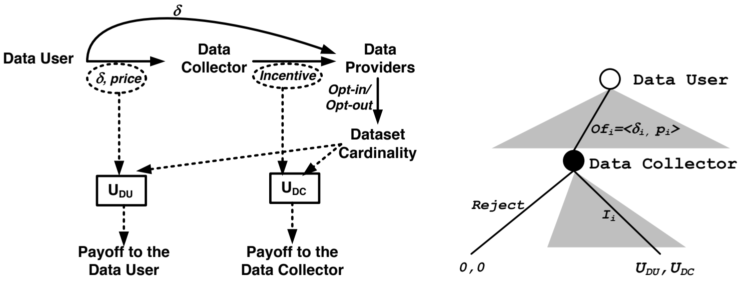
4.2 Game Rules

We model interactions between the data collector and the data user as a sequential game with *perfect* (players are aware of the actions taken by previous players) and *complete* (players are aware of the available strategies and payoff functions of all other players) information. More specifically, both players know data the provider's behavior model. The data user also knows the data collector's available actions and preferences[1].

The game starts with an offer from the data user to the data collector. In the offer, the required value for privacy parameter δ and the price p (per each record) must be

[1] Our assumption of complete information does not mean that the data collector and the data user know privacy/incentive trade-off functions of each data provider because individual data providers are not directly modeled as players in the trimmed game tree.

specified. We denote an offer by $Of = \langle \delta, p \rangle$. Once the data collector receives the offer he can either reject or accept it. In case of a rejection, the game terminates with payoff zero to both the data user and the data collector. If the data collector decides to accept then he needs to announce an incentive in exchange for collecting personal information. Here, we assume that I represents monetary value of the incentive and its domain is $\mathbb{R}_{\geq 0}$. The terminal histories of this game are either of the form (Of, I) or $(Of, Reject)$. At any terminal history, the number of data providers who will opt-in is determined by plugging the values of δ and I into Eq(1). Consequently, preferences of the data user and the data collector over all terminal histories are determined based on the payoff function defined over cardinality of dataset and values of δ, p, and I.



(a) (b)

Fig. 1. (a) The dynamics of setting a stable level for privacy protection. (b) Trimmed game tree.

The interactions and mutual effects of players' decision are captured in Fig. 1(a). Based on the game's dynamics, Fig. 1(b) illustrates the game tree (triangles represent ranges of possible offers and incentives).

4.3 Payoffs

Payoff to the Data Collector: The data collector receives some money, p, from the data user for each data record. The total number of data records in the dataset is the same as the number of data providers who participate in the data collection procedure and is defined by N in Eq(1). Consequently, the income of the data collector is:

$$income_{DC} = p\,N \tag{2}$$

Data collection procedure, data anonymization, and storing the dataset are costly and we denote these costs by C. Moreover, the data collector has to pay some incentive, I, to each data provider. As a result, the expenses to the data collector can be defined as:

$$expenditure_{DC} = I\,N + C \tag{3}$$

For simplicity of analysis we have assumed a fixed cost C for data collector. This assumption can be dropped easily by defining cost as a function of the size of the dataset and privacy level δ without any significant modification to our analysis. The payoff to the data collector is therefore defined as:

$$U_{DC} = income_{DC} - expenditure_{DC} = (p - I)\,N - C \tag{4}$$

Payoff to the Data User: The data user wants to run some data analysis on the privacy protected dataset T^*. As the cardinality of this dataset increases, the dataset will have higher value to the data user. Let a denote the economic value of each record to the data user, *i.e.*, a represents the net revenue of a data record if the data user gets the record for free. If the number of data records collected from individuals is denoted by N we can initially define the data user's income as $a * N$. However, after anonymization the utility of data drops due to imprecision introduced to results of the queries. We use parameter $0 \leq Precision \leq 1$ as a coefficient of the data user's income to show how the value of the dataset decreases as data become less precise. The income of the data user is:

$$income_{DU} = a\;N\;Precision \tag{5}$$

To estimate the precision of query results on a private dataset, various parameters must be considered. These parameters include the semantics of the query, the anonymization method and algorithm used, database schema, level of privacy protection δ, number of data records N, and *etc.*. For each instance of the game, all of these parameters except for δ and N are fixed (and assumed to be a common knowledge of the game). Therefore, $Precision = prec(\delta, N)$ is defined as a function of two variables δ and N. The main characteristic of the $Precision$ function is that for any **fixed** number of data records N, $Precision$ is a decreasing function of δ [2].

If the data user pays price p per record, his expenditure is $p\;N$ and therefore his payoff can be defined as:

$$U_{DU} = a\;N\;Precision - p\;N \tag{6}$$

5 General Approach to Find Subgame Perfect Equilibria

In this section we explain the steps involved in the process of finding the game's subgame perfect equilibria using backward induction [6]. In the next section, we show the details of this process for k-anonymity as an example.

5.1 Equilibrium Strategies of Data Collector

The first step to find subgame perfect equilibria is to find the optimal actions of the data collector in each subgame of length 1. Subgames of length 1 are represented by subtrees at which the data collector has to move based on a history of the form (Of). Where $Of = \langle \delta, p \rangle$ is an offer made by the data user.

The data collector can estimate the expected cardinality of the dataset for each δ and I based on Eq(1). If we plug this equation into the U_{DC} formula from Eq(4), the data collector's payoff after accepting $Of = \langle \delta, p \rangle$ will be:

$$U_{DC} = (p - I)(\beta_0 + \beta_1 h_1(\delta) + \beta_2 h_2(I)) - C \tag{7}$$

For each offer $Of = \langle \delta, p \rangle$, the values of δ and p are fixed. The data collector needs to find the optimum I (denoted by $\hat{I}$) for which the function U_{DC} attains its maximum value. To find $\hat{I}$ we must find the argument of the maximum:

$$\hat{I} = \arg\max_{I} U_{DC} = \arg\max_{I} (p - I)(\beta_0 + \beta_1 h_1(\delta) + \beta_2 h_2(I)) - C \tag{8}$$

[2] Notice that N is also an increasing function of δ (see Eq(1)) and therefore $\frac{\partial\, prec}{\partial\, \delta}$ is not always greater than or equal to zero.

Subject to the constraint that $\hat{I} \geq 0$.

If the maximum U_{DC}, $\hat{U}_{DC}$, is greater than zero the data collector accepts the offer. If $\hat{U}_{DC} = 0$ then the data collector will be indifferent between accepting and rejecting and in the case where $\hat{U}_{DC} < 0$ the data collector rejects. Therefore, the data collector's best response, BR_{DC}, to an offer $Of = \langle \delta, p \rangle$ is:

$$BR_{DC}(\delta, p) = \begin{cases} Reject & if\ (p - \hat{I})(\beta_0 + \beta_1 h_1(\delta) + \beta_2 h_2(\hat{I})) - C \leq 0 \\ Accept\ with\ \hat{I} & if\ (p - \hat{I})(\beta_0 + \beta_1 h_1(\delta) + \beta_2 h_2(\hat{I})) - C \geq 0 \end{cases} \tag{9}$$

The optimum incentive $\hat{I}$ must only be calculated when the data collector accepts the offer. This means $\hat{I} \leq p$, otherwise $\hat{U}_{DC} < 0$. Since U_{DC} is continuous in the closed and bounded interval $[0, p]$ (the domain of I), according to the Extreme value theorem [25], U_{DC} reaches its maximum at least once and therefore $\hat{I}$ is guaranteed to exist.

5.2 Equilibrium Strategies of Data User

The next step to find the subgame perfect equilibria is to find the most profitable action of the data user; Knowing the data collector's best response (Sect. 5.1) to each $Of = \langle \delta, p \rangle$, what combination of δ and p maximizes the data user's payoff? When the data collector accepts an offer $Of = \langle \delta, p \rangle$, he chooses the optimum incentive $\hat{I}$. Depending on the exact function definitions used in Eq(8), if $\hat{I}$ is unique for every combination of δ and p, then $\hat{I}$ can be defined as a function of δ and p (*i.e.*, $\hat{I} = \hat{i}(\delta, p)$). Without loss of generality, we assume that this is the case. If multiple values of I maximize U_{DC}, the one that also maximizes the data user's payoff is in the equilibria of the game.

According to Sect. 5.1, if the data collector accepts the offer he starts collecting personal information at privacy level δ with incentive $\hat{I} = \hat{i}(\delta, p)$. Otherwise, no dataset will be provided to the data user. As a result, the anticipated number of records N can be determined as:

$$N = n(\delta, \hat{I}) = \begin{cases} \beta_0 + \beta_1 h_1(\delta) + \beta_2 h_2(\hat{I}) & if\ \hat{U}_{DC} \geq 0 \\ \\ 0 & Otherwise \end{cases} \tag{10}$$

Plugging the function definition of $\hat{I} = \hat{i}(\delta, p)$ into Eq(10), $N = n_2(\delta, p)$ becomes a function of δ and p as well. Recall that $Precision = prec(\delta, N)$ is defined as a function of δ and N. Since N is a function of δ and p, we can define $Precision = prec_2(\delta, p)$ as a function of δ and p as well. After substituting N and $Precision$ with $n_2(\delta, p)$ and $prec_2(\delta, p)$, the U_{DU} function from Eq(6) becomes a function of two variables δ and p. The most profitable strategy for the data user is to choose values of δ and p that maximize his payoff:

$$\langle \hat{\delta}, \hat{p} \rangle = \arg\max_{\delta, p} U_{DU} = \arg\max_{\delta, p} \left(a\, prec_2(\delta, p) - p\right) \left(n_2(\delta, p)\right) \tag{11}$$

By definition, the lower bounds on p and δ is zero, *i.e.*, $p \geq 0$ and $\delta \geq 0$. Moreover, since $Precision \leq 1$ then $(a * prec_2(\delta, p)) \leq a$. Choosing a value $p > a$ leads to a negative payoff to the data user and he can always do better by choosing $p = 0$

(which leads to payoff zero). Therefore, the upper bound for p is a. Parameter δ is not necessarily bounded from above. Consequently, we cannot use the Extreme value theorem to guarantee an equilibrium.

If U_{DU} has an absolute maximum subject to the bounds defined on δ and p, the game has subgame perfect equilibria of the forms $((\hat{\delta}, \hat{p}), reject)$ or $((\hat{\delta}, \hat{p}), \hat{I})$. The first form occurs when the data collector cannot find any profitable amount of incentive (regardless of δ and p chosen by the data user) and the negotiation is unsuccessful. The second format occurs in games where there are at least one combination of δ and p of which the data collector can make profit. The two types of equilibria provide a means to determine whether an anonymization technique is *practical* or *impractical* given other problem settings. If the cost of implementing an anonymization technique is too high and the public's trust in the method is not high enough, the game might become an instance of unsuccessful negotiations and we have a case of impractical anonymization.

6 Game Theoretic Analysis for *k*-Anonymity

To demonstrate the details of the steps explained in Sect. 5, we use k-anonymity as the anonymization technique and provide a $Precision$ function for it. The game solution is described and a simulation of the results is provided at the end of this section.

6.1 k-Anonymity Overview

A dataset to be released contains some sensitive attributes, identifying attributes, and *quasi-identifying* attributes. Even after removing the identifying attributes, the values of quasi-identifying attributes can be used to uniquely identify at least a single individual in the dataset via linking attacks. Every subset of tuples in dataset that share the same values for quasi-identifiers is often referred to as an *equivalence class*. A released dataset is said to satisfy k-anonymity, if for each existing combination of quasi-identifier attribute values in the dataset, there are at least $k - 1$ other records in the database that contain such a combination.

There are several methods to achieve k-anonymity. Our work is built on Mondrian algorithm [26]. This greedy algorithm implements *multidimensional* recoding (with no cell suppression) which allows finer-grained search and thus often leads to a better data quality. In Mondrian algorithm all the identifying attributes are suppressed first. Then records are recursively partitioned into $d-$dimensional rectangular boxes (equivalence classes), where d is the number of quasi-identifiers. To partition each box, a quasi-identifier attribute (a dimension) is selected and the *median* value along this attribute is used as a binary cut to split the box into two smaller boxes. Once partitioning is done, records in each partition are generalized so that they all share the same quasi-identifier value, to form an equivalence class. A copy of this algorithm is provided in Fig. 3(b).

6.2 Data Providers' Privacy Model

Based on Sect. 4.1, we assume a regression model to explain data providers' reaction (at an aggregate level) to each combination of privacy protection levels and incentives.

This model is explained in Eq(1). In k-anonymity, privacy parameter is k. Here, we consider the identity function for the incentive (because of its simplicity) and logarithmic function for parameter k. In other words :

$$N = n(k, I) = \beta_0 + \beta_1 log_2(k) + \beta_2 I \tag{12}$$

To understand our choice of log function for h_1, notice that when k-anonymity is used, it is assumed that the probability of re-identifying an individual is $\frac{1}{k}$. For example, when k is 1, the probability of re-identification is 1 and the guaranteed privacy is 0. When k becomes 2, the probability of re-identification becomes $\frac{1}{2}$ and the amount of uncertainty about the identity of the individual increases from 0 ($log1$) to 1 ($log2$). However, this increase in uncertainty about the identity of individuals (privacy) is not the same as k changes from 99 to 100 because the probability changes from $\frac{1}{99}$ to $\frac{1}{100}$. For this reason we use entropy ($logk$) of this uniform probability distribution ($p = \frac{1}{k}$) as the indicator for privacy protection.

6.3 Precision Estimate

To determine the payoff to the data user (see Eq(6)) we need a metric to calculate $Precision$. A reasonable estimate on the amount of imprecision caused by anonymization depends on the data application. We have briefly discussed the nature of imprecision that can be introduced to the results of any SELECT query executed against an anonymized dataset elsewhere [24] . In this paper we provide the precision estimates for a specific SELECT query type and consider this query as the data analysis purpose. Our SELECT query is of the following form:

$\mathtt{Q}_i$ ≡ SELECT sensitiveAtt FROM T* WHERE q = $\mathtt{v}_i$

In this query sensitiveAtt represents the value of sensitive attribute, T^* is the anonymized dataset, q is one of the quasi-identifiers, and v_i is the i^{th} possible value for attribute q. For example, a query $\mathtt{Q}_{20}$ can be the following:

$\mathtt{Q}_{20}$ ≡ SELECT disease FROM T* WHERE age = 20

Let $|Q_i(T)|$ denote cardinality of the result set of query Q_i on dataset T. When Q_i is run against T^*, the result set $Q_i(T^*)$ contains two groups of records: a subset of them satisfy the condition q = $\mathtt{v}_i$ and the rest of them are just included in the result because they are partitioned into the same equivalence class as the points with q = $\mathtt{v}_i$. The latter introduce some quantity imprecision in the result. LeFevre *et al.* [27] introduce an imprecision metric to find the best cuts while running the Mondrian algorithm [26] on experimental datasets. After normalizing this metric, we define $Precision$ as:

$$Precision(Q_i, T^*, T) = \frac{|Q_i(T)|}{|Q_i(T^*)|} \quad \text{(where } |Q_i(T^*)| > 0) \tag{13}$$

As a result, to calculate $Precision$ we first need to estimate $|Q_i(T)|$ and $|Q_i(T^*)|$. Let Pr_i denote the portion of the records in the dataset that have value $\mathtt{v}_i$ for quasi-identifier q. Then the expected value of $|Q_i(T)|$ is:

$$|Q_i(T)| = Pr_i\, N \tag{14}$$

Through Theorems 1 and 2 we provide an estimate for $|Q_i(T^*)|$. In Mondrian algorithm the minimum and maximum number of records in each equivalence class are k

and $2d(k-1)+m$, where m denotes the maximum number of records with identical values for all quasi-identifiers [26]. Since the distribution of equivalence class sizes are not known *a priori*, with a simplifying assumption of uniform distribution, we can estimate the average number of records in each equivalence class, ec_{AVG}, as:

$$ec_{AVG} = \frac{2d(k-1)+m+k}{2} \tag{15}$$

Theorem 1. *If the average size of each equivalence class is determined by Eq(15), then the depth of the recursive calls, l, in Mondrian algorithm [26] can be estimated as:*

$$l = log_2(\frac{2N}{2d(k-1)+m+k}) \tag{16}$$

Proof. (*sketch*) Mondrian algorithm starts with the original dataset as a single equivalence class and chooses the *median* value of one of the dimensions to recursively cut each equivalence class into two smaller ones. It stops when there is no more possible cuts for any of the equivalence classes. For this estimate, we assume that the algorithm stops at the point where the size of each class reaches ec_{AVG} from Eq(15). By solving the recursive definition, we get Eq(16). A complete proof is available [24].

Theorem 2. *If N denotes the number of records in a dataset T, the cardinality of the result set of query Q_i on T^* can be estimated as:*

$$|Q_i(T^*)| = (1-\frac{1}{2d})^l \; N \tag{17}$$

where d is the number of quasi-identifiers and l is the depth of recursive calls estimated in Theorem 1.

Proof. (*sketch*) The core idea of this proof is to note that during the partitioning process, for each equivalence class if the dimension $\mathtt{q}$ is chosen as the cutting dimension then half of the records in the class will be partitioned into a new class that will not be included in the result set of Q_i. Otherwise the cut does not reduce the size of the result set. A complete proof is available [24].

Consequently, $Precision$ is defined as:

$$Precision = \frac{pr_i \; N}{(1-\frac{1}{2d})^l \; N} = \frac{pr_i}{(1-\frac{1}{2d})^l} \tag{18}$$

We can also use Theorem 2 to define pr_i based on the parameters. In real instances of the problem pr_i is independent of any specific algorithm and estimates; it is a property of the dataset. However, since we have made some simplifying assumptions for other estimates the assumptions should also be applied to pr_i to produce a meaningful estimate. Theorem 2 provides an estimate on $|Q_i(T^*)|$. When $k=1$, there are no irrelevant records in the result set. Therefore, $|Q_i(T^*_{k=1})|$ provides an estimate on the number of records that satisfy the condition $\mathtt{q} = \mathtt{v}_i$ and $|Q_i(T^*_{k=1})|/N$ can be used as an estimate for pr_i.

Consequently, we can refine Equation(18) as:

$$Precision = \frac{(1-\frac{1}{2d})^{log_2 \frac{2N}{m+1}}}{(1-\frac{1}{2d})^l} \tag{19}$$

6.4 Subgame Perfect Equilibria

As explained in Sect. 5.1, the first step to find the game's subgame perfect equilibria is to determine the optimum incentive $\hat{I}$ from Eq(8). If the data collector accepts the offer $Of = \langle k, p \rangle$ with incentive I, his payoff will be:

$$U_{DC} = (p-I)(\beta_0 + \beta_1 log_2(k) + \beta_2 I) - C \tag{20}$$

Calculating the derivative of U_{DC} with respect to I and setting it to zero reveals the maximizing I:

$$\frac{dU_{DC}}{dI} = -(\beta_0 + \beta_1 log_2(k) + \beta_2 I) + \beta_2(p-I) = 0 \ \Rightarrow \hat{I} = \frac{\beta_2 p - \beta_1 log_2(k) - \beta_0}{2\beta_2} \tag{21}$$

$\hat{I}$ is the local maximum since the second derivative of the function is negative. The restriction here is $I \geq 0$. If $\hat{I} < 0$, the maximizing I will be zero. The lower bound on I leads us to consider two separate cases:

Case 1: $\beta_2 p \geq \beta_1 log_2(k) + \beta_0$- In this case the amount of incentive that maximizes U_{DC} is $\hat{I} = \frac{\beta_2 p - \beta_1 log_2(k) - \beta_0}{2\beta_2}$. Plugging $\hat{I}$ into Eq(20) gives us the maximum payoff to the data collector for Case 1 (denoted as $\hat{U}^1_{DC}$):

$$\hat{U}^1_{DC} = \frac{\beta_2}{4}(p + \frac{\beta_1 log_2(k) + \beta_0}{\beta_2})^2 - C \tag{22}$$

The data collector will accept the offer $Of = \langle k, p \rangle$ if $\hat{U}^1_{DC} \geq 0$. In other words, the data collector accepts if:

$$p + \frac{\beta_1 log_2(k)}{\beta_2} \geq \sqrt{\frac{4C}{\beta_2}} - \frac{\beta_0}{\beta_2} \tag{23}$$

Case 2: $\beta_2 p < \beta_1 log_2(k) + \beta_0$- The optimum incentive in this case would be $\hat{I} = 0$. With this incentive the maximum payoff to the data collector (denoted as $\hat{U}^2_{DC}$) is:

$$\hat{U}^2_{DC} = p(\beta_0 + \beta_1 log_2(k)) - C \tag{24}$$

The data collector will accept this offer if $\hat{U}^2_{DC} \geq 0$. More precisely, the data collector accepts the offer if:

$$p(\beta_0 + \beta_1 log_2(k)) \geq C \tag{25}$$

If the values of $\hat{I}$ (from the two cases) are plugged into Eq(10), we can define the cardinality of the private dataset as a piecewise function of k and p:

$$N = \begin{cases} \frac{\beta_0 + \beta_1 log_2(k) + \beta_2 p}{2} & if\ \beta_2 p \geq \beta_1 log_2(k) + \beta_0 \ \wedge p + \frac{\beta_1 log_2(k)}{\beta_2} \geq \sqrt{\frac{4C}{\beta_2}} - \frac{\beta_0}{\beta_2} \\ \beta_0 + \beta_1 log_2(k) & if\ \beta_2 p < \beta_1 log_2(k) + \beta_0 \ \wedge p(\beta_0 + \beta_1 log_2(k)) \geq C \\ 0 & Otherwise \end{cases} \tag{26}$$

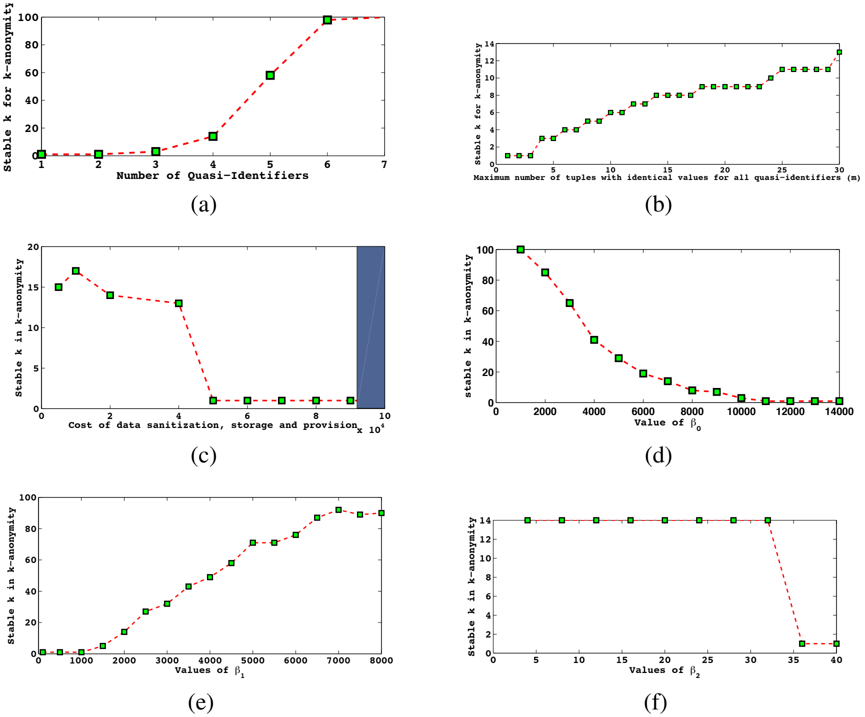


Fig. 2. Changes to the stable k due to an increase in: (a) the number of quasi-identifiers d; (b) the maximum number of data providers with identical values for their quasi-identifiers m; (c) the cost of data anonymization and storage C; (d) the number of privacy unconcerned data providers β_0; (e) the effect of privacy protection level on data providers' decision β_1; (f) the effect of incentive on data providers' decision β_2.

If the new definition of N is plugged into the $Precision$ function, precision becomes a function of k and p. As a result, U_{DU} from Eq(6) becomes a function of k and p. The best strategy for the data user is to compute $\hat{k}$ and $\hat{p}$ according to Eq(11). The optimum offer is $Of = \langle \hat{\delta}, \hat{p} \rangle$ and this completes the process of finding perfect equilibria.

6.5 Simulation Results

If the players of the game are rational and have the required information, the equilibria of the game would always conform to what Sect. 6.4 suggests because we used an analytical method to find the game's equilibria. In our proposed method, a dataset does not exist before the game is complete and the specifications of the collected dataset depend on the parameters chosen while the game is played. Therefore, running experiments on real databases does not provide meaningful results for this work. Alternatively, we choose to simulate the game and visualize the results by testing multiple parameter settings using MATLAB R2008a. In every setting, the effect of one of the parameters a, C,d, m, and β is examined on the stable values of k (while the values of the rest of the parameters are fixed to $\beta_0 = 7000$, $\beta_1 = 2000$, $\beta_2 = 20$, $a = \$10$, $C = \$20,000$, $m = 5$, and $d = 4$). The results are shown in Fig. 2.

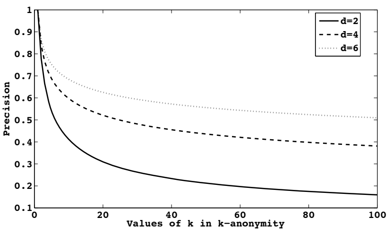
The values for a and C are randomly selected as an estimate of reasonable values commonly used in real instances of the problem. We assumed a population size of 55,000 potential data providers and the values selected for parameters β_0, β_1, and β_2 are chosen to reflect Westin's privacy indexes [28]. Based on the maximum values of k ($k = 100$) and p ($p = a$), β_1 and β_2 are chosen such that the effect of maximum privacy is almost the same as maximum incentive. The value of β_0 is chosen such that 17% of the data providers fall in the *privacy unconcerned* category [28].

Figure 2(a) shows how stable values of k increase as the number of quasi-identifiers increase. To understand the reason, we have provided another diagram in Fig. 3(a) which illustrates the precision curves for different values of d. According to this figure, with fewer quasi-identifiers the precision curve decreases at a higher rate. Therefore, as the number of quasi-identifiers increase, offering larger values for k becomes a better option for the data user since it can increase the size of the dataset without severely affecting data quality.

In Fig. 2(b) we can see the effect of m (maximum number of data providers with identical quasi-identifier values) on the stable values of k. We have chosen the values of m from $\{1, ..., 30\}$. As the value of m increases the stable value of k increases. To understand this counter-intuitive result, notice that as m increases less generalization will be needed to group the tuples in equivalence classes of size k. Therefore, compared to the cases with smaller m, the same precision can be achieved with higher values of k. Larger values of k attract more data providers without largely affecting the precision of query results and consequently, the data user can make more profit in this case.

The effects of anonymization, and maintenance cost (C) on stable values of k are illustrated in Fig. 2(c). Based on the settings chosen for other parameters, after a certain point the cost becomes too high for condition of the Eq(25) to be satisfied and case 1 (from Sect. 6.4) happens. In this case, the data collector is receiving a payment high enough to announce non-zero incentives. This incentive convinces several privacy concerned data providers to participate even with a low privacy protection level. As a result, the data user simply asks for no privacy protection since he is confident that enough data providers will participate to receive the incentive. Finally, after a certain value for C, the game reaches a point (demonstrated by a shaded rectangle) where no combination of $\langle k, p \rangle$ can be found that is acceptable by the data collector and $U_{DU} \geq 0$. This situation represents an instance of *impractical* anonymization.

Figures 2(d), 2(e), and 2(f) represent the effects of data providers' privacy attitude on stable values of k. According to Fig. 2(d) as the number of *privacy unconcerned* group (data providers who provide their personal information without any privacy or incentive) increase, the data user can receive larger volume of data without asking for sanitized dataset. By increasing the value of β_1 we model a privacy aware population. As can be seen in Fig. 2(e), when privacy has more significant impact on data providers' decisions, data will be sanitized with larger values of k. In Fig. 2(f) we showed how the value of β_2 impacts stable values of k. If β_2 is less than a certain level then it mostly affects the price of information and not the level of privacy protection. However if the weight of incentive on data providers' privacy decisions becomes heavier than a certain point, case 1 (refer to Sect. 6.4) happens and the data user can maximize his benefit by just increasing the price and asking for no privacy. These diagrams show how public's privacy awareness can force the firms to protect privacy of data providers.



(a)

```
Anonymize($partition$)
    if (no allowable multidimensional cut for $partition$)
        return $\phi : partition \rightarrow summary$
    else
        $dim \leftarrow$ choose_dimension()
        $fs \leftarrow$ frequency_set($partition, dim$)
        $splitVal \leftarrow$ find_median($fs$)
        $lhs \leftarrow \{t \in partition : t.dim \leq splitVal\}$
        $rhs \leftarrow \{t \in partition : t.dim > splitVal\}$
        return Anonymize($rhs$) $\cup$ Anonymize($lhs$)
```

(b)

Fig. 3. (a) Precision curves for different number of quasi-identifiers d. The value of m is fixed by 5. (b) Mondrian Algorithm.

7 Conclusions and Future Work

In this paper we modeled the process of private data collection as a sequential game to achieve consensus on the level of privacy protection. We explained the general approach to solve the game and as an example provided the details of game analysis for k-anonymity. Players of the game are a data user, a data collector, and a group of data providers. We use the method of backward induction to explore the game's subgame perfect equilibria. Equilibria of the game suggest stable values of the privacy parameter that are unlikely to be changed when other parties move according to their equilibria strategies. For the k-anonymity case, we found the stable values of k and showed that these values are related to number of quasi-identifiers, maximum number of identical tuples (in their quasi-identifier values), cost of data sanitization and storage, and coefficients of public's privacy behavior model. Our results illustrate the significant impact of the number of quasi-identifiers on the decision about the value of k.

We are plannig to analyze other privacy definitions such as l-diversity [3] and differential privacy [5] and for each privacy definition, distinguish the settings which make it the most profitable option to the players of the game. We are also planning to improve the model by dropping the assumption about the amount of information available to the data collector and data user. Our goal is to design a new evaluation framework that uses our game theoretic model to compare different anonymization methods and distinguish the settings that make one anonymization method more appealing than another.

References

1. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, p. 188. ACM Press (1998)
2. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 557–570 (2002)
3. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1(1), 24 pages (2007)
4. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: ICDE 2007, pp. 106–115 (2007)
5. Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006, Part II. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)

6. Osborne, M.J.: 8,9,16. In: An Introduction to Game Theory. Oxford University Press, USA (2003)
7. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization. In: KDD, pp. 277–286 (2006)
8. Bayardo Jr., R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: ICDE, pp. 217–228 (2005)
9. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: ICDE, pp. 205–216 (2005)
10. Sramka, M., Safavi-Naini, R., Denzinger, J., Askari, M.: A practice-oriented framework for measuring privacy and utility in data sanitization systems. In: EDBT/ICDT Workshops (2010)
11. Machanavajjhala, A., Korolova, A., Sarma, A.D.: Personalized social recommendations - accurate or private? CoRR abs/1105.4254 (2011)
12. Loukides, G., Shao, J.: Data utility and privacy protection trade-off in k-anonymisation. In: PAIS 2008, pp. 36–45. ACM (2008)
13. Anderson, H.E.: The privacy gambit: Toward a game theoretic approach to international data protection. bepress Legal Series (2006)
14. Böhme, R., Koble, S., Dresden, T.U.: On the viability of privacy-enhancing technologies in a self-regulated business-to-consumer market: Will privacy remain a luxury good? In: WEIS 2007 (2007)
15. Kleinberg, J., Papadimitriou, C.H., Raghavan, P.: On the value of private information. In: TARK 2001, pp. 249–257. Morgan Kaufmann Publishers Inc. (2001)
16. Calzolari, G., Pavan, A.: Optimal design of privacy policies. Technical report, Gremaq, University of Toulouse (2001)
17. Preibusch, S.: Implementing Privacy Negotiations in E-Commerce. In: Zhou, X., Li, J., Shen, H.T., Kitsuregawa, M., Zhang, Y. (eds.) APWeb 2006. LNCS, vol. 3841, pp. 604–615. Springer, Heidelberg (2006)
18. Gianini, G., Damiani, E.: A Game-Theoretical Approach to Data-Privacy Protection from Context-Based Inference Attacks: A Location-Privacy Protection Case Study. In: Jonker, W., Petković, M. (eds.) SDM 2008. LNCS, vol. 5159, pp. 133–150. Springer, Heidelberg (2008)
19. Kargupta, H., Das, K., Liu, K.: Multi-party, Privacy-Preserving Distributed Data Mining Using a Game Theoretic Framework. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 523–531. Springer, Heidelberg (2007)
20. Acquisti, A., Grossklags, J.: Privacy and rationality in individual decision making. IEEE Security & Privacy 3(1), 26–33 (2005)
21. Culnan, M.J., Armstrong, P.K.: Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. Organization Science 10, 104–115 (1999)
22. Singer, E., Mathiowetz, N.A., Couper, M.P.: The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. census. The Public Opinion Quarterly 57(4), 465–482 (1993)
23. Milne, G.R., Gordon, M.E.: Direct mail privacy-efficiency trade-offs within an implied social contract framework. Journal of Public Policy & Marketing 12(2), 206–215 (1993)
24. Adl, R.K., Askari, M., Barker, K., Safavi-Naini, R.: Privacy consensus in anonymization systems via game theory. Technical Report 2012-1021-04, University of Calgary (2012)
25. Sydsaeter, K., Hammond, P.: Mathematics for economic analysis. Prentice-Hall International (1995)
26. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: ICDE 2006, p. 25. IEEE Computer Society (2006)
27. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization techniques for large-scale datasets. ACM Trans. Database Syst. 33, 17:1–17:47 (2008)
28. Kumaraguru, P., Cranor, L.F.: Privacy indexes: A survey of westin's studies. ISRI Technical Report (2005)