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, 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 protect individuals' privacy before data publishing. Each anonymization method has at least one parameter to adjust the level of privacy protection. Choosing a desirable level of privacy protection is a crucial decision because it affects the volume and usability of collected data differently. In this paper, we demonstrate how to use game theory to model different and conflicting needs of parties involved in making such decision. We describe a general approach to solve such games and elaborate the procedure using k-anonymity as a sample anonymization method. Our model provides a generic framework to find stable values for privacy parameters within each anonymization method, to recognize the characteristics of each anonymization method, and to compare different anonymization methods to distinguish the settings that make one method more appealing than the others.

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

1 Introduction

The rapid growth of the Web and information technologies in recent years has resulted in massive data collection about individuals who use these technologies as part of their daily lives. Data collectors generally have two options to avoid privacy lawsuits: *privacy policy declaration* or *privacy protection* via data anonymization mechanisms. In the former approach, before collecting data, a data collector announces a privacy policy and data providers have to agree to the policy before providing their information. In the latter approach, a data collector collects private information from individuals by promising to protect their privacy. Then the data collector uses a data anonymization technique to transform original data before making it available to users of the data. On the one hand, a privacy policy declaration approach (such as P3P [1] and EPAL [2]) mainly limits the recipients and purposes of data usage. However, data privacy is often poorly protected against malicious recipients (or not protected at all). On the other hand, most anonymization methods [3–7] remove identifiers and apply some data perturbation, generalization, and/or suppression techniques on data

records. Due to a trade-off between privacy and *utility* [8–11] in data anonymization, the danger of privacy breaches is never completely removed. Therefore, even after data anonymization, the data collector is constantly faced with the risk of a privacy lawsuit filed against him.

To avoid the shortcomings of the classic approaches, in this work we adopt an intermediary approach: *the data providers' consent to some level of privacy protection are sought before collecting their personal information. data anonymization methods are then used to provide the promised level of privacy.* We believe that this approach is more ethically responsible and by requesting data providers' permission for privacy protection levels a market differentiator is achieved by the the data collector. With this approach, data providers are brought into the cycle of data anonymization and a new criterion is introduced to privacy/utility trade-offs. We use a game theoretic approach to model and analyze this balancing act.

We consider three 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*. We assume that each data provider is willing to contribute his/her data *at a specific level of privacy protection* if they are given sufficient *incentive* by the data collector. The incentive can be in the form of services that data providers receive or even direct payment. Data users are willing to pay higher prices for larger volumes of data with less changes to the original dataset (caused by the anonymization procedure). The data collector has costs including the cost of storage and infrastructure to run his business. Moreover, the data collector has to pay some incentive to the data providers to collect their data. These costs have to be justifiable by the amount of payment he receives from the data user.

In our model, the goal of the three parties is to maximize their "profit" (payoff). The collective outcomes of these maximizing efforts produce *stable states* in our trade-off system. These states are known as *equilibria* in game theoretic terminology [12].

Our game-based approach to deduce stable levels of privacy is general and can be used for different anonymization methods. We explain how to solve the game for any arbitrary anonymization method and demonstrate the procedure using k-anonymity [3, 4] as an arch type.

We consider the following interactions in the process of private data collection: A data user makes an *offer* to a data collector for a dataset. Since the data user knows he cannot access the original database (with no privacy protection), based on his most preferable balance between data quality and quantity, he asks for a privacy protection level that facilitates such balance. If the data collector accepts the offer, he announces some incentive to collect data at a specific privacy protection level. The combination of privacy protection level and incentive affects the number of data providers who opt-in. Once the database is collected, the data collector anonymizes it with the agreed level of privacy and provides the privacy protected database to the data user.

The underlying ordering of parties' actions is captured using a *sequential form game between rational players with perfect information* [12]. The data providers are considered players in the game who play after the data user and data collector. These players do not necessarily know the data collector and third-party's payoffs and strategies. We model the behavior of data providers as a group behavior. That is, although we do allow each data provider to have his/her own model of the world and accept different combinations of privacy-incentive, we assume that in steady states of the game the behavior pattern of the group is predictable by a regression model (assumed to be common knowledge of the game). The regression model explains the effects of increasing privacy protection and incentive on the number of data providers who are willing to share their private information. This allows the analysis of the game to focus on the data user and the data collector's strategies, taking into account the expected behavior of data providers. We define the sequential game between the parties, and construct payoff functions for the data user and the data collector.

We find the game's subgame perfect equilibria using backward induction [12]. In the end, the result of game analysis comes down to maximizing the data user's payoff function with respect to the privacy parameter and price per record. Notice that the stable values of the privacy parameter and the price found in this way do not necessarily result in maximum profit to both the data user and the data collector at the same time (*i.e.*, a Nash equilibrium is not necessarily Pareto optimal). They are the *outcome* of the game if both the data user and the data collector *try* to maximize their benefits. Therefore, the equilibria provide a means to predict the expected level of privacy that is provided in each situation.

Every equilibrium of the game is a combination of the data user and the data collector's strategies. An equilibrium strategy of the data user is a combination of values for the privacy parameter and the price he chooses to offer. An equilibrium strategy of the data collector is the amount of incentive he chooses in response to an offer from the data user (in case of an acceptance). We explain the steps involved in finding such equilibria (independent of a specific anonymization method) and demonstrate the details of the process for k-anonymity. Using k-anonymity as an example, we show how our game model provides a better understanding of an anonymization method. For instance, our synthesized simulations on k-anonymity show that for datasets with more quasi-identifier attributes or more identical records, higher privacy would be provided. These correlations are justifiable by our estimation of imprecision.

We have also shown how players' strategies change as the privacy protection/storage cost increases. When this cost increases, the data user must offer a higher price per record. This eventually leads to a higher incentive to data providers. The incentive convinces several privacy concerned data providers to provide personal information even with a lower privacy protection level. Therefore, higher cost of data maintenance and privacy protection indirectly leads to lower level of privacy protection. Finally, the results visualize how data provider's privacy awareness can encourage data collectors to improve their promised privacy protection level. If the data providers' decisions are mainly derived by the

privacy protection level, the data user maximizes his profit by asking for higher levels of privacy protection. On the contrary, when the role of the incentive becomes more prominent in data providers' decisions, the most profitable data user strategy is offering a higher price per record and asking for a lower privacy protection.

Contributions: We introduce a hybrid approach where an anonymization mechanism is used at a *consensual* protection level. In this approach the data collector provides the agreed upon privacy protection via an anonymization technique but is not liable for any further privacy breaches due to an adversary's background knowledge. This is the first work to use game theory to analyze trade-offs in a private data collection system by considering preferences of data providers. The model provides a framework to examine privacy trade-offs based on precision of the dataset, privacy protection level, and data providers behavior. Paying specific attention to the reaction of the data providers to different privacy protection levels is one of the novelties of this work.

We provide a high-level step by step approach to find stable levels of privacy protection for arbitrary data anonymization methods. The generic framework can be used as a benchmarking platform to compare data anonymization methods from various perspectives. As a demonstration, the concrete analysis of the game is described in the case of k-anonymity as the data anonymization technique. We used the Mondrian algorithm [13] for k-anonymity and discuss the nature of imprecisions that can potentially occur in the results of a general `SELECT` query. We then provide two detailed precision estimates for a specific `SELECT` query. The estimates are used in the data user's payoff function. These results are new and are of independent interest. We chose the Mondrian algorithm because its low time complexity and high quality results make it more likely to be used in practice.

Paper organization: We start by describing how this work relates to relevant literature and how it is distinct from previous work (Sect. 2). We then review some basic definitions in game theory in Sect. 3. The description of our game model is provided in Sect. 4 where we define the players, rules of the game and utility functions. A step-by-step approach to find the game's subgame perfect equilibria in provided in Setc. 5. We demonstrate how the model can be used through an example in Sect. 6. In our example we use k-anonymity as the anonymization method and by solving the game we show the effects of each parameter on the stable value of k. Finally, Sect. 7 provides the conclusion and suggests future directions and improvements to this work.

2 Related Work

The issue of protecting individual's privacy while collecting personal information has motivated several interesting research projects in literature. One of the most classic approaches to the problem is known as data anonymization. Data anonymization methods such as k-anonymity [3,4], l-diversity [5], t-closeness [6], and differential privacy [7] are built upon a simple philosophy; If a privacy pro-

tection mechanism is applied to the dataset no individual could be re-identified and providing such a privacy protected dataset to other parties does not raise privacy issues. However, since none of the methods can fully remove the risk to privacy, data sanitization alone cannot address all privacy issues.

We recognize anonymization techniques as necessary means to protect individuals' privacy but we also believe that data providers have the right to be informed about the amount of potential risks to their privacy before being asked for their personal information. The privacy parameter such as k in k-anonymity is a suitable indicator of privacy risks. Consequently, we use anonymization methods with a consensual privacy protection level.

Data anonymization methods provide data privacy at the cost of losing some information. Several methods have been proposed to evaluate the trade-off of privacy/utility either through the anonymization process or after it has been done. When data usage is unspecified, similarity between the original data and the privacy protected data is considered as information loss. For example, in k-anonymity, average size of equivalence classes [14] and discernibility [15] are two generic metrics which take equivalence class size into account to measure utility of a sanitized dataset. However, most scholars have noticed that more reliable utility measures must be defined in the context of data application such as data mining and queries. Various measures of utility such as information-gain-privacy-loss ratio [8], clustering and partitioning based measure [9], and risk-return trade-off [11] have been proposed to determine the next generalization step within anonymization algorithms. Sramka *et al.* [16] developed a data mining framework that considers the trade-off between the privacy and utility measure not in the process of anonymization but after the anonymization has been done using a mining utility. Machanavajjhala *et al.* [17] defines an accuracy metric for differential privacy in the context of social recommendation systems and analyzes the tradeoff between accuracy and privacy. Our approach differs from these classic trade-off measures since it considers the effects of the announced level of privacy protection on data providers' decision and hence the volume of 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 [18] and economic perspectives [19–22]. In an effort to measure value of private information, Kleinberg *et al.* [20] describe three scenarios modeled as coalition games [12] and use core and shapely values to find a "fair" reward allocation method in exchange for private information. The underlying assumption in the scenarios is that *any* amount of reward compensates for the loss of privacy protection. We believe this assumption over-simplifies the nature of privacy concerns and is not compatible with our perception of privacy. In another interesting study, Calzolari and Pavan [21] use game theory to explore the optimum flow of customers' private information between two interested firms. The perspective of this work is possibly closest to ours but the model is substantially different from our work since privacy protection is defined as revealing detailed customers' information

(microdata) to another party with some probability. Game theory has also been used as a means to address more technical aspects of privacy such as attacks on private location data [23] and implementation of dynamic privacy [22]. Our work builds up on a commonly accepted definition of privacy among computer and social science scholars and adopts a game theoretical approach to find stable 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

The focus of this work is to propose a generic game-theoretic framework to find acceptable level(s) of privacy protection for any arbitrary data anonymization mechanism. Our game model provides a means to analyze different characteristics of an anonymization method such as the expected amount of privacy, precision, database size, and each party's profit.

In our model, the data providers are fully informed about having their personal information collected. We assume no inappropriate behavior. In other words, the data collector is trustworthy in the sense that he fulfills his promises. Every instance of the game is modeled according to a chosen data anonymization method such as k-anonymity [3,4], l-diversity [5], t-closeness [6], and differential privacy [7]. A common factor between these methods 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 and explain the solution, we use the letter δ to denote the privacy parameter. For any chosen data anonymization method, 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 the rest of the paper.

3.1 Game theory

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 [12], we use the term "stable" through the rest of the paper to denote the strategies found in the equilibrium.

Noncooperative games are usually modeled in either the *normal* form or the *extensive* form. While normal form games capture situations where each player makes a decision without knowing other players' moves, extensive form is used

to model those games with a pre-specified order for players' turn to move captured by a tree. In this tree each node is a point of choice for a player and the branches correspond to possible actions. Every possible sequence of actions from the root to the leaves is called a *terminal history* and a path from the root to an intermediate node is simply referred to as a *history* [12]. Preferences of players over each terminal history are defined by payoff functions and 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" [12] 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 [12]. The principle of *Backward induction* is a common method to deduce subgame-perfect equilibria of extensive form 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 the right value for parameter δ within an arbitrary anonymization method can be viewed as multiple optimization problems to be solved by different decision makers and there exists some ordering on their turn to decide. As a result we model the problem using an extensive form game with perfect information (when a player chooses an action, he knows all the decisions made by other players who has moved before him). The next section describes the ingredients and rules of the game.

4 Game Description

The value of parameter δ must be chosen by considering its effects on quantity of records and utility of the anonymized dataset. As δ increases more generalization, perturbation, and/or suppression is applied to the records and hence the private database has lower utility to a data user. However, higher values of δ guarantee higher level of privacy protection and convince more data providers to share their personal information.

We can analyze the challenge of setting the value of δ from different viewpoints; data users who are interested in using the database for data analysis purposes, a data collector who gathers personal information and provides the private database to the data user, and data providers who provide their personal information to the data collector. A data user prefers to have a larger volume of data with less changes due to data anonymization. He offers values for δ and for price p (per data record). Based on the received offer, the data collector announces the value of δ and the amount of incentive he is willing to pay each data provider in exchange for their personal information. Finally, data providers decide whether they want to share their personal information (with the specified privacy protection level and incentive) or not. As the value of δ and incentive increase more data providers are willing to share their information.

Our game is modeled with a single data user. This choice for modelling does not hinder us from capturing multiple data users and data reuse. Our modelling strategy for data reuse is explained in Sect. 4.4. The interactions and mutual effects of each party's decision are illustrated in Fig. 1(a). Based on the dynamics, we define an extensive game model to analyze behaviors of the data user and the data collector. 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 those individuals who 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 [24–26], 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. Notice that this assumption does not cancel the effect of data providers' privacy decisions on the *stable* levels of privacy protection, and hence the game is different from a model with no data providers. 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 [24, 26, 27]. A regression model that explains the effects of δ (for each data anonymization method) and incentive seem 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.

Notice that we do not assume complete information for data providers (*i.e.*, data providers might not know the available actions and preferences of other players). Moreover, the 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.

We can look at the regression model as the summary of data providers, reaction to each sequence of actions taken by the data user and data collector. With this

perspective, our assumption *trims* the game tree by removing the data providers from the analysis of the game.

Data collector- In our game, a data collector is the entity who initiates some personal data collection procedure to provide it to some data users. We assume that the data collector knows data providers' behavior model. The data collector receives some offers from the data users, and based on their needs and the expected cardinality of the collected dataset announces a privacy protection level and some incentive (monetary value) to collect data from individuals. Once a data collector collects a dataset of personal information, he protects 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 defined as an entity interested in accessing personal information for some data analysis purposes. We assume that a data user is aware of data providers' behavioral model and data collector's available actions and preferences.

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 in Sect. 4.4. Description of data user's payoff is provided in Sect. 4.3.

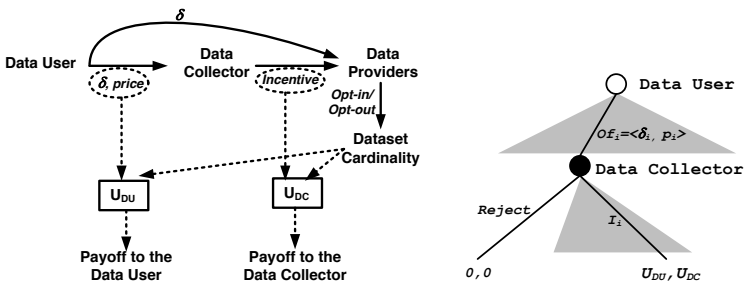


Fig. 1. (a) The dynamics of setting a stable level for privacy protection. (b) Trimmed game tree.

4.2 Game Rules

The game starts with an offer from the data user to the data collector. For each instance of the game we assume that the schema of the original relation and regression model of public's privacy behavior are known to both the data user and the data collector. The data user moves first by proposing an offer. In the offer, the required value for privacy parameter δ and price per each record are specified. Therefore, we can 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 in 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. Figure 1(b) illustrates the game tree. In this image we use the triangles to show 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 table 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:

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

However, data collection procedure, data anonymization, and storing the database are costly and we denote all these costs by C. Moreover, the data collector has to pay some incentive, I, to data providers. 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 C to be a fixed cost. 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:

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

According to Eq(1) N is a function of δ and I (*i.e.*, $N = n(\delta, I)$). Therefore, the payoff to the data collector is a function of δ, I, and p (*i.e.*, $U_{DC} = u_{dc}(\delta, I, p)$).
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 N we can initially define the data user's income as $a * N$. However, after anonymization data utility drops due to imprecisions 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 privacy definition and privacy protection algorithm used, database schema, level of privacy protection δ, number of data records N, *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. However, as mentioned in Eq(1) N is also an increasing function of δ and therefore $\frac{\partial\ prec}{\partial\ \delta}$ is not always greater than or equal to zero.

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

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

Note that N is calculated via Eq(1).

4.4 Modelling Data Reuse

If a second data user asks for an existing dataset that has already been gathered for some privacy protection level δ', depending on the precision requirement of the secondary data user, two situations can happen:

(a) The new data user chooses to offer with a privacy protection level $\delta' \geq \delta$: In this case the dataset can be provided to the new data user without the need for a new consent from data providers. A simple game can represent this situation. In the equilibrium, the data user asks for privacy level $\delta' = \delta$ (the data set size is fixed and higher privacy just causes lower data utility and lower payoff) with a minimum price that covers the extra cost C' of providing a collected dataset to a new data user.

(b) The new data user chooses to offer with a privacy protection level $\delta' < \delta$: Since a lower privacy protection is requested, it is the data collector's responsibility to ask the data providers again for a data reuse with lower privacy. This case becomes a new instance of the game where cost, C' is probably lower and the data provider's behavior model has different coefficients.

Knowing these two cases, a new data user can find the equilibria of the game in each case and after comparing his expected payoffs, choose the case that provides him with a higher overall benefit. A more comprehensive approach, must capture the competition between the data users. This model is in our plan for future extension to this work.

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 [12]. In the next section, we show the details of this process for k-anonymity as an example.

5.1 Equilibrium Strategies of Data Collector

According to the principle of backward induction, the first step to find subgame perfect equilibria is to find the optimal actions of the data collector in each subgame of length 1. Every subgame following the history of the form (Of) is a subgame of length 1, where Of is an offer made by the data user. At these subgames, the data collector has to move based on the $Of = \langle \delta, p \rangle$ received from the data user.

The data collector can estimate the expected cardinality of the dataset for each δ and I using 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 received 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:

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

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 the offer 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 as follows:

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

Notice that 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 [28], 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 (*i.e.*, the most profitable offer) considering the anticipated reaction of the data collector to each combination of δ and p (See Sect. 5.1). Based on Eq(8), 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.*, $\hat{I} = \hat{i}(\delta, p)$). Otherwise, Eq(8) defines a "relation" between $\langle \delta, p \rangle$ and $\hat{I}$. In the latter case, for each combination of $\langle \delta, p \rangle$ we have different incentives $\hat{I}_1,\ \hat{I}_2,\ ...,\ \hat{I}_j$ that maximize U_{DC} and therefore, the data user must examine all $\hat{I}_1,\ \hat{I}_2,\ ...,\ \hat{I}_j$ for each potential offer $Of = \langle \delta, p \rangle$. The incentive $\hat{I}_i$ that leads to the highest payoff to the data user is in the equilibria of the game. Here, without loss of generality we consider the case where $\hat{I} = \hat{i}(\delta, p)$ is a function (single output for each input) of $\langle \delta, p \rangle$ and the latter case becomes a simple extension of this case.

By analyzing data collector's best responses, we know that if the data collector accepts the offer he starts collecting personal information at privacy level δ with incentive $\hat{I} = \hat{i}(\delta, p)$. In this case, the number of data records in the collected dataset is expected to be $N = \beta_0 + \beta_1 h_1(\delta) + \beta_2 h_2(\hat{I})$. However, if the data collector rejects the offer then no dataset will be provided to the data user, *i.e.*, $N = 0$. As a result, the anticipated number of records N can be determined as:

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

Eq (10) defines N as a function of $\hat{I}$ and δ. As we discussed earlier $\hat{I}$ can be explained as a function of δ and p (*i.e.*, $\hat{I} = \hat{i}(\delta, p)$). Plugging the function definition of $\hat{I}$ in Eq(10), N becomes a function of δ and p as well ($N = n_2(\delta, p)$). Recall that $Precision$ is a function of δ and N ($Precision = prec(\delta, N)$). If we plug in the definition of $N = n_2(\delta, p)$ into the definition of the $Precision$ function we have $Precision = prec_2(\delta, p)$ defined 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:

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

By definition, the lower bounds on p and δ are zero, *i.e.*, $p \geq 0$ and $\delta \geq 0$. Moreover, since $Precision \leq 1$ then $(a * prec_2(\delta, p)) \leq a$. We can easily see that 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. However, 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 type of equilibria happens in the games where 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 type of equilibria happens 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 Sample Game Formulation for *k*-Anonymity

To demonstrate the details of the steps explained in Sect. 5, we use k-anonymity definition for privacy and provide a $Precision$ estimate for it. The game solution is described and a simulation of the results are 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 (and are indistinguishable from each other) 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. The basic techniques use hierarchical generalizations and cell suppressions. In all of these methods, the released anonymized dataset has all the identifying attributes suppressed and contains unmodified sensitive attributes. Our work is built on Mondrian algorithm [13] for k-anonymity. This greedy algorithm implements *multidimensional* recoding which allows finer-grained search and thus often leads to a better data quality. In Mondrian algorithm there is no cell suppression and the generalizations are not restricted by predefined generalization hierarchies. Instead, 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 based on a quality metric 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 :

$$h_1(k) = log_2(k) \; and \; h_2(I) = I \tag{12}$$

Consequently, the regression model becomes:

$$N = n(k, I) = \beta_0 + \beta_1 log_2(k) + \beta_2 I \tag{13}$$

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 $\frac{1}{k}$. As k increases this probability decreases. 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 data anonymization depends on the data application. Here, we briefly discuss the nature of imprecisions that can be introduced to the results of any `SELECT` query executed against an anonymized dataset. We then provide a precision estimate for a specific `SELECT` query type and consider this query as the data analysis purpose. A common `SELECT` query is of the following form:

```
SELECT select_list FROM table_names
[WHERE clause_group1]
[GROUP BY clause_group2]
[HAVING clause_group3]
```

The result set of such query can potentially have two types of imprecisions if it is executed on the anonymized dataset T^*: *value imprecision* and *quantity imprecision*. A value imprecision happens when the returned *value* of an attribute in the `select_list` or the output of an aggregate function is different if the query is executed on T^* instead of T (the original dataset). For example, if values of the attribute `age` are generalized to age ranges in T^* and the query asks for the values of `age` or `AVG(age)` then some value imprecision is introduced in the result set.

The `WHERE`, `GROUP BY`, and `HAVING` clauses generally help restrict the number of records included in the result set or simply organize the results. If the conditions specified in `clause_group1` or `clause_group3` are not aligned with the partitioning criteria in T^*, or `clause_group2` contains attributes that are generalized in T^*, then the number of records returned in the result set or in each group of the result set may be incorrect when the query is executed on T^*. We refer to this type of imprecision as quantity imprecision.

Estimates on these two types of imprecisions must consider the anonymization algorithm. Here, we only demonstrate the calculations for a specific `SELECT` query with potential quantity imprecision problem and use Mondrian algorithm for k-anonymization. Quantifying the amount of value and quantity imprecision for other types of queries is still an open question and on our agenda for future work. Suppose that a `SELECT` query of the following form is used by the data user:

$\mathtt{Q}_i \equiv$ `SELECT sensitiveAtt FROM` $\mathtt{T}^*$ `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:

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

Let $|Q_i(T)|$ denote cardinality of the result set of running 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 $\mathtt{q} = \mathtt{v}_i$ and the rest of them are just included in the result because they are partitioned into the same equivalence class as the points with $\mathtt{q} = \mathtt{v}_i$. The latter introduce some quantity imprecision in the result. LeFevre *et al.* [29] introduce an imprecision metric to find the best cuts while running the Mondrian algorithm [13] on experimental datasets. After normalizing this metric, we define *Precision* as:

$$Precision(Q_i, T^*, T) = \frac{|Q_i(T)|}{|Q_i(T^*)|} \tag{14}$$

Without loss of generality, we can assume $|Q_i(T^*)| > 0$ for the following reason: When data is partitioned into equivalence classes, the summary statistics of the equivalence classes (in our example, range of the attribute values) may refine attribute domains. For instance, If value $\mathtt{v}_i$ for attribute `q` does not match with the summary statistics of any of the equivalence classes then $\mathtt{v}_i$ is not in the refined domain of the attribute `q`. To measure precision, we only consider queries that seek for information within the refined domain of attribute `q`. For these queries we can still have $|Q_i(T)| = 0$ but we are guaranteed to have $|Q_i(T^*)| > 0$.

Part of the contribution of this paper is to give an estimate for *Precision* based on the value of k. This estimate is calculated for Mondrian algorithm [13] and is based on the estimates on the size of each equivalence class and the depth of the recursive calls in an execution of the algorithm.

Precision: 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:

$$|Q_i(T)| = Pr_i \; N \tag{15}$$

In Theorem 1 we use some facts about Mondrian algorithm [13] to estimate the depth of the recursive calls during anonymization. This estimate is then used in Theorem 2 to estimate $|Q_i(T^*)|$.

According to Lefevre *et al.*'s [13] second theorem, the maximum number of records in each equivalence class is $2d(k-1)+m$, where m denotes the maximum number of records with identical values for all quasi-identifiers. Moreover, in a k-anonymous dataset the minimum number of records in each class is k. 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:

$$ec_{AVG} = \frac{2d(k-1)+m+k}{2} \tag{16}$$

Theorem 1. *If the average size of each equivalence class is determined by Eq(16), then the depth of the recursive calls, l, in Mondrian algorithm [13] can be estimated as:*

$$l = log_2(\frac{2N}{2d(k-1)+m+k}) \tag{17}$$

Proof. The Mondrian algorithm starts with the original dataset as a single equivalence class and finds the best cut along one of the dimensions to cut the equivalence class into two equivalence classes. Since the split value is the median, if this value is not duplicated, splitting a partition with size ec produces two partitions of almost the same size ($ec/2$). If this is not the case, one partition will have the size $ec/2+\epsilon$ and the other one will have the size $ec/2-\epsilon$. In either case, the average size of these two new partitions is still $ec/2$. The algorithm then recursively cuts each of the two produced classes into 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(16).

At level 0, with no recursive call, the size of the class is N (the original dataset). Let $Size_l$ denote the size of each class after l recursive calls. The size of each class after $l+1$ recursive calls would be $Size_l/2$. Solving this recursive definition, we have:

$$Size_l = \frac{N}{2^l} \tag{18}$$

Since we assume that the algorithm stops when $Size_l$ reaches ec_{AVG}, we have:

$$\begin{array}{lll} Size_l = ec_{AVG} & \Rightarrow \\ \frac{N}{2^l} = \frac{2d(k-1)+m+k}{2} & \Rightarrow \\ l = log_2(\frac{2N}{2d(k-1)+m+k}) \end{array} \tag{19}$$

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:*

$$|Q_i(T^*)| = (1-\frac{1}{2d})^l \ N \tag{20}$$

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

Proof. If the depth of the recursive calls is zero then the whole dataset is returned as the result of the query $\mathtt{Q}_i$. Therefore $|Q_i(T^*)|_0 = N$. Let $|Q_i(T^*)|_x$ denote the cardinality of the result set when the depth of the recursive calls is x. If the algorithm goes one level deeper, then each of the overlapping classes from the previous stage are either cut by the median value along dimension `q` or other dimensions. For this estimate, we assume that all dimensions are chosen with equal probability. Therefore, the algorithm chooses dimension `q` with probability $1/d$ and other dimensions with probability $(1 - 1/d)$.

When the depth of the recursive calls is x the size of the result set is $|Q_i(T^*)|_x$. Assume that these records are scattered in s equivalence classes and denote the size of each class e_j as $|e_j|$. We have:

$$|Q_i(T^*)|_x = \sum_{j=1}^{j=s} |e_j| \tag{21}$$

For each of the classes e_j at the depth x, the expected size of the overlapping classes after splitting e_j at depth $x + 1$ can be estimated as:

$$|e_j|_{x+1} = \frac{1}{d}(\frac{|e_j|}{2}) + (1 - \frac{1}{d})|e_j| = (1 - \frac{1}{2d})|e_j| \tag{22}$$

The summation over all overlapping classes at the depth $x+1$, gives us $|Q_i(T^*)|_{x+1}$:

$$\begin{aligned} |Q_i(T^*)|_{x+1} &= \textstyle\sum_{j=1}^{j=s} |e_j|_{x+1} \\ &= \textstyle\sum_{j=1}^{j=s} (1 - \frac{1}{2d})|e_j| \\ &= (1 - \frac{1}{2d})|Q_i(T^*)|_x \end{aligned} \tag{23}$$

By solving the recursive formula we get $|Q_i(T^*)|_l = (1 - \frac{1}{2d})^l \ N$.

Consequently, *Precision* is can be defined as:

$$Precision = \frac{pr_i \ N}{(1 - \frac{1}{2d})^l \ N} = \frac{pr_i}{(1 - \frac{1}{2d})^l} \tag{24}$$

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 `q = v`$_i$. We have:

$$pr_i = \frac{|Q_i(T^*_{k=1})|}{N} = (1 - \frac{1}{2d})^{log_2 \frac{2N}{m+1}} \tag{25}$$

Now we can refine Equation(24) as:

$$Precision = \frac{(1 - \frac{1}{2d})^{log_2 \frac{2N}{m+1}}}{(1 - \frac{1}{2d})^l} \tag{26}$$

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) considering $N = n(k, I)$ (seeEq(13)) as the model of data provider's behavior. if the data collector accepts the offer $Of = \langle k, p \rangle$ with incentive I, his payoff will be:

$$U_{DC} = (p - I)(\beta_0 + \beta_1 log_2(k) + \beta_2 I) - C \tag{27}$$

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

$$\frac{dU_{DC}}{dI} = -(\beta_0 + \beta_1 log_2(k) + \beta_2 I) + \beta_2(p - I) = 0 \quad \Rightarrow \hat{I} = \frac{\beta_2 p - \beta_1 log_2(k) - \beta_0}{2\beta_2} \tag{28}$$

$\hat{I}$ 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 and the expected cardinality will be $\beta_0 + \beta_1 log_2(k)$. 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 value of I which maximizes U_{DC} is $\hat{I} = \frac{\beta_2 p - \beta_1 log_2(k) - \beta_0}{2\beta_2}$ and the maximum payoff to the data collector (in case of an acceptance) will be:

$$\begin{aligned} \hat{U}^1_{DC} &= (p - \tfrac{\beta_2 p - \beta_1 log_2(k) - \beta_0}{2\beta_2})(\beta_0 + \beta_1 log_2(k) + \beta_2 \tfrac{\beta_2 p - \beta_1 log_2(k) - \beta_0}{2\beta_2}) - C \\ &= \tfrac{\beta_2}{4}(p + \tfrac{\beta_1 log_2(k) + \beta_0}{\beta_2})^2 - C \end{aligned} \tag{29}$$

The superscript in the U_{DC} function is just to denote that the acceptance happened in Case 1.

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:

$$p + \frac{\beta_1 log_2(k)}{\beta_2} \geq \sqrt{\frac{4C}{\beta_2}} - \frac{\beta_0}{\beta_2} \tag{30}$$

If the data collector accepts then $I = \hat{I}$ and cardinality of dataset would be:

$$N = \beta_0 + \beta_1\ log_2(k) + \beta_2\ \hat{I} = \frac{1}{2}\ (\beta_0 + \beta_1\ log_2(k) + \beta_2\ p) \tag{31}$$

Case 2: $\beta_2 p < \beta_1 log_2(k) + \beta_0$- As mentioned earlier, the optimum incentive in this case would be $I = 0$. With this incentive the payoff to the data collector is:

$$\hat{U}^2_{DC} = p(\beta_0 + \beta_1 log_2(k)) - C \tag{32}$$

The superscript in the U_{DC} function is just to denote that the acceptance happened in Case 2.

Consequently, the data collector will accept this offer if $\hat{U}_{DC}^2 \geq 0$. More precisely, in case 2 the data collector accepts the offer and announces zero incentive if the following condition holds:

$$p(\beta_0 + \beta_1 log_2(k)) \geq C \tag{33}$$

If the data collector accepts then $I = 0$ and the cardinality of dataset would be:

$$N = \beta_0 + \beta_1 \ log_2(k) + \beta_2 \ 0 = \beta_0 + \beta_1 \ log_2(k) \tag{34}$$

Based on the two cases, the optimum value of incentive $\hat{I}$ can be described as a function of k and p as follows:

$$\hat{I} = \hat{i}(k,p) = \begin{cases} \frac{\beta_2 p - \beta_1 log_2(k) - \beta_0}{2\beta_2} & if \ \beta_2 p \geq \beta_1 log_2(k) + \beta_0 \\ 0 & Otherwise \end{cases} \tag{35}$$

Plugging this definition into Eq(10), we can define the cardinality of the private dataset as a piecewise function of k and p:

$$N = \begin{cases} \frac{\beta_0 + \beta_1 log_2(k) + \beta_2 p}{2} & if \ \beta_2 p \geq \beta_1 log_2(k) + \beta_0 \ \wedge p + \frac{\beta_1 log_2(k)}{\beta_2} \geq \sqrt{\frac{4C}{\beta_2}} - \frac{\beta_0}{\beta_2} \\ \beta_0 + \beta_1 log_2(k) & if \ \beta_2 p < \beta_1 log_2(k) + \beta_0 \ \wedge p(\beta_0 + \beta_1 log_2(k)) \geq C \\ 0 & Otherwise \end{cases} \tag{36}$$

If the new definition of N is plugged into the *Precision* function (see Eq(26)) precision becomes a function of k and p. As a result, U_{DU} from Eq(6) can be rewritten as 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 maximizing values for k and p represent the optimum offer and completes the process of finding subgame 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. A common experiment on data anonymization is to test a hypothesis on an existing dataset. However, in our proposed method, a dataset does not exist before the game is complete and the specifications of the collected dataset depends on the parameters chosen in the game. 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 value of k (while the values of the rest of the

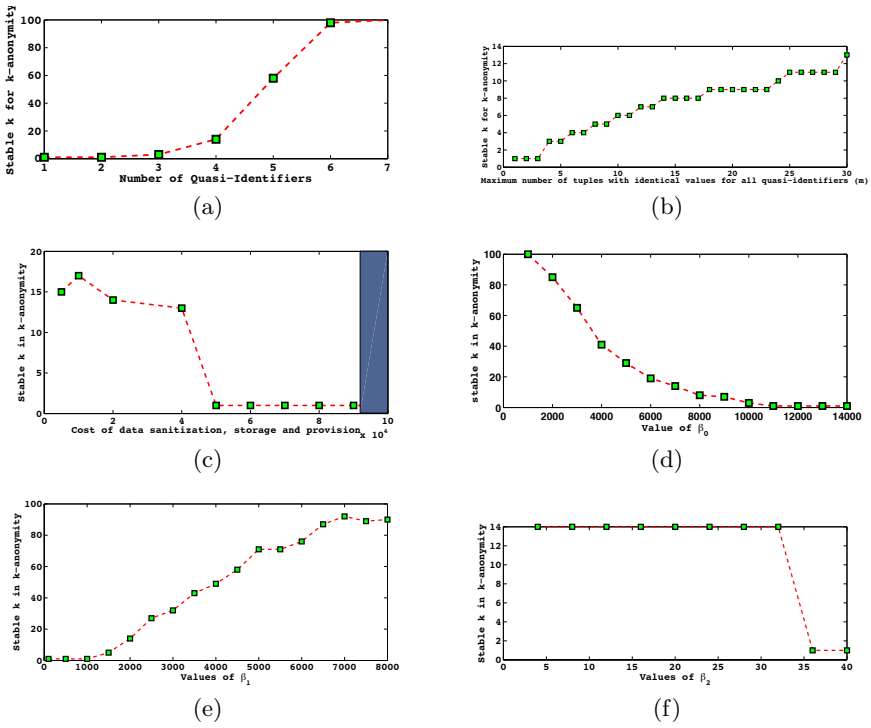
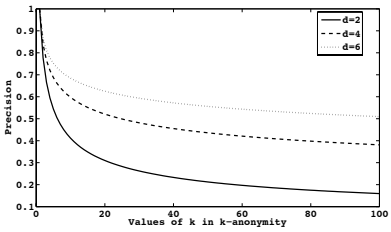


Fig. 2. Changes to the stable k due to an increase in: (a) the number of quasi-identifiers d; (b) the maximum number of data providers who have identical values for their quasi-identifiers m; (c) the cost of data sanitization, storage, and provision C; (d) in the number of privacy unconcerned data providers 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.

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(a), 2(b), 2(c), 2(d), 2(e), and 2(f).

In our implementation, 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 except for diagrams in Figures 2(d), 2(e), and 2(f) the values selected for parameters β_0, β_1, and β_2 are chosen to reflect Westin's privacy indexes [30] over this population. 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 on data providers' decision to opt-in. Moreover, the value of β_0 is chosen such that 17% of the data providers fall in the *privacy unconcerned* category [30].

Figure 2(a) shows how stable values of k increase as the number of quasi-identifiers increase. To understand the reason of such significant impact, we have



(a)

```
Anonymize(partition)
    if (no allowable multidimensional cut for partition)
      return φ : partition → summary
    else
      dim ← choose_dimension()
      fs ← frequency_set(partition, dim)
      splitVal ← find_median(fs)
      lhs ← {t ∈ partition : t.dim ≤ splitVal}
      rhs ← {t ∈ partition : t.dim > splitVal}
      return Anonymize(rhs) ∪ 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.

provided another diagram in Fig. 3(a) which illustrates the precision curves for different values of d. According to this figure, with more quasi-identifiers the *Precision* curve decreases at a higher rate for small values of k. Therefore, as the number of quasi-identifiers increase, the data user cannot expect high quality dataset even with small values of k and he tries to increase his profit by increasing the size of the dataset (offering larger values for k).

In Fig. 2(b) we can see the effects 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 $\{1, ..., 30\}$ since we believe that in practice this number cannot be very large otherwise the term quasi-identifier would not make sense. 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. Consequently, the data user can make more profit by asking for larger values of k.

The effects of different anonymization, and maintenance cost (C) on the 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(33) 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 the stable k. According to Fig. 2(d) as the number of *privacy un-*

concerned 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 anonymized with larger values of k. In Fig. 2(f) we showed how the value of β_2 impacts the stable values of k. Based on this diagram, 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.

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 based on the problem specifications. We explained the general approach to solve the game and as an example provided the details of game analysis when k-anonymity is used as an anonymization method. Players of the game are a data user who requires a private dataset for some data analysis, a data collector who collects private information and provides an anonymized version of the dataset to the data user, 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.

The ultimate piece of the game solution is a function of privacy parameter δ and price per record, which must be maximized. The maximizing values of the variables reveal the equilibria of our model. The game's subgame perfect equilibria provides a solution to the problem of setting a reasonable value for privacy parameter. Moreover, it can reveal valuable information about the characteristics of the anonymization method used. For example, in this paper we showed how the stable values of k (in k-anonymity) are related to number of quasi-identifiers, maximum number of identical tuples (in their quasi-identifier values), cost of data anoymization and storage, and coefficients of publics' privacy behavior model. Our results illustrate the significant impact of the number of quasi-identifiers on the decision about the value of k. We also show that even without government regulations a privacy-aware group of data providers can instigate improvements on privacy protection levels. Therefore, solving privacy/utility trade-offs without considering the opinion of data providers is an over-simplification of the problem. We also recognize the situations (based on data providers' privacy behaviour and anonymization cost) where k-anonymity becomes an impractical anonymization method to use. Our game analysis for k-anonymity is highly influenced by our choice in the query type and anonymiza-

tion algorithm. This fact reveals the underlying dependencies of stable values of k on the data usage query and anonymization algorithm and implies that no single value of k can be prescribed for all general problems.

Applicability of the results are subject to some limitations caused by our simplifying assumptions. The most influential assumptions are the data user and data collector's accurate information about public's privacy behavior and having fixed cost, C, for privacy protection and data maintenance. Dropping the first assumption will have a significant impact on the approach of solving the game. However, defining C as a function of cardinality and δ will only cause some modifications to the final curves but the process remains unaltered.

We are currently working on using the game model to analyze other anonymization methods such as l-diversity [5] and differential privacy [7] and for each method, distinguish the settings which make it the most profitable option to the players of the game. We are also planning on improving the model by dropping the assumption about the amount of information that is available to the data collector and data user (such as other players' payoff function and the public's privacy behavior model).

References

1. Cranor, L., Dobbs, B., Egelman, S., Hogben, G., Humphrey, J., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J., Schunter, M., Stampley, D.A., Wenning, R.: The platform for privacy preferences 1.1 (p3p1.1) specification. `http://www.w3.org/TR/P3P11/` (November 2006) W3C Recommendation.
2. Ashley, P., Hada, S., Karjoth, G., Powers, C., Schunter, M.: Enterprise Privacy Authorization Language (EPAL 1.2). Technical report, IBM (2003)
3. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, ACM Press (1998) 188
4. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems **10**(5) (2002) 557–570
5. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data **1**(1) (March 2007) 3+
6. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, IEEE 106–115
7. Dwork, C.: Differential privacy. In: ICALP (2). (2006) 1–12
8. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: ICDE. (2005) 205–216
9. Loukides, G., Shao, J.: Data utility and privacy protection trade-off in k-anonymisation. In: PAIS '08: Proceedings of the 2008 international workshop on Privacy and anonymity in information society, New York, NY, USA, ACM (2008) 36–45
10. Xu, J., 0009, W.W., Pei, J., Wang, X., Shi, B., Fu, A.W.C.: Utility-based anonymization for privacy preservation with less information loss. SIGKDD Explorations **8**(2) (2006) 21–30
11. Li, T., Li, N.: On the tradeoff between privacy and utility in data publishing. In: KDD, New York, NY, USA, ACM (2009) 517–526

12. Osborne, M.J.: 8,9,16. In: An Introduction to Game Theory. Oxford University Press, USA (August 2003)
13. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering. ICDE '06, Washington, DC, USA, IEEE Computer Society (2006) 25–
14. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization. In: KDD. (2006) 277–286
15. Jr., R.J.B., Agrawal, R.: Data privacy through optimal k-anonymization. In: ICDE. (2005) 217–228
16. 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)
17. Machanavajjhala, A., Korolova, A., Sarma, A.D.: Personalized social recommendations - accurate or private? CoRR **abs/1105.4254** (2011)
18. Anderson, H.E.: The privacy gambit: Toward a game theoretic approach to international data protection. bepress Legal Series (2006)
19. 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: Proceedings of Workshop on the Economics of Information Security (WEIS). (2007)
20. Kleinberg, J., Papadimitriou, C.H., Raghavan, P.: On the value of private information. In: Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge. TARK '01, Morgan Kaufmann Publishers Inc. (2001) 249–257
21. Calzolari, G., Pavan, A.: Optimal design of privacy policies. Technical report, Gremaq, University of Toulouse (2001)
22. Preibusch, S.: Implementing privacy negotiations in e-commerce. In Zhou, X., Li, J., Shen, H., Kitsuregawa, M., Zhang, Y., eds.: Frontiers of WWW Research and Development - APWeb 2006. Volume 3841 of Lecture Notes in Computer Science., Springer Berlin / Heidelberg (2006) 604–615
23. 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., Petkovic, M., eds.: Secure Data Management. Volume 5159 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2008) 133–150
24. Acquisti, A., Grossklags, J.: Privacy and rationality in individual decision making. IEEE Security & Privacy **3**(1) (2005) 26–33
25. Culnan, M.J., Armstrong, P.K.: Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. Organization Science **10** (January 1999) 104–115
26. 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) (1993) pp. 465–482
27. 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) (1993) pp. 206–215
28. Sydsaeter, K., Hammond, P.: Mathematics for economic analysis. Prentice-Hall International editions. Prentice-Hall International (1995)
29. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization techniques for large-scale datasets. ACM Trans. Database Syst. **33** (September 2008) 17:1–17:47
30. Kumaraguru, P., Cranor, L.F.: Privacy indexes: A survey of westins studies. ISRI Technical Report (2005)