University of Nevada, Reno

A Game Theoretic Approach Applied in k- Anonymization for Preserving Privacy in Shared Data

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Computer Science and Engineering

by

Anuraag Kotra

Dr. Shamik Sengupta - Thesis Advisor
May 2020

UNIVERSITY OF NEVADA RENO
THE GRADUATE SCHOOL

We recommend that the dissertation prepared
under our supervision by

Anuraag Kotra

entitled

A Game Theoretic Approach Applied in k- Anonymization for Preserving Privacy in Shared Data

be accepted in partial fulfillment of the
requirements for the degree of

Master of Science

Shamik Sengupta, Ph.D. – Advisor

Shahriar Badsha, Ph.D. – Committee Member

Hanif livani, Ph.D. – Graduate School Representative

David W. Zeh, Ph. D. - Dean, Graduate School

May 2020

Abstract

Privacy preservation is one of the greatest concerns when data is shared between different organizations. On the one hand, releasing data for research purposes is inevitable. On the other hand, sharing this data can jeopardize users' privacy. An effective solution, for the sharing organizations, is to use anonymization techniques to hide the users' sensitive information. One of the most popular anonymization techniques is k-Anonymization in which any data record is indistinguishable from at least k-1 other records. However, one of the fundamental challenges in choosing the value of k is the trade-off between achieving a higher privacy and the information loss associated with the anonymization. In this work, the problem of choosing the optimal anonymization level for k-anonymization, under possible attacks, is studied when multiple organizations share their data to a common platform which is data collector (Cybex) in this case. In particular, we have considered two common types of attacks, namely, Homogeneity attack and Background knowledge attack, which have the capability of compromising k-anonymization technique. To this end, a novel game-theoretic framework is proposed to model the interactions between the sharing organizations and the attacker along with contract theoretic framework to model interactions between organizations and data collector (Cybex). The problem is first formulated as a static game and its different Nash equilibria solutions are analytically derived. Later, we have used a contract theoretic model on interactions between data collector (Cybex) and the organizations. We also show how data collector varies the rewards of the organizations to increase it's utility over the stages.

Dedication

Dedicated to Jayapal Kotra

Acknowledgments

I would like to extend sincere thanks to my advisor, Dr. Shamik Sengupta, who was a constant source of inspiration and enlightenment required for all the phases of my research.

I also thank my committee members, Dr. Shahriar Badsha, and Dr. Hanif Livani, for taking time to review this thesis.

Ultimately, I thank my family, especially my parents for being pillars of financial and emotional support throughout my time in this school. I also thank my friends for their constant support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the Big Data era, vast amounts of data are constantly being generated, collected, and analyzed because of the ease of generating and distributing the data in its digital formats. Companies and organizations use the accumulated data to personalize their services, optimize their decision making, and predict future trends of the users [1]. However, these practices raise many public concerns about the users' privacy, especially as this data contains many pieces of personal and sensitive information. In response, organizations usually deploy powerful security mechanisms to preserve the users' privacy by protecting the stored data against different cyber attacks [2]. Similarly, encryption-based security systems were shown to be effective when data is shared between different locations of the same organization, e.g., patients' remote monitoring [3].

However, as organizations often need to share or publish the stored data, e.g., sharing electronic health records between different organizations, traditional security mechanisms cannot be used to protect the users' privacy as they are applied locally. This shortcoming was the main enabler for using data anonymization to hide the sensitive information within a dataset. For instance, information such as the name,

address, and phone number can be removed before sharing the data. Although, it was shown that the remaining data, after removing the sensitive information, can still be used to identify the users by figuring out the unique characteristics in this data [4]. Therefore, more effective anonymization techniques have been proposed in literature to preserve the privacy while withstanding possible attacks. The key idea behind such techniques is to ensure that the records of the shared datasets are indistinguishable. This is achieved by removing some information from the dataset to decrease the probability of identifying individual records. Examples of such techniques are k-anonymization [5], l-diversity [6] and t-closeness [7]. For instance, in k-anonymization the dataset is anonymized such that each record is indistinguishable from at least $k-1$ other records [5]. Both l-diversity and t-closeness are extensions to k-anonymization which make more changes to the dataset to make it harder to differentiate the records and the attributes.

However, even such models were shown to be prone to specific attacks such as background knowledge attack [8], in which the attacker uses background knowledge such as demographic information and public records to increase its probability of identifying the records. Since such types of attacks affect k-anonymization, l-diversity and t-closeness, and that k-anonymization is the basic technique behind l-diversity, and t-closeness, this work will mainly focus on k-anonymization. Another popular type of attack that affects k-anonymization, is the homogeneity attack [9] in which the attacker can reveal the private information when all the values of sensitive attributes are the same in one equivalence class.

One way to increase the privacy achieved by k-anonymization is to increase the value of k as this implies the need for differentiating each record from a bigger number of records. However, increasing the value of k will increase the information loss, i.e., more information will be removed from the data. This, in turn, will reduce the

value of the shared information, when received by other organizations. Therefore, the organizations need to carefully choose the value of k to maximize the privacy while minimizing information loss, which represents a real challenge. In [10], the authors proposed two algorithms to reduce the information loss associated with using k-anonymization. However, these algorithms depend on the structure of the data and cannot be generalized. To this end, choosing the optimal value of k, in k-anonymization, remains an open problem in privacy preserving.

1.1 Contribution

In this work, we propose a game-theoretic model to determine the value of k, in k-anonymization. Game theory is a powerful mathematical framework that enables studying the interactions between parties with opposing goals [11]. The key idea, here, is that each organization will choose the value of k that maximizes its outcome based on the expected attacks. Meanwhile, attackers can choose between different types of attacks based on their expected outcomes, when an organization chooses a specific k. While game theory has been used, in literature, to study privacy [12] and [13], such works do not apply to data anonymization and, hence, the problem of finding k requires its own analysis.

In this work, we developed a novel game-theoretic framework that allows the organizations to determine the optimal value of k, in k-anonymization. In particular, we consider a scenario in which more than one organization shares their data to a common platform. Each organization gets a reward from the common platform based on the level of anonymization, i.e., a higher level of anonymization will increase the information loss, and, hence, decrease the reward. The framework considers two types of de-anonymization attacks which are background knowledge and homogene-

ity attack. We formulate the problem as a static non-zero-sum game in which the organizations are considered as defenders that seek to optimize the choice of k, and an attacker that optimizes the selection of its attack, based on the choice of k. For the formulated problem, we analyzed the different cases of achieving a Nash equilibrium by considering both the cases in which pure equilibrium is possible, and the general case of mixed-strategy Nash equilibrium. Simulation results show that the proposed approach will enable the organizations to determine their optimal k value in the face of the expected attacks.

As, the sharing of data depends upon the rewards given by data collector (Cybex), we have formulated a Contract theoretic framework using optimal contract problems in [14], [15]. Contract theory is the study of how organizations construct and design legal arguments. Contract theory is modelled upon principles of economic and financial behaviour as different parties have different rewards on performing particular actions. The main goal is that the data collector (Cybex) should offer appropriate contracts in order to make organizations accept the contract and share the data. Initially, the data collector starts with zero utility, i.e., distributes it's income to the organizations by building contracts with maximum rewards in order to establish a connection. Later, it increases it's utility by reducing the rewards given to organization. An optimal repeated game heuristic is provided in Chapter 7 on how data collectors vary the rewards given to the organizations.

1.2 Contents

The following chapters of this thesis are as follows: Chapter 2 provides some background information on K-anonymization and Game theory. Chapter 3 provides the game formulation and defines the defender's and attacker's utilities. In Chapter 4, the

equilibrium analysis is derived for the formulated game. Contract-theoretic framework on how data collector designs the contracts is in Chapter 5 and Simulation results are discussed in Chapter 7. Finally, conclusions are drawn in Chapter 8.

Chapter 2

Background

In this Chapter, we will discuss few topics which are related to this research.

2.1 k- Anonymization

The research topic of privacy preserving data publishing has received much attention from various research communities and many such anonymization techniques have been proposed. In this work, we have considered the most popular technique, $k-$anonymization, as the strategy of data collector.

k-anonymization: In $k-$anonymization technique, the value of k is fixed before anonymization. Once the anonymization is done, information loss can be calculated. The original dataset is anonymized in such a way that each record is indistinguishable from at least $k-1$ other records [16], [17].For example, let us consider data of few students which consists zip Code, Age, Nationality as Non-sensitive attributes and their majors as sensitive attribute as shown in Table 2.1. In Table 2.2, we can observe that four records are grouped together and are anonymized in a way that they all are indifferent.

Table 2.1: Student Information

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Major
1	89503	28	Russian	Computer Science
2	89510	29	American	Computer Science
3	89510	21	Japanese	Statistics
4	89503	23	American	Statistics
5	91710	50	Indian	Chemical Engineering
6	91703	55	Russian	Computer Science
7	91710	47	American	Statistics
8	91703	49	American	Statistics
9	89503	31	American	Chemical Engineering
10	89503	37	Indian	Chemical Engineering
11	89510	36	Japanese	Chemical Engineering
12	89510	35	American	Chemical Engineering

2.2 De-Anonymization Techniques

In $k-$anonymization technique, the attributes are generalized or suppressed in a way that each row is indifferent with at least $k-1$ other rows. Thus, it is guaranteed that the probability of identification of data is at most $\frac{1}{k}$. Therefore, it can be attacked by either homogeneity attack or background knowledge attack.

Background Knowledge Attack: In the background knowledge attack, the attacker may have some background knowledge which might lead to eliminate some values from the dataset, so that, the attacker can compromise the k-anonymity with high probability [18]. For instance, in Table 2.2, if an attacker wants to know about a user, Bob, who is her neighbour and whose zip code is 895**, and the attacker also knows that he is 28 years old, then the probability of Bob having the sensitive attribute as **Computer Science** will be $\frac{1}{2}$ which is higher than **Statistics** and **Chemical Engineering** as it is only $\frac{1}{4}$.

Homogeneity Attack: An equivalence class may contain at least k indistinguishable records, there is a possibility that all records may contain same sensitive

Table 2.2: 4-Anonymous Student Information

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Major
1	895**	<30	*	Computer Science
2	895**	<30	*	Chemical Engineering
3	895**	<30	*	Statistics
4	895**	<30	*	Computer Science
5	917**	>40	*	Chemical Engineering
6	917**	>40	*	Computer Science
7	917**	>40	*	Statistics
8	917**	>40	*	Statistics
9	895**	>30	*	Chemical Engineering
10	895**	>30	*	Chemical Engineering
11	895**	>30	*	Chemical Engineering
12	895**	>30	*	Statistics

information, which will be very evident to the attacker to infer that sensitive information [6]. From Table 2.3, for example, Alice wants to know about Bob, whose zip code is 917**, as all the records grouped under zip code 917** have same sensitive attribute which is **Statistics**. Therefore, the attacker will be successful in breaching the data.

Table 2.3: 4-Anonymous Student Information-2

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Major
1	895**	<30	*	Computer Science
2	895**	<30	*	Chemical Engineering
3	895**	<30	*	Statistics
4	895**	<30	*	Chemical Engineering
5	917**	>40	*	Statistics
6	917**	>40	*	Statistics
7	917**	>40	*	Statistics
8	917**	>40	*	Statistics
9	895**	>30	*	Computer Science
10	895**	>30	*	Computer Science
11	895**	>30	*	Chemical Engineering
12	895**	>30	*	Chemical Engineering

2.3 Game theory

In [19], the author defined Game theory as a study of mathematical models of conflict and cooperation between intelligent rational decision-makers. In order to find the solution for a game theoretic model, one has to find an equilibrium point. Nash equilibrium can be a solution which can be defined as a set of actions of the players in which none of the players achieve higher utility by deviating without reducing the other player's utility.

Before we continue further, let us discuss a few game theoretic terms which are used in this work.

(i) Pure strategy and mixed strategy: A mixed strategy Nash equilibrium can be defined as at least one player playing a randomized strategy and no player being able to increase his or her expected payoff by playing an alternate strategy. A Nash equilibrium in which no player randomizes is called a pure strategy Nash equilibrium [20].

(ii) Static games and Dynamic Games: In static games, each player will get to play only once. The solution of the static game can be obtained by finding the nash equilibrium. Where as, in dynamics games or repeated games, the game is played in multiple stages. When playing a dynamic game, the static game strategy might not be an optimal solution. The players can get better payoffs in long run, by cooperate or defecting.

Chapter 3

Game Formulation

3.1 Players

Organizations: We consider a scenario in which two organizations share anonymized datasets with a common platform, a data collector as shown in 3.1.

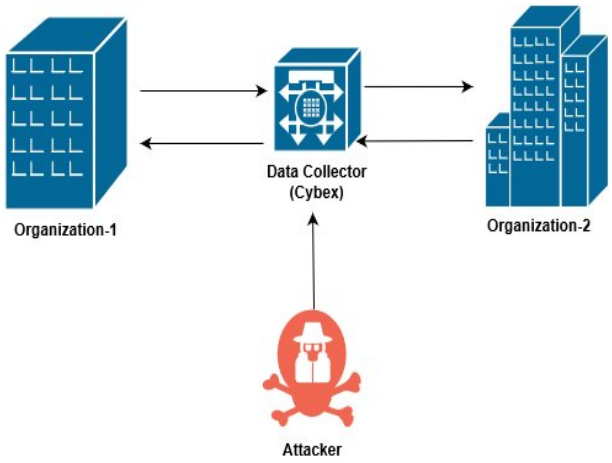

Figure 3.1: Interaction between two organizations of sharing data which also contains sensitive information and possibility of being attacked simultaneously

All organizations are assumed to use $k-$anonymization technique to make their shared data anonymous. The goal of organization i is to choose the best value of k_i

to maximize its payoff, given other organization's k values and the possible attacks on the data. Let $\mathcal{D}$ be the set of all organizations' actions.

Attacker: Attacker will attack the data, at the data collector side, in order to reveal the private data. We assume the attacker can anticipate the level of anonymization used, by analyzing the structure of the dataset. The attacker has three actions to choose from. Let $a \in \mathcal{A} = \{B, H, N\}$ represents the attacker's action which can be B, performing *background knowledge* attack, H, performing *homogeneity attack*, or N, no-attack.

Data Collector: In this scenario, Organizations share the anonymized data to a common platform in order to get some reward which is called as data collector. The data collector performs data mining on this data and gets benefits from this data. This can be done by any data collector like CYBEX. In [21] and [22], the authors mentioned that multiple organization share cyber security information in a CYBEX framework and privacy is one of the major concerns of the organization while sharing.

3.2 Payoffs

Each player wants to maximize its payoff function based on its actions and the other players' actions. Each player's payoff is given by its utility function which defines its outcome in light of its action and other players' actions.

Organizations: The utility of each organization is given as a function in the reward it gets from the data collector, $r_i(K_i)$, the cost for applying the anonymization technique, $c_i(k_i)$, and the probability of data breach, $b_i(k_i, k_{-i}, a)$, where k_{-i} refers to the other organization's action. Let u_i be the utility of organization i, it can then be given by:

$$u_i(k_i, k_{-i}, a) = r_i(k_i) \cdot (1 - b_i(k_i, k_{-i}, a)) - c_i(k_i) + T(k_i), \tag{3.1}$$

where the first term represents the probability of receiving the reward based on all the players' actions.

Next, we discuss, in detail, each term in (3.1). First, the reward function $r_i(k_i)$ is defined as a declining function in k_i such that when the level of anonymization increases, the data collector will give less reward to the organization as the data will be less informative. We propose to define $r_i(k_i)$ as:

$$r_i(k_i) = \frac{1}{k_i} R_i, \tag{3.2}$$

where R_i is the value of the information at organization i. By using (3.2), when $k_i = 1$, i.e., no anonymization, the organization can obtain the full value of the reward as $r_i(1) = R_i$. For every $k_i > 1$, the reward will be declining such that, for large values of k_i, e.g., $k_i > 10$, any increase in k_i will cause a small decrease in r_i. This can be interpreted as when the anonymization level increases, the information will be less useful up to some point where the increased k_i will have a very small effect on the information loss (reward). This can be captured by the heavy tail of the function in (3.2).

Next, we study effect of choosing k_i on the cost function $c_i(k_i)$. This cost represents the computational cost for each organization as it executes the k-anonymization procedure. This cost was shown in [23] to depend on the nature of the data under consideration and was proven to be [23]:

$$O(nm + 2^{t_{\text{in}}.t_{\text{out}}}(t_{\text{in}}.t_{\text{out}}.m + (t_{\text{in}} + t_{\text{out}})\log(t_{\text{in}} + t_{\text{out}})(t_{\text{in}}.t_{\text{out}} + (t_{\text{in}} + t_{\text{out}})\log(t_{\text{in}} + t_{\text{out}})))), \tag{3.3}$$

where n denotes the total number of rows, m is the total number of columns, t_{in} is the number of different input row types, t_{out} is the number of different output row

types such that $\frac{t_{\text{in}}}{t_{\text{out}}} = k$.

Here, we use this time complexity to represent the cost associated with k-anonymization in (3.1) as follows:

$$c_i(k) = \beta\Big(n.m + 2^{t_{\text{in}}.t_{\text{out}}}\Big(t_{\text{in}}.t_{\text{out}}.m + (t_{\text{in}} + t_{\text{out}}).\log(t_{\text{in}} + t_{\text{out}})(t_{\text{in}}.t_{\text{out}} + (t_{\text{in}} + t_{\text{out}})\log(t_{\text{in}} + t_{\text{out}}))\Big)\Big), \tag{3.4}$$

where β is a conversion factor from the time complexity to a monetary value.

Substituting $\frac{t_{\text{in}}}{t_{\text{out}}} = k$ in (3.4), the cost can be written as:

$$c_i(k) = \beta\Big(n.m + 2^{k.(t_{\text{out}})^2}.\Big(k.(t_{\text{out}})^2.m + t_{\text{out}}(k+1).\log(t_{\text{out}}(k+1))(k.(t_{\text{out}})^2 + t_{\text{out}}(k+1).\log(t_{\text{out}}(k+1)))\Big)\Big) \tag{3.5}$$

Next, we discuss about the trust factor, which is the incentive for the organization for maintaining the security of the cybex network and depends upon the value of k_i:

$$T(k_i) = \gamma \cdot k_i \tag{3.6}$$

where, γ is the co-efficient of trust.

Finally, we consider the breach probability for each organization's shared data, $b_i(k_i, k_{-i}, a)$. In [24], it was shown that the information breach probability is given by:

$$b(a, k_i) = \frac{p(a)}{\alpha k_i + 1}, \tag{3.7}$$

Where, $p(a)$ is probability of successful attack, based on the attack type, and $\alpha > 0$ is a measure of information security. Note that, in [24], the probability of breach is given as a function in the organization's investment. Here, we assume k_i represents the organization's investment in protecting its shared data.

Equation (3.7) represents the organization's own probability of breach. In the case of multiple organizations sharing to the same platform, this will increase the probability of successful attack as an attacker can link information from different datasets to identify the records [25]. Here, we propose to model this interdependency similar to the model in [26] such that the interdependent risk between the organizations is given as:

$$b(a, k_i, k_{-i}) = 1 - (1 - \frac{p(a)}{\alpha k_i + 1})(1 - \frac{p(a)}{\alpha k_{-i} + 1}). \tag{3.8}$$

Substituting (3.8) in (3.1), the utility of any organization can then be given as:

$$u_i(k_i, k_{-i}, a) = \frac{R_i}{k_i} \cdot (1 - \frac{p(a)}{\alpha k_i + 1}) \cdot (1 - \frac{p(a)}{\alpha k_{-i} + 1}) - c_i(k_i) + \gamma \cdot k_i. \tag{3.9}$$

Note that, in (3.9), increasing one organization's k_i will reduce its reward; however, this will increase the probability of achieving this reward at this and other organizations.

Attacker: For the case of two organizations, the attacker's utility can be given in terms of its probability of achieving the reward from the information and the cost to apply its attack. Thus, we define the attacker's utility u_a as follows:

$$u_a(k_1, k_2, a) = b(a, k1, k_2)R_a - c_a(a), \tag{3.10}$$

where R_a is the reward from revealing the real data and this reward can be achieved based on the combined breach probabilities of the datasets and $c_a(a)$ is the cost of performing each type of the attack.

Note that (3.8) can be rewritten as:

$$b(a, k_1, k_2) = \frac{p(a)}{(\alpha k_1 + 1)} + \frac{p(a)}{(\alpha k_2 + 1)} - \frac{p(a)}{(\alpha k_1 + 1)} \frac{p(a)}{(\alpha k_2 + 1)}, \tag{3.11}$$

and thus the attacker's utility in (3.10) can be given as:

$$u_a(k_1, k_2, a) = \Big(\frac{p(a)}{(\alpha k_1 + 1)} + \frac{p(a)}{(\alpha k_2 + 1)} - \frac{p(a)}{(\alpha k_1 + 1)}\frac{p(a)}{(\alpha k_2 + 1)}\Big)R_a - c_a(a), \quad (3.12)$$

Here, according to the nature of homogeneity attack, the attacker will benefit if the two organizations are using the same anonymization level. This is because of the similar structure of the shared data. For an example, both organization 1 and organization 2 choose the hyper parameter as $k = 4$ to anonymize their data, such as in Table 2.2 and Table 2.3. If the attacker wants to know about Bob whose zip code is 917**, probability of Bob's sensitive attribute to be **Statistics** is $\frac{1}{2}$. Then, when put together probability of bob's sensitive attribute as **Statistics** increases from $\frac{1}{2}$ to $\frac{3}{4}$. In this case, the probability of successful attack will be higher.

Let $p(H_s)$ be the success probability of the homogeneity attack when the organizations use the same anonymization level. Similarly, let $p(H_d)$ be the success probability of the homogeneity attack when the organizations use different anonymization levels, such that $p(H_s) > p(H_d)$. We assume $p(B) > p(H_d) > 0$, i.e., the success probability of background attack is higher than that of the homogeneity attack with different anonymization levels, that is because the attacker can link between the shared data and have extra information (background knowledge). However, $p(B)$ can be higher or lower than $p(H_s)$.

The cost of performing the background knowledge attack is assumed to be higher than that of the homogeneity attack, i.e., $c_a(B) > c_a(H) > 0$. This is because the attacker will spend more time collecting the background information and linking the similar information. Note that, when the attacker chooses not to attack, its utility $u_a(N, k_1, k_2)$ will equal zero. This choice will be superior to the attacker if the cost of performing the attack exceeds the reward from revealing the information.

After considering the success probabilities of the different attack types, we reconsider the organizations' utilities in (3.9). We notice that each organization can obtain

a fraction of the reward R_i that depends on the attack's success probability. Let

$$\delta = \frac{1}{k_i} \cdot (1 - \frac{p_{\max}(a)}{\alpha k_i + 1}) \cdot (1 - \frac{p_{\max}(a)}{\alpha k_{-i} + 1}) \tag{3.13}$$

be the minimum fraction of R_i that an organization can achieve based on the maximum success probability of the available attacks, i.e., $p_{\max}(a)$. We refer to δR_i as the minimum profit factor.

To this end, we define a game $\mathcal{G}$= $\{\mathcal{N}, \mathcal{D}, \mathcal{A}, \mathcal{U}\}$ such that $\mathcal{N}$ is the set of the players which include all the organizations as well as the attacker, $\mathcal{D}$ is the set of defender's strategy, $\mathcal{A}$ is the set of attacker's strategy, and $\mathcal{U}$ is the set of the all players' utilities. The goal of each player is to take actions to maximize their utility given the actions of other players. When no player can improve its utility by unilaterally changing its actions, the game is said to be at equilibrium. The notion of equilibrium, in game theory, is referred to as Nash equilibrium [27]. Nash equilibrium can either be pure Nash equilibrium, or mixed-strategy Nash equilibrium. A pure strategy equilibrium is when every player has only one action/ strategy at equilibrium. On the other hand, a mixed Nash equilibrium represents a probability distribution over each player's set of available actions [28]. Next, we study the possible cases of equilibrium, both pure and mixed strategies for the proposed game.

Chapter 4

Proposed Game Solution

The studied game is a finite static non-zero-sum game which is known to have a Nash equilibrium, either pure or mixed-strategy. For the sake of analytical tractability, we consider the case where each organization can choose between two k values, i.e., k_L and k_H. These values represent choosing low and high values for k, respectively. Based on these values, each organization will have two minimum profit factors δ_H and δ_L corresponding to the choice of k_L and k_H, respectively.

Let p_1 be the probability for the first organization to choose k_L such that it chooses k_H with a probability $1-p_1$. Similarly, the second organization can choose k_L and k_H with the probabilities p_2 and $1-p_2$, respectively. The attacker, on the other hand, will have a probability distribution of q_B, q_H, q_N for choosing the actions B, H, and N, respectively. We start the analysis by considering the cases in which the game $\mathcal{G}$ can have a pure strategy Nash equilibrium.

Proposition 1. *Let* $k_i^* = \arg\max_{k_i} \frac{R_i}{k_i} - c_i(k_i) + \gamma \cdot k_i$. *Then, the tuple* (k_i^*, k_{-i}^*, N) *constitute a pure strategy Nash equilibrium for* $\mathcal{G}$ *when the attacker is not able to achieve a positive utility.*

Proof. We note that, the attacker's utility for no-attack is zero, i.e., $u_a(k_1, k_2, N) = 0$. The attacker can only turn to this choice if all the other actions yield a negative utility, i.e., all the utility instances for choosing B and N with the different combinations of k_L and k_H, for each organization, will be result in a negative attacker's utility. Therefore, choosing the action N will be a dominant strategy for the attacker. In this case, each organization's utility will be:

$$u_i(k_i, k_{-i}, N) = \frac{R_i}{k_i} - c_i(k_i) + \gamma \cdot k_i, \tag{4.1}$$

which clearly depends only on the organization's action and not on the other players' actions. In this case, each organization will chose the value of k that maximizes its utility in (4.1). Hence, $k_i^* = \arg\max_{k_i} \frac{R_i}{k_i} - c_i(k_i) + \gamma \cdot k_i$ will represent the optimal organization's choice under no-attack scenario. In this case, no player will have an incentive to change its choice and, therefore, the actions tuple (k_i^*, k_{-i}^*, N) is a pure strategy Nash equilibrium for the game. □

From Proposition 1, the attacker's probability q_N of choosing the action N will be either 1 or 0 based on whether the action N dominates or it is being dominated by other actions. Therefore, we will consider only two actions for the attacker, i.e., B and H which can be selected by the probabilities q and $1 - q$, respectively. Similarly, for the organizations, we note the similarity in their actions and utilities, thus, they will have the same equilibrium profile which can be given by p for selecting k_L and $1 - p$ for selecting k_H.

Similar to the attacker, each organization can have a dominant strategy under some circumstances and, hence, the probability p can be either 0 or 1 based on the

dominant strategy.

Proposition 2. *Each organization will have a dominant strategy when the value of R_i is large enough such that the minimum profit factor is the dominant term in the organization's utility, i.e., $\delta_H R_i > \gamma \cdot k_H - c_i(k_H)$ and $\delta_L R_i > \gamma \cdot k_L - c_i(k_L)$. The dominant strategy can then be given as the solution of:*

$$k_i^* = \arg\max_i \delta_i R_i - c_i(k_i) + \gamma \cdot k_i, \quad i \in \{L, H\}$$

Proof. The values of $\delta_H R_i$ and $\delta_L R_i$ represent the minimum fractions of the reward each organization can achieve, under the attacker's maximum probability of success. When the values of R_i are large enough to make these minimum profit factors higher than the rest of the utilities, each organization can expect that any other attacker's action will not lower its utility. Thus, the organization can determine its dominant strategy while neglecting the attacker's effect. □

Note in Proposition 2, a high reward can eliminate the attacker's effect, however, it cannot be used solely to determine the organization's action as this is affected by the other factors in the organization's utility.

To this end, when no player has a dominant strategy, the players will randomize over their strategies using the probability distributions of the mixed-strategy Nash equilibrium. These mixed strategies can be calculated when the players are indifferent between choosing their actions, i.e., the expected utility of choosing each action will be the same. For instance, the organizations can choose their p such that the attacker's expected utility from choosing the action B will equal to that of choosing the action

H. The attacker's expected utility from choosing the action B can be given by:

$$\begin{aligned}\mathbb{E}(u_a(k_1,k_2,B)) = p\cdot p\cdot u_a(k_L,k_L,B) + p\cdot(1-p)\cdot u_a(k_L,k_H,B) + (1-p)\\ \cdot p\cdot u_a(k_H,k_L,B) + (1-p)\cdot(1-p)\cdot u_a(k_H,k_H,B).\end{aligned} \tag{4.2}$$

Similarly, the expected utility of choosing the action H can be given by:

$$\begin{aligned}\mathbb{E}(u_a(k_1,k_2,H)) = p\cdot p\cdot u_a(k_L,k_L,H) + p\cdot(1-p)\cdot u_a(k_L,k_H,H) + (1-p)\\ \cdot p\cdot u_a(k_H,k_L,H) + (1-p)\cdot(1-p)\cdot u_a(k_H,k_H,H).\end{aligned} \tag{4.3}$$

For the attacker to be indifferent between its actions, the utility in (4.2) must equal the utility in (4.3). Solving both equations together, the organizations' probabilities of choosing k_L, i.e., p can then be given as the solution of the equation:

$$\begin{aligned}&\Big(\frac{2p^2(B)-2p^2(H)}{(\alpha k_L+1)(\alpha k_H+1)} - \frac{p^2(B)-p^2(H)}{(\alpha k_L+1)^2} - \frac{p^2(B)-p^2(H)}{(\alpha k_H+1)^2}\Big)R_a p^2\\ &+\Big(\frac{p(B)-p(H)}{(\alpha k_L+1)} - \frac{p^2(B)-p^2(H)}{(\alpha k_L+1)(\alpha k_H+1)} + \frac{p^2(B)-p^2(H)}{(\alpha k_H+1)^2}\\ &-\frac{p(B)-p(H)}{(\alpha k_H+1)}\Big)2R_a p + \Big(\frac{2p(B)-2p(H)}{(\alpha k_H+1)} - \frac{p^2(B)-p^2(H)}{(\alpha k_H+1)^2}\Big)R_a\\ &- c_a(B) + c_a(H) = 0.\end{aligned} \tag{4.4}$$

After calculating the probability p, the attacker's probability q can be calculated in a similar way by considering the expected utility of one of the organizations. Note that, due to the symmetry between the organizations, considering the utilities of both organizations will be redundant. To this end, the first organization expected utility from choosing k_L can be given by:

$$\begin{aligned}\mathbb{E}(u_1(k_L,k_2,a)) = p\cdot q\cdot u_1(k_L,k_L,B) + p\cdot(1-q)\cdot u_1(k_L,k_L,H) + (1-p)\\ \cdot q\cdot u_1(k_L,k_H,B) + (1-p)\cdot(1-q)\cdot u_1(k_L,k_H,H).\end{aligned} \tag{4.5}$$

Similarly, the first organization's expected utility from choosing k_H can be given by:

$$\begin{aligned}\mathbb{E}(u_1(k_H, k_2, a)) = p \cdot q \cdot u_1(k_H, k_L, B) + p \cdot (1-q) \cdot u_1(k_H, k_L, H) + (1-p) \\ \cdot q \cdot u_1(k_H, k_H, B) + (1-p) \cdot (1-q) \cdot u_1(k_H, k_H, H).\end{aligned} \tag{4.6}$$

For the organization to be indifferent between its actions, the utility in (4.5) must equal the utility in (4.6). Solving both equations together, the attacker's probabilities of choosing B, i.e.,q can then be given as the solution of the equation:

$$\begin{aligned}q = \Big(u_1(k_L, k_H, H) + u_1(k_H, k_H, H) - p\Big(u_1(k_L, k_L, H) \\ + u_1(k_H, k_L, H) + u_1(k_L, k_H, H) + u_1(k_H, k_H, H)\Big)\Big) \Big/ \\ \Big(u_1(k_L, k_H, H) + u_1(k_H, k_H, H) - u_1(k_H, k_H, B) - u_1(k_H, \\ k_L, B) + p\Big(u_1(k_H, k_H, B) + u_1(k_H, k_H, B) + u_1(k_L, k_L, B) + \\ u_1(k_H, k_L, B) - u_1(k_H, k_H, H) - u_1(k_L, k_H, H) - u_1(k_L, \\ k_L, H) - u_1(k_H, k_L, H)\Big)\Big)\end{aligned} \tag{4.7}$$

Given the value of p, the value of q can be uniquely computed from (4.7). The Nash equilibrium mixed strategies can then be given as $(p, 1-p)$ for the organizations and $(q, 1-q)$ for the attacker. Next, we discuss how data collector (Cybex) interacts with the organizations to establish a connection with them.

Chapter 5

Optimal Contracts

The goal of the data collector is to collect the data from organizations and make profits by performing data mining on it and the goal of the organizations is to get maximum utility by sharing data to the data collector while preserving the privacy. We propose a contract theoretic framework to decide the rewards for the organizations based on the utility of the data collector (Cybex). The utility function for data collector (cybex) can be written as:

$$U_d(k_i, k\_i) = \sum_{i=1}^{\mathcal{N}} p_i \cdot (\mathcal{V}_i - r_i(k_i)). \tag{5.1}$$

Here, $\mathcal{V}_i$ is the evaluated function of profit earned by the data collector from the data collected from i^{th} organization, where $i = 1, 2..., \mathcal{N}$. p_i is the mixed strategy equilibrium of an organization choosing it's strategy as k_i. The contract on which both data collector and organizations work on should guarantee an incentive to their

utility and should satisfy the two key constraints, namely, Individual Rationality (IR) and incentive Compatibility (IC).

1. **Individual Rationality (IR):** As both organizations and the data collector are rational, they would agree on a contract if and only if it ensures non-negative utility.

$$r_i(k_i) \cdot \delta - c_i(k_i) + T(k_i) > 0 \tag{5.2}$$

Therefore, the minimum reward for organization to accept the contract is as follows:

$$r_{min}(k_i) \cdot = \frac{c_i(k_i) - T(k_i)}{\delta} \tag{5.3}$$

2. **Incentive Compatibility (IC):**The contract should also ensure that the organizations can achieve maximum utility only if they agree on a particular contract considering the following two constraints:

According to equation, 3.2 and 3.6, if $k_1 < k_2$

$$r_i(k_1) - r_i(k_2) > 0 \tag{5.4}$$

$$T(k_2) - T(k_1) > 0 \tag{5.5}$$

In order to formulate optimal contracts, the data collector should come up with contracts with maximum rewards which are given to the organizations. Thus, the

optimization problem would be:

$$\begin{aligned} \max_{r(k)} \quad & \sum_{i=1}^{\mathcal{N}} p_i \cdot (\mathcal{V}_i - r_i(k_i)) \\ \text{s.t} \quad & r_i(k_i) \cdot \delta - c_i(k_i) + T(k_i) > 0, \\ & T(k_j) > T(k_i), \\ & r_i(k_i) > r_j(k_j), . \end{aligned}$$

Using Lagrange's analysis along with KTT conditions, we can solve the optimization problem subjected to 3 inequality constraints, 5.2, 5.4, 5.5, in order to find the optimal contract.

$$\begin{aligned} L\Big(r(k), \mu\Big) = & \sum_{i=1}^{\mathcal{N}} p_i \cdot (\mathcal{V}_i - r_i(k_i)) + \\ & \sum_{i=1}^{\mathcal{N}} \mu_i \cdot \Big(r_i(k_i) \cdot \delta - c_i(k_i) + T(k_i)\Big) + \mu_{(\mathcal{N}+1)} \cdot \Big(r_1(k_1) \\ & - r_2(k_2)\Big) + \mu_{(\mathcal{N}+2)} \cdot (T(k_2) - T(k_1)) \end{aligned} \tag{5.6}$$

In this work, we have two organizations involved in a network along with the data collector (cybex). So, the above equation 5.6 can be re-written as:

$$\begin{aligned} L\Big(r(k), \mu\Big) = & p_1 \cdot \Big(\mathcal{V}_1 - r_1(k_1)\Big) + p_2 \cdot \Big(\mathcal{V}_2 - r_2(k_2)\Big) \\ & + \mu_1 \cdot \Big(r_1(k_1) \cdot \delta - c_1(k_1) + T(k_1)\Big) \end{aligned}$$

$$+ \mu_2 \cdot \Big(r_2(k_2) \cdot \delta - c_2(k_2) + T(k_2) \Big)$$
$$+ \mu_3 \cdot \Big(r_1(k_1) - r_2(k_2) \Big) + \mu_4 \cdot (T(k_2) - T(k_1)) \tag{5.7}$$

This optimization problem is similar to optimal contract problem in [14] [15]. The solution is feasible if and only if the following condition is satisfied, $r_i(k_i) > r_{min}(k_i)$. Next, we discuss how the data collector updates the contracts to improve his utility.

Chapter 6

Repeated Game

In Chapter 4 and Chapter 5, we have showed how this game is solved in a static game. In this chapter, we are going to discuss how this game is played in a multi-stage game or repeated game. To this end, we can say that the data collector's utility in Stage-1 is 0, i.e, he is going to give all his income earned from performing data mining to the organizations in order to establish a network. In the further stages, the data collector gradually increases his utility by reducing the rewards which should also satisfy the IR and IC constraints. Therefore,

$$U_d(k_i, k\_i) = \sum_{i=1}^{\mathcal{N}} p_i \cdot \Big(\mathcal{V}_i - (1-d)^{t-1}(r_i^{(t-1)}(k_i))\Big); \tag{6.1}$$

Where, d is the rate of reduction by which the data collector decreases the reward of the organization at stage, $t = 1, 2, 3, \ldots$

and $r_i^{(t)}(k_i)$ is the reward of the organization according to the contract at stage t. At

stage one, i.e, when $t = 1$, the data collector starts with zero utility. Therefore, as $(1 - d)^0 = 1$, we can say that,

$$r_{max}^{(0)}(k_i) = \mathcal{V}_i \tag{6.2}$$

To this end, $r_{max}^{(0)}(k_i)$ is the maximum reward earned by the organization for sharing the data. From stage two, the data collector reduces the reward of the organization by d, where $0 < d < 1$ and also designs a new contract which also satisfies the IR constraint, such that, $r_i(k_i) > r_{min}(k_i)$.

In Algorithm 1, we have provided a heuristic on how this game is played in a repeated game scenario. At the first stage, the data collector designs the contracts using $r_{max}^{(0)}(k_i)$, if the contracts fails to satisfy the IR and IC constraints then network cannot be established or else the contracts are offered to the organizations. From the second stage, the data collector re-designs the contracts with updated rewards which are obtained by solving equation 6.1. If the updated reward is less than $r_{min}(k_i)$ then the data collector uses the contracts which are designed in the previous stage as the new contracts fails to satisfy the IR constraint or else offers the contracts with updated rewards to the organizations.

The following is the heuristic on how the data collector updates the reward of the organization in a multi- stage game :

Algorithm 1: Optimized repeated game algorithm

Input: $r_{min}(k_i)$, $\mathcal{V}_i$, d.
1.Data collector asks the organizations to share the data.
2.Organizations, before sharing perform anonymization in order to preserve the privacy of the data holders.
3.**for** `t = 1 to n` **do**
  if $t = 1$ **then**
    Solve for optimal contracts
    if *there exists a feasible solution, i.e,* $r_i^{(1)}(k_i) > r_{min}(k_i)$ **then**
      Contracts are offered to the organization
    else
      Network cannot be established
    end
  else
    re-design the contracts using equation 6.1
    if $r_i^{(t)}(k_i) > r_{min}(k_i)$ **then**
      Offer the new contract to the organization
    else
      Use previous contract, i.e, $r_i^{(t-1)}(k_i)$
    end
  end
end

Chapter 7

Simulation Results and Analysis

For our simulations, we set the value of lower k, $k_L = 4$ and the other value is chosen to be slightly higher, i.e., higher k, $k_H = 7$ to better highlight the effect of these values on the utilities, and are kept same for all the experiments. Other system parameters, such as measure of information security, $\alpha = 1$, co-efficient of trust, $\gamma = 0.6$, and conversion factor which converts from time complexity to monetary value is set to, $\beta = 10^{-6}$. We assume similar dataset structures between the organizations, so that, the cost function is affected only by the choice of k. In Fig. 7.4, 7.5, each stage represents all possible interactions between the players and it outcomes for 10 stages.

First, we solve the formulated game $\mathcal{G}$ using the analysis in chapter 4. We allow the values of attacker's reward R_a to vary from 16 to 25. These values represent the monetary rewards the data collector will give to the organizations as a reward for sharing the data. Here, we use abstract values, however, in a real-life scenario, the data collector needs to estimate these values to be proportional to the cost. The

Table 7.1: Attacker's equilibrium strategies

$\mathbf{R_a}$	16	17	18	19	20	21	22	23	24	25
B	0	0	0.0971	0.4290	0.7175	0.97	1	1	1	1
H	0	0	0.9029	0.5710	0.2825	0.03	0	0	0	0
N	1	1	0	0	0	0	0	0	0	0

Table 7.2: Defender's equilibrium strategies

$\mathbf{r_i(k_i)}$	16	17	18	19	20	21	22	23	24	25
$\mathbf{k_L}$	0	0	0.2187	0.2125	0.2070	0.2019	0.1973	0.1937	0.1893	0.1857
$\mathbf{k_H}$	1	1	0.7813	0.7875	0.7930	0.7981	0.8027	0.8063	0.8107	0.8143

equilibrium strategies for both the attacker and defender are shown in Tables 7.1 and 7.2, respectively. We note that, when the values of R_a are less than 17, the attacker cannot achieve a positive utility, and, hence, it will choose not to attack. This situation corresponds to the case of Proposition 1 and the defender's utility is calculated using (4.1). In this case, the defender will have a pure strategy of choosing k_H. For the values of R_a and $r_i(k_i)$ between 18 and 21, both the attacker and the defender will have mixed strategies, i.e., choosing their actions with certain probabilities. Finally, for large values of R_a and $r_i(k_i)$, the attacker will benefit if it performs the background knowledge attack, in this case the defender can choose between the two values of k with k_H being superior.

Fig. 7.1 shows the expected utilities calculated using the equilibrium strategies in Tables 7.1 and 7.2. These utilities are compared to the case where one player chooses random probabilities while the other plays its equilibrium strategy. From Fig. 7.1, we can see that when a player deviates from the equilibrium strategy to a random strategy, it cannot increase its utility as it will be lower or equal to the equilibrium utility. This corroborates the importance of finding Nash equilibrium strategies as

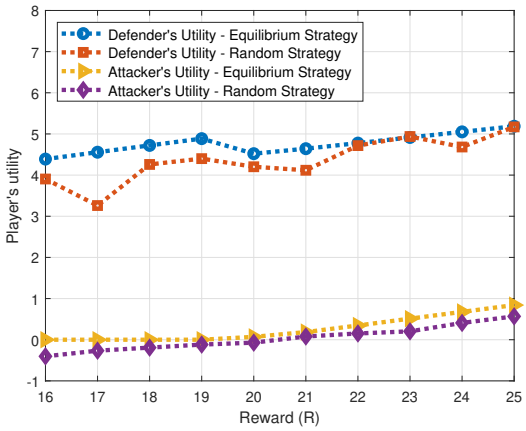

Figure 7.1: The defender's and the attacker's utilities at equilibrium at different reward R values.

they represent the best each player can do given their opponent's actions. We can also see from Fig. 7.1 that the players' utilities do not exhibit a monotonic increase in rewards of attacker and organization as at each value of R_a and $r_i(k_i)$ the utility depends on the players' strategies.

In Fig. 7.2, we show the effect of the success probability of the background knowledge attack, i.e., $p(B)$ on the equilibrium strategies of the players. Note that the values of $p(B)$ are chosen to start at 0.4 to satisfy the assumption $p(B) > p(H_d)$. The equilibrium strategies in Fig. 7.2 are calculated in a similar way to the values in Tables 7.1 and 7.2. The simulation parameters are the same as Fig. 7.1. This value was chosen as the attacker has almost equal probability of choosing its actions under this value. From Fig. 7.2, we can see that when $p(B)$ is slightly higher than $p(H_d)$ i.e., $p(B) = 0.4$ the attacker will have a high value of q which corresponds to the probability of choosing the background knowledge attack. At the same point, the defender will be choosing k_H with slightly higher probability. However, as the value

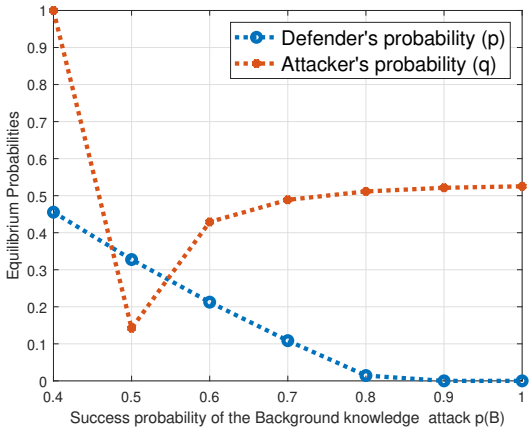

Figure 7.2: The defender's and the attacker's equilibrium probabilities at different success probabilities for background knowledge attack $p(B)$ values.

of $p(B)$ increases, the defender will prefer to use k_H more which lowers the attacker's utility and forces it to switch to the homogeneity attack because of its lower cost. This can be seen as the value of q decreases for $p(B) = 0.5$ and this represents the maximum probability of choosing the homogeneity attack. As $p(B)$ increases more, it will become closer to $p(H_s)$ and in this case, the defender will stick more to choosing k_H as it achieves more trust in protecting the data. Meanwhile, the attacker will choose the background knowledge attack with slightly higher probability.

In Fig. 7.3, we study the effect of the success probability of the homogeneity attack, at similar values of k, i.e, $p(H_s)$ on the equilibrium strategies of the players. Similar to Fig. 7.2, the values of $p(H_s)$ are starting at 0.4 so that $p(H_s) > p(H_d)$. The simulation parameters are the same as Fig. 7.2 and $p(B) = 0.5$. From Fig. 7.3, we can see that when $p(H_s)$ is less than $p(B)$ i.e., $p(H_s) < 0.6$, so the attacker will have a higher probability of choosing the background knowledge attack. This probability will decrease as $p(H_s)$ is equal to $p(B)$ or higher. In this case, the attacker will prefer

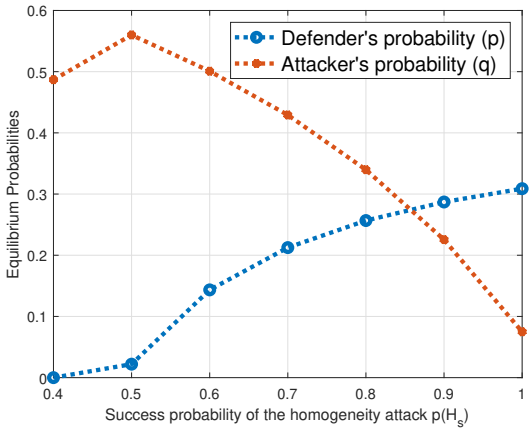

Figure 7.3: The defender's and the attacker's equilibrium probabilities at different success probabilities for homogeneity attack $p(H_s)$ values.

to choose the homogeneity attack with higher probability especially with the increase in its success probability. For the same range of probabilities, the defender will be choosing k_H with higher probability. However, this probability will be decreasing as $p(H_s)$ increases.

In Fig. 7.4, we study the effect of rate of reduction on the reward of the organizations over the stages, i.e., effect of d on the organization's reward. We assumed the values of d to be 0.2, 0.3, 0.4, as they are supposed to be $0 < d < 1$. From equation 5.3, we get the values of $r_{min}(k_L)$ and $r_{min}(k_H)$, which are 4.42, 3.56 , respectively. Now, using the algorithm: 1, we have calculated the rewards of organizations and plotted them in Fig. 7.4. Here, we can observe that at stage-1, the organizations are offered contracts with maximum rewards. Later, the reward is gradually decreased and after a few stages there is no change in rewards as $r_i^{(t)}(k_i) < r_{min}(k_i)$ and the organizations will not accept the contracts if it doesn't satisfy the IR Constraints. Therefore, the data collector (cybex) uses the previous contract.

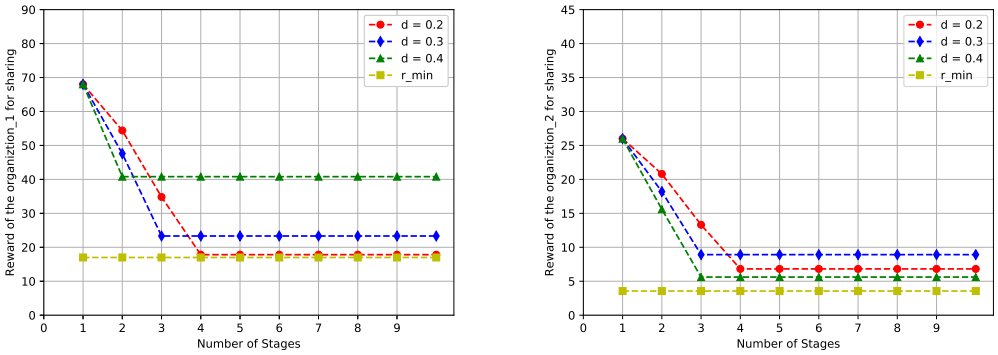

Figure 7.4: Reward of organization 1 and organization 2 for sharing the data over the stages.

In Fig. 7.5, we have plotted the utility of data collector (cybex) over different stages. Here, the value of equilibrium strategy p is calculated in a similar way to the values in Tables: 7.1 and 7.2 and used to calculate the utility of data collector using equation 6.1 . We can observe that the utility of data collector (Cybex) is 0 at stage-1, as it is distributing the most of it's income to the organization and later, increases it's utility by decreasing the reward of the organizations.

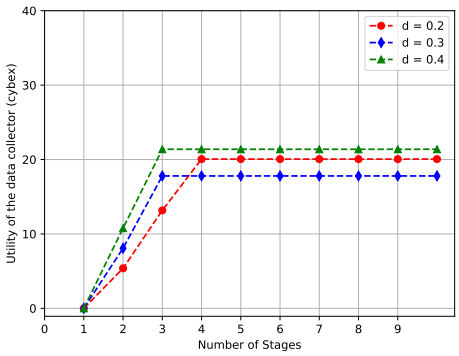

Figure 7.5: Utility of the data collector(cybex) over the stages.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

In this work, we have studied the problem of determining the optimal value of k for the $k-$anonymization technique. We have formulated the problem using a game-theoretic model that involves three players which are an attacker and two organizations sharing data with a common platform, a data collector. In particular, we have considered two common types of attacks that can affect k-anonymization techniques. We have defined the players' utilities resulting from the interactions between the three players. Then, we have provided the mathematical model for deriving the different Nash equilibria for the proposed game.

From the game-theoretic model, we get the minimum reward for the organizations. We have showed how the data collector(Cybex) builds the contracts with rewards which are greater than minimum reward in a static game scenario. We have

also provided a heuristic on how the data collector establishes a network along with the organizations and updates the reward to increase it's utility in a repeated game scenario.

8.2 Future Work

In the future, this model can be extended by implementing upon l-diversity and t-closeness which are extended versions of k- anonymization technique. For better analysis, we can use huge datasets which follow different trends and learn how these anonymization techniques can be compromised in order to provide better security. Although, we have considered the trust factor to be constant, future work would be required on how the trust factor changes over the stages.

Bibliography

[1] G. Zyskind, O. Nathan *et al.*, "Decentralizing privacy: Using blockchain to protect personal data," in *2015 IEEE Security and Privacy Workshops*. IEEE, 2015, pp. 180–184.

[2] S. Badsha, I. Vakilinia, and S. Sengupta, "Privacy preserving cyber threat information sharing and learning for cyber defense," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2019, pp. 0708–0714.

[3] A. Eldosouky and W. Saad, "On the cybersecurity of m-health iot systems with led bitslice implementation," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2018, pp. 1–6.

[4] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[5] L. SWEENEY, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and*

Knowledge-Based Systems, vol. 10, no. 05, pp. 571–588, 2002. [Online]. Available: https://doi.org/10.1142/S021848850200165X

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 2006, pp. 24–24.

[7] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[8] T. Li, N. Li, and J. Zhang, "Modeling and integrating background knowledge in data anonymization," in *2009 IEEE 25th International Conference on Data Engineering*. IEEE, 2009, pp. 6–17.

[9] Q. Wang, Z. Xu, and S. Qu, "An enhanced k-anonymity model against homogeneity attack." *JSW*, vol. 6, no. 10, pp. 1945–1952, 2011.

[10] Z. Liang and R. Wei, "Efficient k-anonymization for privacy preservation," in *2008 12th International Conference on Computer Supported Cooperative Work in Design*. IEEE, 2008, pp. 737–742.

[11] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game theory in wireless and communication networks: theory, models, and applications*. Cambridge university press, 2012.

[12] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Bacundefinedar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Comput. Surv.*, vol. 45, no. 3, Jul. 2013. [Online]. Available: https://doi.org/10.1145/2480741.2480742

[13] C. T. Do, N. H. Tran, C. Hong, C. A. Kamhoua, K. A. Kwiat, E. Blasch, S. Ren, N. Pissinou, and S. S. Iyengar, "Game theory for cyber security and privacy," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–37, 2017.

[14] A. Eldosouky, W. Saad, C. Kamhoua, and K. Kwiat, "Contract-theoretic resource allocation for critical infrastructure protection," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[15] P. Bolton, M. Dewatripont *et al.*, *Contract theory*. MIT press, 2005.

[16] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *21st International conference on data engineering (ICDE'05)*. IEEE, 2005, pp. 217–228.

[17] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," in *International Conference on Database Systems for Advanced Applications*. Springer, 2007, pp. 188–200.

[18] E. K. Wang, B. Jia, and N. Ke, "Modeling background knowledge for privacy preserving medical data publishing," in *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*. IEEE, 2017, pp. 136–141.

[19] R. B. Myerson, *Game theory*. Harvard university press, 2013.

[20] M. J. Osborne and A. Rubinstein, *A course in game theory.* MIT press, 1994.

[21] S. Badsha, I. Vakilinia, and S. Sengupta, "Privacy preserving cyber threat information sharing and learning for cyber defense," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 2019, pp. 0708–0714.

[22] I. Vakilinia, D. K. Tosh, and S. Sengupta, "Privacy-preserving cybersecurity information exchange mechanism," in *2017 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, 2017, pp. 1–7.

[23] R. Bredereck, A. Nichterlein, R. Niedermeier, and G. Philip, "The effect of homogeneity on the computational complexity of combinatorial data anonymization," vol. 28, no. 1, Jan 2014, pp. 65–91. [Online]. Available: https://doi.org/10.1007/s10618-012-0293-7

[24] L. A. Gordon and M. P. Loeb, "The economics of information security investment," *ACM Trans. Inf. Syst. Secur.*, vol. 5, no. 4, pp. 438–457, Nov. 2002. [Online]. Available: http://doi.acm.org/10.1145/581271.581274

[25] A. S. Sattar, J. Li, J. Liu, R. Heatherly, and B. Malin, "A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments," *Knowledge-based systems*, vol. 67, pp. 361–372, 2014.

[26] H. Ogut, N. Menon, and S. Raghunathan, "Cyber insurance and it security investment: Impact of interdependence risk." in *WEIS*, 2005.

[27] S. Sengupta, M. Chatterjee, and K. Kwiat, "A game theoretic framework for power control in wireless sensor networks," *IEEE Transactions on Computers*, vol. 59, no. 2, pp. 231–242, 2009.

[28] A. Eldosouky, W. Saad, and D. Niyato, "Single controller stochastic games for optimized moving target defense," in *2016 IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.