2013-04-16

Stable Privacy Parameter Settings Using Game Theory

Karimi Adl, Rosa

UNIVERSITY OF CALGARY

Stable Privacy Parameter Settings

Using Game Theory

by

Rosa Karimi Adl

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

CALGARY, ALBERTA

April, 2013

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a dissertation entitled "Stable Privacy Parameter Settings Using Game Theory" submitted by Rosa Karimi Adl in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Supervisor, Dr. Ken Barker
Department of Computer Science

Dr. Jörg Denzinger
Department of Computer Science

Dr. Payman Mohassel
Department of Computer Science

Dr. Jonathan Sillito
Department of Computer Science

Internal/External Examiner,
Dr. Mark Bauer
Department of Math Sciences

External Examiner,
Dr. Raymond T. Ng
University of British Columbia

Date

Abstract

Privacy protection appears as a fundamental concern when personal data is collected, stored, and published. Several privacy protection methods have been proposed to address privacy issues in private datasets. Each method has at least one parameter to adjust the guaranteed level of privacy protection. As the privacy protection level increases, the dataset loses more information utility due to further application of data manipulation methods and/or access restriction rules. Consequently, balancing the tradeoff between privacy and utility is a crucial step and so far no systematic mechanism exists to provide directions on how to establish values for privacy parameters such that a balanced privacy/utility tradeoff is induced.

A balanced privacy/utility tradeoff can be described as a level on which the stakeholders of data reach a consensus (in the sense that no single party would be wiling to act differently to change the agreed upon level). Game theory provides a natural solution to finding such balanced tradeoffs. In this thesis, we capture the essence of establishing balancing values for privacy parameters as an extensive-form game with incomplete and imperfect information. A high-level step-by-step guideline is provided on how to solve the generic game. We instantiate the generic game model for three different privacy protection methods and analytically solve each game. The games' solutions are further simulated for sample problem settings to study the effects of various problem parameters on the balancing values of privacy parameters.

The game model and its solution contribute to the fulfilment of our objective of establishing balancing values for privacy parameters (of a chosen privacy protection method). In addition to our main objective, the proposed game model can be consulted to choose the most profitable privacy protection method based on the problem requirements. Benchmarking frameworks can also benefit from our game solutions by using the balancing privacy parameter values as the reference points for the comparisons between different privacy protection methods. We believe that a first step towards improving the data collection and

privacy protection procedures is to understand how much privacy is currently sacrificed to achieve information utility (at the steady states). The game-based solution provided in this thesis promotes a deeper understanding of how privacy and utility reach a balanced trade-off within the current privacy protection methods.

Acknowledgements

The process of completing a PhD program entails many challenges far and beyond the extensive study, research, and problem solving. It is, in fact, an overwhelming and lonely affair. Pushing through the ups and downs and the occasional disappointments was not an easy task and I am certain that if it had not been for the invaluable assistance of my supervisor, committee, and many other individuals, this dissertation would not have been possible.

First and foremost I offer my sincerest gratitude to my supervisor Dr. Ken Barker who has patiently provided me with continuous technical, emotional, and financial support during the past five years. By giving me the honour of starting a PhD degree under his supervision, he opened up a world of opportunities for me to pursue my dream of being an independent researcher. Besides his abundant guidance in technical aspects of my research, he taught me how to explore open problems in realms of privacy and he helped me choose between alternative methodologies to address the research problem. He has been my inspiration as I hurdled all the obstacles in the completion of this dissertation and one simply could not wish for a better or friendlier supervisor.

I would also like to express my special appreciation and thanks to Dr. Denzinger for meticulous assessment of this dissertation and valuable advice on how to improve the quality of this work. From the early stages all the way to the last days of writing the dissertation, he has been promoting constructive discussions to add to the depth and applicability of this project.

It gives me great pleasure in acknowledging the support and help of the rest of my supervisory committee and examiners Dr. Payman Mohassel, Dr. Jonathan Sillito, Dr. Mark Bauer, and Dr. Raymond T. Ng who pointed out critical issues to be addressed in the thesis and provided me with different perspectives to further progress this work in the future.

I am also indebted to my Master's degree supervisor Mr. Rouhani Rankouhi from whom I learned the essential characteristics of being a successful researcher and teacher. He encouraged me to learn philosophy of science and simply taught me how to live the academic dimension of my life. Words cannot express my gratitude for the lessons I have learned from him.

I share the credit of this work with my colleagues, especially Dr. Mina Askari, Dr. Mishtu Banerjee, Leanne Wu, and other PSEC group members who participated in a collaborative environment to share and develop ideas.

I am thankful to the Department of Computer Science for providing the support and equipment I needed to produce and complete my thesis and to the AITF for partially funding my studies.

I am grateful to my friends Dr. Yaghoubi, Dr. Haghighat, Dr. Barzin, Emmet Moradi, Dr. Epp, and Thomas Burt for their endless support. During the past five years they have always been there for me and held my hands through all the sad and happy moments of my life.

Last but not least, I wish to express my love and gratitude to Cindy and my supportive family; for their understanding and encouragement through the duration of my studies. They believed in me even when I was doubting myself and this thesis would have remained a dream had it not been for their kindness and support.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
DC	A data collector, page 77
DP	A data provider, page 77
DU	A data user, page 77
pur	The purpose of data collection and usage, page 88
IDES	Iterated Elimination of Dominated Strategies, page 59
α_o	Probability of the event that a data provider opts in for a privacy policy according to offer o with zero incentive, page 111
β_0	The constant term in data providers' OptIn probability function, page 94
β_1	The coefficient of $h(\delta)$ in data providers' OptIn probability function, page 94
β_2	The coefficient of $g(I)$ in data providers' OptIn probability function, page 94
ΔQ	The sensitivity of query Q, page 24
δ	The privacy parameter represented as a vector of privacy components, page 75
ϵ	The privacy parameter in differential privacy, page 24
ϵ	The privacy parameter in differential privacy, page 173
γ	The coefficient of $g(I)$ in data providers' OptIn probability function, page 111
$\hat{I}$	The incentive that maximizes data collector's payoff for each offer, page 96
$\hat{U}_{DC}^{\langle case \rangle}$	The maximum payoff to the data collector if she accepts an offer in case $\langle case \rangle$, page 114

$\hat{U}_{DU}^{\langle case \rangle}$	The data user's maximum payoff by making an *acceptable* offer in case $\langle case \rangle$, page 119
κ	The randomized function to provide differential privacy, page 24
λ_i	The sensitivity of attribute A_i (or the retention component), page 107
$\langle \hat{\delta}, \hat{p} \rangle$	The data user's most profitable offer, page 98
$\overline{v}$	The upper bound of the function $w_1 \cdot h(\delta) + w_2 \cdot g(I)$, page 79
$\overline{v}'$	The upper bound of the function $h(\delta)$, page 81
ρ	The total number of queries that the data user intends to ask, page 177
σ	The depth of the recursive calls in the Mondrian Algorithm, page 161
τ_0	The constant term in data providers' OptIn probability function, page 111
τ_i	The coefficient of $h_g(g_i)$ (privacy gain from the granularity level of attribute A_i) in data providers' OptIn probability function, page 111
θ	The coefficient of $h_r(r)$ (privacy gain from the retention value) in data providers' OptIn probability function, page 111
$\|T\|$	Cardinality of a data table (or query result set) T, page 25
ξ	Accuracy of the `COUNT`-query when attribute A_j is revealed at a partial granularity level, page 108
A	The share of data anonymization cost for each linkable data field, page 109
a	The economic value of a record when values of attribute A_j are revealed at a partial granularity level, page 108
A_i	An attribute (or data field) i, page 12

A_j	The attribute over which the predicate of the data user's query is defined, page 107
AC_i^j	The set of actions available to player i at point j of the game, page 53
B	The base cost of storing the dataset and enforcing the privacy policy, page 109
b	The economic value of each record to the data user, page 83
BR_{DC}	The best response of the data collector, page 96
BR_{DP_i}	The best response of data provider i, page 91
C	The cost of data collection, storage, and privacy protection, page 84
C_o	The cost of data collection, storage and privacy protection according to an offer o (when cost is a function of δ), page 109
$CG(g_i)$	The cost of protecting data field A_i at granularity level g_i, page 109
d	Number of quasi-identifying attributes, page 15
D_i	The domain of the i^{th} quasi-identifying attribute, page 18
DF	The set of all data fields to be collected, page 104
DP_i	The data provider i, page 77
dpv	The vector of data providers' decisions, page 86
EC	An equivalence class, page 12
ec_{AVG}	The average number of records in each equivalence class in k-anonymity, page 161
$EN(N)$	The effective cardinality of a dataset with cardinality N, page 83

$F^f(l)$ Fraction of the privacy fundamentalists with threshold of at most l, page 82

$F^p(l)$ Fraction of the privacy pragmatists with threshold of at most l, page 80

f^f The probability density function of the privacy fundamentalists' thresholds, page 82

f^p The probability density function of the privacy pragmatists' thresholds, page 80

G The cost of applying a generalization method to a data field, page 109

$g(I)$ The perceived rewarding effects of the incentive I, page 79

g_i The granularity level of data field A_i, page 105

Gr The set of all possible granularity levels, page 104

$h(\delta)$ The perceived amount of privacy gain from the privacy parameter δ, page 79

$h_g(g_i)$ The perceived amount of privacy gain from choosing granularity level g_i for attribute A_i, page 106

$h_r(r)$ The perceived amount of privacy gain from choosing the retention period r, page 106

I The amount of incentive in monetary value, page 78

k The privacy parameter in k-anonymity, page 155

$l_i^{f,j}$ The privacy/incentive threshold level of subtype $t_i^{f,j}$, page 81

$l_i^{p,j}$ The privacy/incentive threshold level of subtype $t_i^{p,j}$, page 79

m The maximum number of records with identical values for all quasi-identifying attributes., page 161

Max	The maximum number of data records necessary for data analysis, page 82
Min	The minimum number of data records necessary for data analysis, page 82
N	The expected cardinality of the data table, page 83
n	The total number of data providers, page 77
oe	An offer that only asks for data field A_j at the exact granularity level, page 119
$Of = \langle \delta, p \rangle$	An offer made by the data user, page 85
op	An offer that only asks for data field A_j at the partial granularity level, page 119
p	The price that the data user pays for each record, page 83
p_1	The probability of a data provider being a privacy unconcerned, page 78
p_2	The probability of a data provider being a privacy pragmatist, page 78
p_3	The probability of a data provider being a privacy fundamentalist, page 78
PP	A privacy policy, page 104
Pr_i	The ratio of the records that have value $\mathtt{v}_i$ for the quasi-identifying attribute $\mathtt{q}$ in the dataset., page 161
$Precision$	The precision of the dataset after applying the privacy protection mechanism, page 83
ps	A privacy statement, page 104
$Q(T)$	The result of query Q on table T, page 25
QI_T	A quasi-identifier of table T, page 12

R	The set of all possible retention periods, page 104
r	The chosen retention period, page 105
s	A strategy profile, page 52
S_I	The set of all pure strategies available to player i, page 53
s_i	The strategy of player i in strategy profile s, page 53
s_{-i}	Strategies of all players except for player i in strategy profile s, page 58
T	A data table, page 12
$T(A_1, ..., A_j)$	A table T with attributes A_1, ..., A_j, page 12
$t[A_i, ...A_m]$	The sequence of values for attributes A_i,...A_m on row t, page 12
T^*	The anonymized version of data table T, page 159
t_i^x	The type x of player i, page 63
$t_i^{f,j}$	The subtype j of a privacy fundamentalist data provider i, page 81
$t_i^{p,j}$	The subtype j of a privacy pragmatist data provider i, page 79
$U_i^{f,j}$	The utility of a privacy fundamentalist data provider i, page 81
$U_i^{p,j}$	The utility of a privacy pragmatist data provider i, page 79
U_i^u	The utility of a privacy unconcerned data provider i, page 78
U_{DC}	The payoff to the data collector, page 85
$U_{DC}[o]$	The data collector's maximum payoff by accepting offer o, page 116
U_{DU}	The payoff to the data user, page 84

V	The set of all possible visibility levels, page 104
w_1	The weight of privacy gain in privacy pragmatists' indifference curve, page 79
w_2	The weight of incentive gain in privacy pragmatists' indifference curve, page 79
z	The parameter of the Laplace noise, page 175
`q`	The quasi-identifying attribute over which the predicate of data user's query is defined, page 160

Chapter 1

Introduction

1.1 Background

Privacy is an essential requisite for people and society's welfare. The right to privacy is known to be necessary for individual's physical, psychological, social, and financial well-being [Sol08]. Privacy is also considered important to promote human relationships, freedom, and democracy. Although human desire for the right to privacy has a long history (some instances date back to ancient civilizations in Greece and China to conceal sexual activities), day to day advances in technology introduce new dimensions of privacy requirements and increase privacy concerns in public. Extensive information on individual's medical histories, financial records, residence and geographic information, biological material, and insurance claims are collected and stored across many (seemingly isolated) databases worldwide. Data integration and digital analysis on these large volumes of personal information are further facilitated by advanced networking and information technologies.

"Technology... is a queer thing. It brings you great gifts with one hand, and it stabs you in the back with the other."[1] While massive collections of personal information are precious resources for data mining, extracting undiscovered patterns, and predicting future trends, they pose significant threats to respondents' privacy. In fact, Solove [Sol08] has recognized multiple privacy threatening activities involved in three different phases of information collection, information processing (with potential threats of data aggregation, identification, insecurity, secondary use, and exclusion), and information dissemination (with potential threats of disclosure, exposure, increased accessibility, blackmail, appropriation, and distortion).

To this date, various privacy protection methods have been proposed to protect privacy

[1] C.P. SNOW, *New York Times*, 15 March 1971

of respondents while maintaining the usefulness of data for applications that are not privacy invasive. Two commonly used classes of these methods are known as privacy policy declaration methods and sanitization methods. Privacy policy declaration methods (such as P3P[CLM+02b], EPAL[AHK+03], and XACML[Mos05a]) focus on establishing a concise set of statements as a privacy policy and seeking for data providers' consent before collecting their information. They adopt the perspective that believes as long as data practices adhere to the agreed upon policy, collectors and users of data are not legally and ethically responsible for privacy breaches and their consequences. Data sanitization methods (such as k-anonymity [Swe02b], l-diversity [MKGV07], t-closeness [LLV07], and differential privacy [DMNS06]) apply different techniques (such as data generalization, suppression, and perturbation) to remove the risk of privacy breaches before it is made available to users of the data. These techniques postulate that once a private data table is sanitized, publishing it does not invade the privacy of the respondents. Therefore, data providers' consents are assumed to be unnecessary.

Providing access to a dataset such that no individual can learn anything beyond what she already knows, is known to be impossible [Dwo06]. Consequently, none of the proposed privacy protection methods can guarantee absolute privacy. They all include at least one privacy parameter (for example, privacy statements in privacy policy declaration methods, k in k-anonymity, and ϵ in differential privacy) that adjusts the guaranteed level of privacy.

1.2 Problem Statement

Different choices for values of privacy parameters result in different levels of privacy. As more privacy is promised, utility of information drops since more data manipulations and restricting rules are applied to the data table. On the one extreme, respondents can achieve full privacy when nothing about the collected data is revealed to data users. On the other extreme, the maximum information utility level is attained when the collected dataset is

provided to data users in its entirety with no privacy considerations. In the continuum between absolute privacy and absolute utility, the question is where the line should be drawn? In this dissertation the research problem is balancing privacy against the countervailing need for information utility.

In some literature (on data sanitization methods) [LDR08, XT06, LWFP06, BA05], this problem is approached by choosing a fixed privacy protection level and searching for an optimum transformation for the data table that offers the maximum utility while fulfilling the required level of privacy protection. More recent studies have targeted the privacy/utility tradeoff problem as an independent challenge [LS08, LL09, SSNDA10, MKS11, ZO10]. Some of these target a specific privacy protection method with a predefined workload and subjectively explain utility loss as higher privacy is provided [SSNDA10, MKS11]. Other work focuses on guiding the process of balancing the privacy/utility tradeoff. However, they often provide very broad guidelines (*e.g.*, [LL09]) or do not consider the true motivation behind the need for better privacy. Without considering the motive for privacy promises, it is unclear how to weight the relative merits of privacy protection versus that of information utility. Therefore, deficiency in the latter inevitably leads to relying on the data collector's intuition when choosing appropriate weights for privacy and utility requirements (see [ZO10] and [LS08]). Clearly, providing a concrete framework to balance the privacy and utility tradeoff is still an open question.

Any party who wishes to collect and use personal information must first choose a privacy protection method and then decide on values for its privacy parameters. This thesis proposes a generic solution to produce values for privacy parameters such that a balanced privacy/utility tradeoff is induced. The provided model can also be used to guide data collectors in choosing a privacy protection method based on its profitability. Besides their direct application in setting privacy parameters, we believe that balanced levels of privacy/utility tradeoff can provide reasonable reference points for comparing different privacy protection

methods. Exploring these balanced levels is the first step in designing mechanisms to direct the process (of establishing privacy parameters) towards solutions that offer higher profits to the society.

1.3 Objectives

This thesis aims at answering the question "How to establish values for privacy parameters to achieve a balanced tradeoff between privacy and utility?" Consequently, the objectives of this work can be summarized as:

- To provide a better understanding of the dynamics of private data collection and clarify the true motivation behind providing better privacy.
- To present a reasonable interpretation for "balanced" tradeoff between privacy and utility.
- To propose a generic solution that receives the specifications of a privacy protection method as the input and produces the balancing values for its privacy parameters.
- To demonstrate the application of the provided solution in capturing the effects of different problem settings on the balancing values found for the privacy parameters.
- To recommend reference points for comparing different privacy protection methods.

1.4 Scope

Data privacy is a multi-dimensional challenge. To address a data privacy problem one must clarify "who's privacy is protected?", "What types of data are collected (that impose a risk to privacy)?", and "What types of privacy protection method are used?". The answers to these questions delimit the scope of this study.

Depending on who's privacy concerns are addressed, three different dimensions have been recognized [DFS09] for information privacy. These dimensions are *respondents privacy*, *owner privacy*, and *user privacy*. The respondent privacy dimension aims at protecting privacy of respondents who donate information to a dataset. The focus in owner privacy dimension is on executing multiparty computation such that a query is collectively answered without leaking the underlying information. Finally, in user privacy dimension, the goal is protecting privacy of the queries submitted by a data user to an interactive database. In this study, the focus is restricted to respondent privacy.

We concentrate on privacy challenges when statistical data is to be disclosed. A statistical dataset is a collection of personal information organized in rows (tuples) and columns (attributes). Each row is a personal record provided by a respondent and lists values for each of the attributes, such as name, age, salary, disease, *etc.* of the respondent. In most statistical data disclosure problems the respondents are assumed to be aware of the data collection. We only consider those situations where data providers (respondents) know that their personal information is to be collected and can choose whether to provide such information or not.

Several methods have been proposed to protect privacy of statistical databases. Two common approaches are data sanitization and privacy policy declaration. In a non-interactive data sanitization method, data is perturbed before it is published. An interactive data sanitization method can be either a query restriction method or output perturbation method. Privacy policy declaration approach mainly focuses on the agreement on privacy considerations and implements the promises using a combination of privacy and security control mechanisms. This study does not address the challenges involved in query restriction methods. Moreover, security control mechanisms employed in privacy policy declaration methods are beyond the scope of this work.

1.5 Methodology

To answer the research question, we choose to analyze the problem from an economic perspective with the simple principle that says "all rational agents attempt to maximize their profits."

The real motivation behind providing better privacy, can be explained by how it affects one of the party's profits. A higher level of privacy protection satisfies more respondents and leads to a larger (and hence more valuable) dataset. Consequently, as an agent seeking a higher profit, the data collector must attempt to establish higher privacy protection.

Instead of searching for "optimum" combinations, this thesis finds the "balanced" combinations of privacy and utility. An optimum combination provides the maximum profit to one of the parties, but in finding the balanced combinations we consider the decisions made by three different parties who can impact the chosen value for the privacy parameter. These parties are: data users, data collectors, and data providers (respondents). We define a balancing tradeoff between privacy and utility as the outcome of a state where the attempts of these three parties' in maximizing their gain reach an equilibrium. We believe this is a more reasonable tradeoff concept since even if a combination of privacy and utility is optimal for one party, the combination may never be achieved because of the inter-dependencies between the decisions made by all stakeholders of the data.

The notion of equilibrium suggests the use of game theory. The interactions involved in establishing a privacy parameter value for an arbitrary privacy protection method are modeled as a game that captures the inter-dependencies between parties' decisions. The game's solution is produced analytically for the generic model and for three different instantiations of it when the underlying privacy protection method is chosen to be: (1) a privacy declaration method, (2) a non-interactive data sanitization method, and (3) an interactive data sanitization method. The implications of the analytical results are then studied by simulating them for some synthetic problem settings.

To sum up, this thesis applies an inter-disciplinary research approach by borrowing concepts from game theory (and microeconomics such as indifference curves) in solving a challenging problem in data privacy. The research method is partly modelling (providing the generic game model), partly analytical reasoning (solving the game with backward induction), and partly simulation (studying the effects of different problem settings on balancing values of privacy parameters).

1.6 Outline of the Dissertation

We start the thesis with a literature review on different privacy protection mechanisms in Chapter 2. We organize these mechanisms into two main classes: data sanitization (interactive and non-interactive) and privacy policy declaration. The positive and negative points in each of these two classes are explained which encourage an intermediary approach adopted in this thesis. A brief review on how the problem of privacy/utility tradeoff has been approached in different literature is also offered. Since our approach relies on data providers' privacy preferences, we briefly summarize the findings of some major social science studies on individual's privacy behaviour.

In Chapter 3, the required game theory definitions and concepts are provided. Using this knowledge, we review other works that also use game theory to address a privacy challenge. Consequently, the literature review in the thesis is provided partly in Chapter 2 and partly in Chapter 3 (since some basic knowledge about game theory was required).

Chapter 4 explains the backbone of our proposed model. In this chapter, a generic extended-form game with incomplete and imperfect information is defined. Then, a step-by-step method to find the game's equilibria (and hence the balancing privacy parameter values) is described based on the best responses of data providers, the data collector, and the data user.

Chapter 5 demonstrates how the generic game model can be used to find balanced pri-

vacy policy settings within a sample privacy policy declaration method. With the concrete function definitions that are instantiated for the game, a detailed analysis of the game's equilibria is offered and the results are explained in two lookup tables, each with seven different cases. We then choose a simpler scenario to fix some of the variables in data user's multi-variable utility function. For this simplified scenario, we show how the problem space can be partitioned into 22×22 different classes. Each class explains the game's equilibria for the problem settings that match the class's criteria. Applications of the game's solution are illustrated with two examples.

Chapter 6 describes the game's instantiations for k-anonymity as a non-interactive sanitization method, and differential privacy as an interactive sanitization method. For each instantiation, precision estimate functions are defined carefully based on how the method is implemented. Following the steps explained in Chapter 4, each game is analytically solved. These solutions are simulated for some synthetic problem settings and the effects of privacy protection cost, number of data providers, data providers' privacy awareness, and privacy trust on the balancing values of privacy parameters are illustrated. At the end of this chapter, we provide some directions on how the game results can be used as reference points within the process of comparing different privacy protection methods. As an example, we discuss how k-anonymity and differential privacy can be compared.

In Chapter 7, we conclude the thesis with explaining how the research objectives are achieved, our main contributions, limitations of the model, and future application of the results.

Chapter 2

Literature Review

Before explaining our game theoretic model to find balanced tradeoffs between privacy and utility, we review the related literature on privacy protection methods, the challenge of privacy/utility tradeoff, privacy behaviour of data providers (since their decisions influence the outcome of privacy settings), and game theory applications to solve privacy challenges. In this chapter, we summarize the literature on all of the specified related topics except for the last one. Literature review on application of game theory in data privacy is deferred to the next chapter after we explain some of the basic definitions in game theory.

2.1 Privacy Protection in Data Repositories

In the past few decades, the rapid growth of the Internet and information technologies and the massive data collection about individuals has raised privacy concerns among those who use these technologies as part of their daily lives. Two common approaches to address data privacy concerns in this context are *data sanitization* and *privacy policy declaration*. In the former approach, a data collector collects private information from individuals with an implicit promise of full privacy protection. A data anonymization/perturbation method is implemented by the data collector to transform raw data before making it available to users of the data. In the latter approach, before collecting private information, a data collector announces a privacy policy and data providers have to agree to the policy before providing their information.

Dwork [Dwo06] proves the impossibility of providing access to a dataset such that no individual can learn anything beyond what she already knows (before accessing the dataset). More specifically, one can always find some piece of auxiliary information to combine with

seemingly harmless knowledge in order to learn private information. Therefore, contrary to the initial promise, data sanitization methods cannot provide full privacy protection to data providers. The privacy policy declaration approach (such as P3P[CLM+02b], EPAL[AHK+03], and XACML[Mos05a]) attempts to overcome this problem by *partially* protecting data providers' privacy up to a level promised by an agreed upon privacy policy. Although these methods control the recipients and purposes of data usage, they do not protect privacy of the collected information against malicious (but authorized) recipients (or a very poor protection is provided). In the following two sections we review some of the major techniques in both data sanitization and privacy policy declaration approaches.

2.2 Data Sanitization

Data sanitization aims at removing the risk of a privacy breach by modifying the original private information before providing it to users of the data. One essential question to be answered by any data sanitization method is "what constitutes a privacy breach?". Generally, any specific information revealed about an individual is considered as a privacy violation whereas aggregate information about the participants in a dataset is acceptable information to be released [MT07]. Some researchers [Swe02b, MKGV07, LLV07] believe that only a subset of attributes, known as *sensitive attributes*, in a dataset should be considered private and the values of other attributes collected about an individual are safe to be released. Other researchers [SSNDA10, LDR06b] suggest that the association rules between non-sensitive attributes and sensitive attribute values should also be protected. However, Dwrok [Dwo11] and the subsequent literature on differential privacy (see Section 2.2.2.1) postulate that aggregate information (including the association rules) that are produced by a dataset is safe to release as long as the participation of any single individual in the dataset does not largely affect the aggregate results.

Depending on the chosen definition for a privacy breach, various privacy goals have been

set for data sanitization methods. In some data sanitization schemes (such as l-diversity [MKGV07]), the goal is to limit the difference between an adversary's view on an individual before and after publishing the private dataset. A number of authors [LLV07, LL09] believe that the distribution of the sensitive attribute values must be treated as public information and the privacy protection method has to limit the difference between this knowledge and the adversary's knowledge after publishing the private dataset. Dwork [Dwo11, DMNS06] argues that depending on the adversary's background knowledge, publishing a dataset might reveal private information about some individuals who are not even in the dataset. She believes that datasets are published to increase public's knowledge and hence this type of privacy violation is inevitable. For instance, if an adversary knows that John drinks heavily and the dataset teaches that esophageal cancers have high correlation with excessive alcohol consumption then the improvement in the adversary's guessing odds is not a privacy breach. Consequently, Dwork [Dwo11, DMNS06] revises the privacy goal as to limit the difference between an adversary's view on an individual before and after *the individual joins the private dataset*.

Adopting different definitions for a privacy breach and the privacy goal, various data sanitization method have been proposed [Swe02b, MKGV07, LLV07, XT06, Dwo06]. Data sanitization methods operate in two different modes: interactive and non-interactive. Methods that are designed for non-interactive settings are also known as *anonymization* techniques [DMNS06]. These techniques use generalization, suppression, and/or input perturbation methods to publish a publicly available sanitized version of the dataset. Interactive settings do not provide the dataset itself to any data user. Instead, they offer a query-response interface that answers queries in a private way. Interactive sanitization mechanisms (also known as *data perturbation* methods) usually add noise to query result sets.

Each data sanitization mode (interactive and non-interactive) has its own advantages and disadvantages. The non-interactive mode provides a "table" view of the dataset to

data users which is strongly preferred by statisticians. Moreover, data utility remains the same regardless of the number of users and number of queries asked from the dataset. The main difficulty with this method is that the ultimate use of data is usually unknown at the point of anonymization. Therefore, data utility is mostly optimized according to generic utility metrics (see Section 2.2.3.1). On the contrary, the interactive mode can provide a higher data utility if very limited queries are allowed. However, the tabular structure of data is not provided to the data user and to avoid inferences based on previous queries, more noise is added to the results as the number of queries increase (even if the same query is repeated). In this chapter, as a non-interactive sanitization method, we review one of the pioneer anonymization techniques known as k-anonymization and the sequel methods that refine it. From the interactive mode we discuss differential privacy as the most recent and strongest privacy protection mechanism in this category.

2.2.1 Data Anonymization

To publish a publicly available and yet private version of a dataset, a naive approach suggests to remove all explicit identifiers such as name and address. But even after removing such identifiers, privacy of the data providers might still be compromised by matching their data to other data sources (known as linking attacks) [Swe02b]. In fact, Sweeney [Swe00] shows that 87% of the US population are very likely to have a unique combination of values for the attributes *ZIP code*, *gender*, and *date of birth* that makes them very vulnerable to linking attacks. As a result, most data anonymization techniques [SS98, Swe02b, MKGV07, LLV07, Dwo06] first remove the identifiers and then apply some data generalization and/or suppression methods to the data records to protect privacy of the data providers.

2.2.1.1 Preliminary Definitions

Before explaining the details of different anonymization techniques, we need to clarify some definitions:

Data table: A data table organizes data in rows and columns. A column represents an attribute with a specific domain and a row contains personal information about an individual for each of the attributes (often referred to as a tuple or a data record). A table T with attributes A_1, ..., A_j is represented as $T(A_1, ..., A_j)$ and the sequence of values for attributes A_i,...A_m on row t is denoted by $t[A_i, ...A_m]$. In this work we use the terms data table and dataset interchangeably.

Quasi-identifier: A subset QI_T of the attributes $\{A_1, ..., A_q\}$ is called a quasi-identifier if there exists a row t in T such that $t[QI_T]$ can be linked to external data to re-identify the individual who has provided row t [Swe02b]. It is assumed that quasi-identifiers can be accurately identified by the entity who collects personal information.

Equivalence Class: An Equivalence Class [Swe02b] EC with respect to a quasi-identifier $QI_T = \{A_1, ..., A_q\}$ is a subset of tuples $EC = \{t_1, t_2, ..., t_m\} \subseteq T$ such that $t_1[QI_T] = t_2[QI_T] = ... = t_m[QI_T]$.

The last two definitions are first introduced in k-anonymity [Swe02b] but also used in subsequent non-interactive sanitization methods. Partitioning a data table into equivalence classes of indistinguishable tuples (based on their quasi-identifier values) provides a means to to prevent identity discloser.

2.2.1.2 Essential Anonymization Operations

The essence of most anonymization techniques is an operation known as data generalization. Generalization (recoding) is defined as substituting the value of an attribute in a row with a less specific (but semantically consistent) value [Swe02a]. This operation is sometimes combined with an optional data suppression step. Data suppression involves withholding contents of a row or an attribute in a row from being published. Data generalization and

suppression operations provide the advantage of maintaining data integrity while anonymizing private datasets.

The proposed generalization models can be categorized based on at least two criteria: Global *vs.* Local Recoding and Hierarchy-Based *vs.* Partition-Based Generalization [LDR05a]:

- Hierarchy-Based Generalization - In Hierarchy based generalization, a domain generalization hierarchy is defined for each quasi-identifying attribute (in case of single dimensional recoding) or for the combination of domains of all quasi-identifying attributes (in case of multi-dimensional recoding). A domain generalization hierarchy is a structure defined to organizes different levels of generalization of a domain D as a total order. The lowest level in the hierarchy is a set containing all possible values that an attribute can take (the domain). As we move up the hierarchy, the set of possible values becomes smaller since each member of a set represents a group of different values from the set at the immediate lower level. The root of each hierarchy is a set containing one single value that represents all the values in the actual domain. This level of generalization is used to implement data suppression [Swe02a]. More specifically, if domain D' is an ancestor of domain D'' in a domain generalization hierarchy then every value in domain D' is a generalization of a subset of values in domain D'' [LDR05a]. The goal of an anonymization algorithm is to find a generalization level in each domain hierarchy (in case of single dimensional recoding) or in the combined domain hierarchy (in case of multi dimensional recoding) that allows for the required level of privacy without over-generalizing the attributes. Since this model relies on predefined hierarchy structures, it is not as flexible as the partition-based generalization in the sense that only a very small part of the solution space is explored [LDR05a, Swe02a].

- Partition-Based Generalization - In partition-based generalization, the domain of an attribute is assumed to be ordered. A partitioning is defined as a set of cutting points that induce non-overlapping intervals in the domain [BA05]. This method of partitioning can be used only for global recoding.

- Global Recoding - Global recoding maps the domain of quasi-identifying attributes to more generalized values [LWFP06]. Two types of global recoding have been proposed: single dimensional and multidimensional global recoding. In single dimensional global recoding the domain generalization is done for the domain of each quasi-identifying attribute independently. Therefore, once a mapping is chosen, a single value $v \in D$ is always generalized to $v' \in D'$. Where D' is the generalized mapping of domain D. In multidimensional global recoding the cartesian product of domains of all quasi-identifying attributes is considered as one single domain to be mapped to a more generalized domain. More specifically, in a dataset with d quasi-identifying attributes, a multidimensional partitioning defines non-overlapping d-dimensional rectangular boxes to cover the cartesian product of quasi-identifying attributes' domains. Consequently, instances of a single attribute value $v \in D$ can be generalized to two (or more) different values depending on the values of other quasi-identifying attributes in the same record. In other words, instead of generalizing each attribute value separately, the "value vectors of quasi-identifying attributes" are generalized. Every single dimensional partitioning is also a multidimensional partitioning. Therefore, the optimal multidimensional partitioning is at least as good as any optimal single dimensional recoding. However, finding the optimal multidimensional partitioning is computationally expensive and in the case of k-anonymity the problem has been shown to be NP-complete (proven by reduction from integer programming [LDR05b]).

- Local Recoding - In local recoding, generalization is done at the cell level (each record is separately generalized) instead of domain level [LWFP06, LDR05b]. The local recoding model is an instance of the relaxed multidimensional recoding where multidimensional global recoding model is modified to allow for overlapping d-dimensional rectangular boxes. Each record is then mapped to the summary statistics of one of the boxes that contain the record [LDR05b].

Several anonymization techniques have been developed using the explained operations. We describe some of the main techniques with a focus on the most seminal technique known as k-anonymity [Swe02b].

2.2.1.3 k-Anonymity

k-Anonymity [Swe02b] is a seminal data anonymization technique which protects the data providers against identity disclosure. In this technique, the anonymization operators (see Section 2.2.1.2) are applied to data records until each record in the released dataset becomes indistinguishable from at least $k-1$ other records.

k-Anonymity is defined as follows::

k-anonymity A table T with quasi-identifier QI_T is k-anonymous if and only if for every row t in T there are at least $k-1$ other rows, $t_1, ..., t_{k-1}$, such that $t[QI_T] = t_1[QI_T] = ... = t_{k-1}[QI_T]$ [Swe02b]. In other words, a table T is k-anonymous if each equivalence class (with respect to QI_T) has at least k tuples.

Finding an optimum k-anonymization of a data table is an NP-hard problem [BA05]. Several k-anonymization algorithms have been proposed [LDR06b, Swe02a, LDR05a, BA05, LWFP06, LDR05b]. These algorithms generalize (and suppress) records until the required level of privacy is guaranteed while utility of the anonymized dataset is kept close to its optimum. The main difference between these anonymization algorithms lies in how data utility is measured and the heuristic used to find a high quality private data table. Some

of the proposed algorithms are not tied to any specific usage for the published data table. There are also workload dependent algorithms that attempt to maximize data utility with regard to a specific query or data mining task. In the following we provide a brief overview of a number of k-anonymization algorithms:

- **MinGen -** The MinGen algorithm [Swe02a] applies single dimensional hierarchy-based generalization to achieve k-anonymity. This algorithm assumes predefined generalization hierarchies for the domain of each quasi-identifying attribute. The goal of the algorithm is to find a generalization level for each of the hierarchies so that the combination of the selected levels produces a dataset that satisfies k-anonymity with minimal overall generalization (defined based on the height of the combined generalization hierarchy). To find the best solution, the algorithm simply tries all possible combinations of generalization levels and therefore it is not efficient.

- **Incognito -** The Incognito algorithm [LDR05a] attempts to achieve the same goal as the MinGen algorithm in a more efficient way. Instead of considering *all* possible combinations of generalization levels, Incognito starts by considering each quasi-identifying attribute, qi, separately and checks whether the dataset is k-anonymous with respect to qi at each level of the domain generalization hierarchy of qi. It keeps the part of the hierarchy that guarantees k-anonymity with respect to qi. In the next phase, it combines all the stored hierarchies to construct generalization hierarchies for all possible combinations of two quasi-identifying attributes. It checks the newly constructed hierarchies to test the dataset's k-anonymity property at each multi-domain generalization level (each node) and keeps the sub-hierarchies that provide k-anonymity. In the same manner, in each subsequent phase a larger combination of quasi-identifying attributes are considered until a final hierarchy is constructed out

of all of the quasi-identifying attributes. The minimal generalizations are the ones that are at the lowest level of the final hierarchy.

- **KACA -** The KACA (k-Anonymization by Clustering in Attribute hierarchies) algorithm [LWFP06] uses clustering techniques to implement a hierarchy-based local recoding. The algorithm is built on the concept of distance between two equivalence classes. The distance between two equivalence classes is defined by the pairwise distance between each pair of corresponding tuples they contain. To find the distance between two tuples t_1 and t_2, the generalization tuple $t_{1,2}$ is created such that the value of each quasi-identifying attribute in $t_{1,2}$ is the closest common generalization of the two corresponding values in t_1 and t_2. The distance between tuples t_1 and t_2 is defined as the summation of the distortions caused by generalizing t_1 to $t_{1,2}$ and generalizing t_2 to $t_{1,2}$. The amount of distortion is calculated based on the WHD utility metric (see Section 2.2.3.2). The algorithm starts with equivalence classes of size 1 (every single tuple is an equivalence class). At each step two equivalence class (of size less than k) with minimal distance are arbitrarily chosen to merge.

- **Optimal -** To avoid the shortcomings of hierarchy-based recoding, the Optimal algorithm [BA05] implements a partition-based single dimensional global recoding (with tuple suppression). This algorithm assumes a total order for the values of $D_1 \times D_2 \times ... \times D_d$ where there D_i is the domain of the i^{th} quasi-identifying attribute and the total number of quasi-identifying attributes is d. A generalization is defined as finding cutting points along this composite domain. The problem of finding the optimal cutting points is framed as a powerset searching problem. The algorithm works by creating the set enumeration tree and uses pruning and heuristic strategies to reduce the time complexity of examining the utility of partitioned data tables induced by each node of the

tree. This algorithm is developed to optimize utility when the Discernibility utility metric (see Section 2.2.3.1) is used. However, the definitions of pruning techniques are highly dependent on characteristics of the two utility metrics used and without pruning the algorithm has to search the whole problem space. Moreover, scalability of the algorithm is questionable since the experimental analysis is only done for a data table with a single quasi-identifying attribute.

- **Mondrian -** A more efficient algorithm that implements partition-based multi-dimensional global recoding is known as the Mondrian Algorithm[LDR05b]. This greedy algorithm has two phases: partitioning and recoding. At the beginning of the algorithm, the domain space is considered as a single region to be partitioned. Each region is recursively partitioned into two non-overlapping partitions by choosing a split value on one of the heuristically chosen dimensions (quasi-identifying attributes), *dim*. The split value is the median of the values that appear on dimension *dim* in the partition. Once the partitioning phase is done (no more allowable cuts can be found for any of the regions), tuples are mapped to summary statistics of the region they fall in.

- **Workload Dependent Mondrian -** Lefevre *et al.*[LDR06b] also proposed a workload aware version of Mondiran algorithm where data quality is measured with respect to the utility of the data for classification/regression workloads. In this version, they modify the basic Mondiran algorithm [LDR05b] to consider the ultimate data usage while choosing the split value on the domain of one of the quasi-identifying attributes. More specifically, to increase data utility, at each recursive step, their proposed algorithm considers every split value and chooses the one that leads to more homogeneous (based on entropy or Mean Square Error) partitions with respect to diversity of values of a target attribute.

2.2.1.4 Improvements on k-Anonymity

Although it is not explicitly mentioned anywhere, implementations of k-anonymity only consider *a single* set of quasi-identifying attributes. Fung *et al.* propose Top-down specialization [FWY05] as an extension to k-anonymity to anonymize datasets with more than one set of quasi-identifying attributes. More specifically, *all* possible subsets of attributes with a potential of uniquely identifying an individual are considered as *separate* quasi-identifiers. The privacy requirement of the Top-down specialization is specified as a vector $\langle k_1, ..., k_q \rangle$ where k_i defines the minimum size of equivalence classes with respect to quasi-identifier i.

To achieve the anonymity requirement of Top-down specialization method, Fung *et al.*[FWY05] propose an algorithm for classification workload. This algorithm starts from the most generic value for every attribute in the dataset and at each step uses heuristics to choose an attribute and refine its domain to more specific values. The heuristic aims at keeping equivalence class sizes as large as possible while creating more homogeneous equivalence classes with respect to the class labels of a target attribute. The algorithm stops when no more refinement can be found to preserve the privacy requirement.

Another drawback of k-anonymity is its vulnerabilities when multiple records belong to the same individual[XT06]. Xia and Tao [XT06] propose the Personalized Privacy Preservation (PPP) method as an extension to the k-anonymity technique where sensitive attribute values are also generalized based on data providers' preferences. In PPP, data providers are given the choice to select a node in the domain hierarchy of the sensitive attribute as their guarding node. The PPP algorithm implements global recoding to generalize quasi-identifying attributes and local recoding to generalize sensitive attribute values so that an adversary cannot infer an individual's sensitive attribute at a more specific level than her guarding node (with a probability higher than a threshold).

The PPP algorithm starts with the original dataset as a single equivalence class. At each iteration, it examines all possible generalizations of the dataset produced by a single

cut in the domain of one of the quasi-identifying attributes. For each of the newly created generalized datasets, it finds the best generalization for sensitive attribute values and keeps the anonymization with the least amount of information loss to be refined in the next iteration. The algorithm stops when no more cut can be found to further refine the current best anonymization. The range-based information loss metric (see Section 2.2.3.1) is used to measure data utility.

2.2.1.5 l-Diversity

k-anonymity aims at hiding the connection between a tuple and the individual that the tuple belongs to (also known as *identity disclosure* [LLV07]). This type of protection does not necessarily hide the association between an individual and her sensitive attribute value (also known as *attribute disclosure* [LLV07]). More specifically, sometimes the distribution of sensitive attribute values in an equivalence class can reveal the association between an individual and her sensitive attribute value to an adversary even if the adversary cannot distinguish the record that belongs to the target individual [MKGV07]. For example, in a k-anonymous dataset, it is possible for an equivalence class to have the same sensitive attribute value for all of the tuples. In this case, the adversary would be 100% confident about an individual's sensitive attribute only if she knows that the individual's record belongs to a specific equivalence class. In other words, identity disclosure always leads to attribute disclosure but attribute disclosure can occur without identity disclosure [XT06, LLV07].

l-diversity [MKGV07] recognizes the lack of diversity in sensitive attribute values as the cause of k-anonymity's vulnerability to attribute disclosure. To solve this problem, the l-diversity principle requires that at least $l \geq 2$ different values of the sensitive attribute appear in each equivalence class and the most frequent values have similar frequencies. Three implementations of l-diversity have been proposed so far:

- **Distinct l-Diversity -** This version requires that each equivalence class has at least l-different values of the sensitive attribute.

- **Entropy l-diversity -** This version requires that the entropy of fractions of records with the same value for the sensitive attribute be greater than or equal to $log(l)$ in each equivalence class.

- **Recursive (c, l)-diversity -** This version requires that frequency of the most frequent sensitive attribute value in each equivalence class be less than a constant factor, c, of total frequencies of $m - l$ least frequent values in the same class, where m is the number of distinct sensitive attribute values in the equivalence class.

To achieve l-diversity, a modified version of the Incognito algorithm [LDR05a] (see Section 2.2.1.3) can be used [MKGV07]. At each step of the modified version, instead of checking whether the dataset is k-anonymous, the generalized dataset is checked to determine whether it fulfills l-diversity requirements or not.

2.2.1.6 t-Closeness

Li *et al.*[LLV07] show that l-diversity is neither a necessary nor a sufficient privacy concept to prevent information leakage on the association between a data provider and her sensitive attribute value. More specifically, they show that if the sensitive attribute takes a few distinct values with different degrees of sensitivities, l-diversity over-generalizes the dataset. Moreover, the overall entropy of sensitive attribute values in the dataset defines an upper bound on the value of the parameter l (see entropy l-diversity in Section 2.2.1.5). It was also shown that sometimes an attempt to satisfy l-diversity might increase the risk of disclosure. For example, if the probability of having a certain disease is very low and the disease appears very frequently in an equivalence class to satisfy the l-diversity requirement a privacy breach has occurred. Furthermore, when sensitive attribute values are semantically close in an equivalence class, an attacker can learn important information about an individual's sensitive attribute.

To address the shortcomings of l-diversity, Li *et al.* propose an anonymization technique known as t-closeness [LLV07]. The main idea behind t-closeness is to define an adversary's prior belief as the knowledge she learns based on the overall distribution of sensitive attribute values in a fully generalized dataset. The goal of anonymization is defined as minimizing the amount of increase in the adversary's knowledge when she knows which equivalence class her target individual belongs to. To reach this goal, t-closeness suggests that distribution of sensitive attribute values in each equivalence class must closely resemble the overall distribution of sensitive attribute values in the dataset. The distance between the distribution of sensitive attribute values in each equivalence class and its overall distribution in the dataset is measured by the *earth movers distance measure* [RTG00].

To implement t-closeness, a modified version of Incognito [LDR05a] (see Section 2.2.1.3) is used to explore all possible full-domain generalizations of quasi-identifying attributes and choose the ones that satisfy the property of t-closeness with minimal amount of generalization.

2.2.2 Data Perturbation

Instead of publishing a dataset, data perturbation methods suggest the provision of an interface between data users and the data repository to answer queries on a dataset of personal information. This system ensures that answering queries does not cause privacy violations. Two major classes of data perturbation methods are query auditing and output perturbation.

In query auditing methods [NMK$^+$06, KMN05, KPR00], before answering a query the transaction log (history of queries answered in the past) is analyzed to determine if the query can disclose private information or not. A query is rejected if it is considered disclosive. This method has the two drawbacks of being computationally infeasible and disclosing private information even when the query is rejected [Dwo11].

Output perturbation methods rely on modifying the resultset of a query or the query itself

before providing the results to a data user [Dwo11]. A very common output perturbation technique is to add random noise to the output. If this technique is used, data is not truthful anymore and repeating the same query can reveal private information. Dwork [Dwo11] proves the relationship between the noise bound and an adversary's power to reconstruct the original dataset. With this bound she shows that adding noise to data is only practical in interactive privacy protection modes and not in data publishing. Here we only explain differential privacy [DMNS06] as the standard method of output perturbation.

2.2.2.1 Semantics of Differential Privacy

Differential privacy has recently become one of the most popular mechanisms to guarantee privacy. This method is classified as an output perturbation method and provides privacy protection in an interactive mode [Dwo11].

The essence of differential privacy is a simple principle: "running a query on any two datasets that only differ in one individual must provide the same result with a probability higher than a threshold" [DMNS06]. In this method a randomized function κ is applied to the original data table T and returns the answer to a requested query Q after adding some randomized noise to the original answer. The randomized function κ guarantees ϵ-differential privacy if for every pair of data tables T and T' differing in only one row and for all $S \subseteq Range(\kappa)$ the following holds [Dwo11]:

$$Pr[\kappa(T) \in S] \leq exp(\epsilon) \cdot \ Pr[\kappa(T') \in S] \tag{2.1}$$

To design the function κ, the amount of added noise must be chosen according to a property of the query known as *query sensitivity* [Dwo11]. Simply speaking, sensitivity of a query is the maximum difference between the query results on any two datasets differing in only one row. More formally, for all data tables T_1 and T_2 differing in only one entry, sensitivity of a query Q is defined as the smallest number ΔQ such that:

$$|Q(T_1) - Q(T_2)| \leq \Delta Q \tag{2.2}$$

Where $Q(T)$ represents the result of query Q on table T and the notation $|T|$ is used to denote the cardinality of a dataset T. Notice that the definition of query sensitivity is independent of the underlying dataset.

If more than one queries (or even more than one instance of the same query) is asked from the data table then sensitivity of the *query sequence* must be considered. Sensitivity of a sequence of queries Q_1, Q_2, ..., and Q_n can be safely assumed to be $\sum_{1 \leq i \leq n} \Delta Q_i$ [DMNS06]. However, Dwork *et al.*[DMNS06] showed some special cases where the real sensitivity of a sequence of queries is much smaller than their additive sensitivities. These cases include building a histogram and queries that can be accurately approximated by a small sample of the dataset.

2.2.2.2 Achieving ϵ-Differential Privacy

There are two different methods to design a function κ (see Section 2.2.2.1) such that it provides ϵ-differential privacy: the Laplace Mechanism [Dwo11] and the Exponential Mechanism [MT07].

Laplace Mechanism - Dwork [Dwo11, DMNS06, Dwo06] shows that a randomized function κ provides ϵ-differential privacy if it adds independent noise with a Laplace distribution $Lap(\frac{\Delta Q}{\epsilon})$ to the original results of a query Q with sensitivity ΔQ. Let $Q(T)$ and $Ans(Q(T))$ be the answer to query Q on data table T before and after applying the function κ to the results, respectively. We have:

$$Ans(Q(T)) = Q(T) + Noise \qquad where \quad Noise \propto Lap(\frac{\Delta Q}{\epsilon}) \tag{2.3}$$

The symbol $\propto$ is used to denote that the noise is proportional to the specified Laplace distribution. Recall that Lap($\frac{\Delta Q}{\epsilon}$) has density function $h(y) \propto exp(\frac{-|y|\ \epsilon}{\Delta Q})$. Equation 2.3 states that

$$Pr[Ans(Q(T)) = a] = Pr[Noise = a - Q(T)] = h(a - Q(T))$$

For any two noises n_1, n_2, we have:

$$\frac{h(n_1)}{h(n_2)} \leq exp(\frac{\epsilon}{\Delta Q} \cdot |n_1 - n_2|)$$

According to the definition of sensitivity, for any two data tables T_1 and T_2 differing in only one entry $|Q(T_1) - Q(T_2)| \leq \Delta Q$. Consequently we have:

$$\begin{aligned} \frac{Pr[Ans(Q(T_1))=a]}{Pr[Ans(Q(T_2))=a]} &= \frac{Pr[n_1=a-Q(T_1)]}{Pr[n_2=a-Q(T_2)]} \\ &= \frac{h(a-Q(T_1))}{h(a-Q(T_2))} \\ &\leq exp(\tfrac{\epsilon}{\Delta Q} \cdot |a - Q(T_1) - a + Q(T_2)|) \\ &\leq exp(\tfrac{\epsilon}{\Delta Q} \cdot \Delta Q) = e^{\epsilon} \end{aligned} \tag{2.4}$$

This proves how adding noise based on a Laplace distribution provides ϵ-differential privacy. Dwork *et al.*[DMNS06, Dwo11] extended the application of this method to query functions that return results of higher dimensions (not just a scalar number but a vector of arbitrary dimension).

Exponential Mechanism - A more general technique to achieve ϵ-differential privacy is known as the exponential mechanism [MT07]. This mechanism relies on a utility function $u(T, y)$ that measures the optimality of answer y to query Q of the data table T. The exponential mechanism answers query $Q(T)$ with some value $y \in Range(Q)$ with a probability proportional to $exp(u(T, y) \cdot \frac{\epsilon}{\Delta u})$. In other words:

$$Pr[Ans(Q(T)) = y] \propto exp(u(T, y) \cdot \frac{\epsilon}{\Delta u}) \tag{2.5}$$

It can be easily seen that this method provides ϵ-differential privacy. Let T_1 and T_2 be two versions of the data table differing only in a single entry. The difference in probabilities of getting the same output y with both versions of the dataset is as follows:

$$\begin{aligned} \frac{Pr[Ans(Q(T_1))=y]}{Pr[Ans(Q(T_2))=y]} &= \frac{exp(u(T_1,y)\cdot\frac{\epsilon}{\Delta u})}{exp(u(T_2,y)\cdot\frac{\epsilon}{\Delta u})} \\ &= exp((u(D_1, y) - u(D_2, y)) \cdot \tfrac{\epsilon}{\Delta u}) \\ &\leq exp(\Delta u \cdot \tfrac{\epsilon}{\Delta u}) = e^{\epsilon} \end{aligned} \tag{2.6}$$

Unlike the Laplace method, this mechanism is applicable to all kind of queries even when adding noise to query answers does not make sense. McSherry and Talwar [MT07] show how to define a utility function to make the Laplace mechanism a special case of the Exponential mechanism.

2.2.2.3 Advantages and Drawbacks of Differential Privacy

Compared to sanitization techniques, providing ϵ-differential privacy is not very expensive. This method does not require an optimized algorithm to design the privacy protected dataset according to a privacy/utility tradeoff. Moreover, the amount of inaccuracy in the query results are always bounded [1] [Dwo11] and does not depend on the size or contents of the dataset.

However differential privacy is not always effective in preventing privacy breaches. Kifer and Machanavajjhala [KM11] show that the promise of differential privacy is only valid when records in the dataset are independent of each other. Their *No Free Lunch* theorem [KM11] proves that if there are correlations between values of tuples in a dataset, then no data perturbation method with sufficient utility can prevent privacy breaches. Therefore, they suggest that the crucial information to be hidden is not the existence of an individual in a dataset but the *participation* of an individual to produce query results. The notion of participation is case based and there is no general privacy protection paradigm to prevent breaches when correlations exist between records in a dataset. Moreover, Kifer and Machanavajjhala [KM11] argue that protecting a dataset against the most knowledgable attacker does not necessarily protect the dataset against a less knowledgable one. In fact, by assuming a powerful attacker, a privacy protection algorithm only focuses on inducing noise on the small portion of information that is assumed to be unknown to the attacker. Therefore, the rest of the information remains unprotected and leaks information to a less knowledgable attacker. Consequently, they show that the widely accepted belief that asserts "differential privacy

[1] If Laplace noise is used the inaccuracy is equal to the standard deviation of the Laplace noise: $\frac{\sqrt{2}\Delta Q}{\epsilon}$.

provides protection agains *any* kind of attacker" is not correct.

Even in circumstances where differential privacy is capable of fulfilling its promises, it might not be the most desired privacy protection method. Often researchers (data users) would like to be able to look at the dataset as a whole even if the data is not very specific [Dwo11]. In other words, interactive privacy protection modes are generally not desirable. Moreover, as the number of queries on the dataset increases, more noise must be added to the results and after a point the dataset will not be useful anymore. Finally, adding noise to data records violates the integrity constraint, which is a classic requirement in databases.

2.2.3 Privacy/Utility Tradeoff [2]

Achieving privacy via sanitization comes at the cost of losing some information. Often a higher level of privacy protection is only achievable via increasing the amount of data manipulation (generalization, suppression, and/or perturbation) which leads to a lower data utility. Several utility metrics have been proposed to evaluate the tradeoff of privacy/utility either through the sanitization process or after sanitization has been done. A utility metric can be either general purpose or workload dependent. In the following we summarize some of these metrics and explain how they are used to evaluate privacy/utility tradeoff.

2.2.3.1 General Purpose Utility Metrics

Often the specifications of the data analysis procedures to be conducted on a dataset (referred to as the *workload*) is not pre-determined. In these cases, similarity between the original data table and the sanitized version of it is considered as an indicator of data utility. Various

[2]Since the problem of privacy/utility tradeoff is mainly studied in the context of data sanitization methods, we explain the related literature within the data sanitization section. This problem also exists in privacy policy declaration methods. However, it is mostly overlooked in the latter context because data manipulation is usually not considered as a major concern in privacy policy declaration. We believe that any privacy protection mechanism must use some kind of data manipulation and hence balanced privacy/utility tradeoffs must be found even if privacy policy declaration is to be used.

general purpose utility metrics have been proposed so far, some of which are explained in the following.

- **Discernibility Metric** [BA05] **-** This metric measures data utility based on the size of the equivalence classes. Information loss is defined as total number of pairs of tuples that are indistinguishable from each other plus an additional penalty of $|T|$ (cardinality of the original data table) for each suppressed tuple.

- **Normalized Average Class Size** [LDR05b] **-** This metric uses the ratio between total number of tuples and total number of equivalence classes multiplied by k (the minimum size of equivalence classes) as an indicator of information loss.

- **Classification Metric** [BA05] **-** To measure information loss, this metric assigns a penalty of one unit to every suppressed tuple and every tuple that is in minority (with respect to the class label) within its own equivalence class.

- **Modification Rate** [LWFP06] **-** This metric measures information loss as the percentage of records that are modified to sanitize the data table. It does not consider the magnitude of modification in each record.

- **Weighted Hierarchical Distance (WHD)** [LWFP06] **-** In this metric, for each quasi-identifying attribute, the distance between every two consecutive levels in the domain hierarchy is defined in terms of weights. For a value v that is generalized to value v' this metric counts how many levels we must climb up the domain hierarchy to get the value v' and adds up the weights of each level elevation. This weighted sum is normalized by the sum of the weights from the root to the leaves of the domain hierarchy. The distortions of generalization for a tuple is the the summation of the WHD for every quasi-identifying attribute in the tuple. Consequently, the amount of information

loss for a sanitized data table is the summation of the distortions across all tuples of the table.

- **Range-Based Information Loss Metric** [XT06] - This metric measures information loss caused by generalizing a value v to v' in terms of the ratio of the range of values covered by v' to the range of values covered by the original domain of the attribute. The amount of information loss for a tuple is defined as the weighted sum of information loss across all attributes. The total amount of information loss across all tuples determine the information loss of a sanitized dataset.

- **s-Diversity Utility Metric** [LS08] - s-Diversity defines diversity of an equivalence class as the summation of the diversity factors[3] for each quasi-identifying attribute. The amount of information loss in a sanitized data table is the average diversities across all equivalence classes.

2.2.3.2 Workload Dependent Utility Metrics

When the ultimate purpose of data usage is known *a priori* a more reliable utility metric must measure the amount of information loss in the context of data application (such as data mining and queries). However, some scholars argue that when the workload is pre-determined a better option would be to perform the data mining task on the original data table and just publish the result [LL09].

A popular target workload is the data classification/regression models. Often, when the ultimate mining task on a private dataset is classification (or regression), a single nominal (or numeric) attribute other than the sensitive attribute is chosen to be the target attribute and the miner attempts to predict the value of the target attribute based on values of some

[3]Diversity factor of an attribute in an equivalence class specifies how diverse the value of the attributes are in the class.

predictor attributes (also chosen from the quasi-identifying attributes) [LDR06b, SSNDA10]. In the following we explain three different workload dependent metrics:

- **Mining Utility Measure** [SSNDA10] **-** This method measures mining utility of each tuple in terms of how close the estimated value for the target attribute is to its original value in the dataset. Total utility of a mining procedure on a target attribute is the weighted sum of mining utilities over all tuples in the dataset. The decline in mining utility caused by sanitization is measured by the difference between utility of data mining on the original and sanitized versions of the dataset.

- **Region-Based Utility Metric for Select Queries** [LDR08] **-** Data utility for select queries can be measured based on how much the query region (specified in the select predicates) is aligned with bounding boxes of partitions created during anonymization. LeFevre *et al.*[LDR08] adopt the semantics under which a select query (on an anonymized dataset) returns all tuples that are partitioned into any of the regions that overlap with the requested query region. Therefore, imprecision is defined as the difference in size between the query outputs on the anonymized dataset versus original dataset.

- **Distribution-Based Utility Metric for Select Queries** [LL09] **-** In this metric, for each select predicate on quasi-identifying attributes, distribution of the sensitive attribute values across selected tuples is estimated for the sanitized dataset. This estimate is compared against the real distribution of sensitive attribute values in tuples selected from the original dataset. The average of the distribution distances for all sufficiently large result sets determines the amount of information loss.

2.2.3.3 Balancing Privacy and Utility Levels

A common approach to balance utility and privacy levels is to choose a minimum privacy protection level first and incorporate heuristics in the sanitization algorithm to maximize utility while keeping privacy protection level above the minimum. In this approach, one of the specified utility metrics (see Sections 2.2.3.1 and 2.2.3.2) are used *within* the sanitization algorithm to evaluate the fitness of a potential solution.

More recently, researchers [LS08, LL09, SSNDA10, MKS11, ZO10] have been working on addressing the balancing act between privacy and utility as an independent problem. In these articles usually another metric is used to measure the amount of privacy breach. The following lists three such metrics:

- **Mining Privacy Breach** [SSNDA10] **-** In this metric, the adversary is assumed to be interested in predicting the value of a target attribute (usually the sensitive attribute) that is different from the data users' target attribute. The amount of privacy breach for each tuple t is determined based on how accurate an adversary can mine t's value for the target attribute only by accessing the sanitized dataset. Total privacy breach is defined as the weighted sum of privacy breaches over all records.

- **Distribution-Based Privacy Loss** [LL09] **-** This metric measures the amount of privacy loss for an individual t in terms of the difference between the distribution of sensitive attribute values in t's equivalence class (the class that t's tuple belongs to) and the overall distribution of sensitive attribute values in the dataset. The amount of privacy loss in a sanitized dataset is defined as the maximum privacy loss over all individuals.

- **s-Diversity Privacy Metric** [LS08] **-** To assess vulnerability of a sanitized dataset, s-diversity considers range disclosure risks. This metric measures the normalized sum of pairwise distances between sensitive attribute values that

appear in an equivalence class. Vulnerability of a dataset is defined as the average vulnerabilities among all equivalence classes.

Several frameworks have been proposed to evaluate and balance the tradeoff of privacy/utility. Sramka *et al.*[SSNDA10] propose a generic framework to evaluate utility and privacy of anonymized datasets with regard to data mining workloads. This framework is independent of any specific anonymization technique. Both the data user and the adversary are assumed to be data miners with different mining tasks (different target attribute). Data utility and privacy breach are measured as a mining utility measure (see Section 2.2.3.2) and mining privacy breach, respectively. Although the proposed framework is generic enough to cover various anonymization techniques, its applicability is limited to situations where data users and adversaries are data miners. Moreover, several scholars [LL09, Dwo11] believe that not every piece of knowledge that a malicious user might extract from a private dataset should be considered as a privacy breach; after all the ultimate purpose of publishing a dataset is to increase society's knowledge. Li and Li [LL09] describe utility as an aggregate concept and argue that the accumulative knowledge offered by a dataset determines its utility.

Machanavajjhala *et al.*[MKS11] provide an elaborate analysis on privacy/utility tradeoffs when differential privacy [DMNS06] is used in social recommendations. In a social recommendation system, structure of an underlying social network graph is used to recommend objects to users of the system. To prevent information leakage about the edges in a social graph, one solution is to design recommendation algorithms that satisfy $\epsilon-$differential privacy. Machanavajjhala *et al.*[MKS11] formalize the notion of utility in these systems and prove the relationship between utility levels and ϵ values. They further prove stronger lower bounds on the value of ϵ with regard to more specific utility functions. Their work shows an inevitable harsh tradeoff between privacy and utility in this specific application and the authors conclude that based on the application, differential privacy is not always the best

choice.

Li and Li [LL09] realize that despite the large volume of literature on privacy utility tradeoff, none of the existing works provide directions about how to choose an acceptable level for privacy and utility. They propose a generic framework based on Modern Portfolio Theory [EG95] to guide the process of making decisions on choosing an anonymization technique and value of the privacy parameter. In their solution, the target dataset is anonymized with every candidate anonymization method and every candidate privacy parameter value. The amount of privacy loss and utility loss in each anonymized dataset is measured and the pair $\langle$*privacy loss*, *utility loss*$\rangle$ is used to place the candidate on a two dimensional space. In this space the candidates that are higher in both privacy and utility loss compared to at least one other candidate are considered inefficient. The remaining superior candidates are incomparable and the authors do not offer any guideline on how to choose one setting from the winners set of configurations.

Loukides and Shao [LS08] study the privacy/utility tradeoff in settings where privacy protection level is partly [4] and utility level is fully negotiable and attempt to maximize a fitness function that considers both conflicting requirements. They combine clustering and partitioning to anonymize a dataset with the goal of optimizing weighted sum of the average information loss and privacy vulnerabilities over all equivalence classes. Information loss and privacy breach is measured based on s-diversity metrics. Their work aims at balancing the utility/privacy tradeoff at a *micro level* (within the algorithm) while the big decisions such as minimum size of equivalence classes (k) and weights of utility and privacy (within the optimization formula) are still external parameters of the problem and need to be fed to the algorithm as input arguments. The goal of this thesis is to provide an insight on setting the values of these external parameters by analyzing the negotiations that take place before running any anonymization algorithm.

Economic price theory [MY04] has also been used to find the optimum balance between

[4] Equivalence class sizes are still fixed.

privacy and utility [ZO10]. Zielenski and Olivier [ZO10] assume global recoding (see Section 2.2.1.2) as the sanitization method and use the concept of entropy to measure the importance of an attribute value to a data user and a privacy intruder. Based on how important each attribute is to a data user and an intruder, different weights are assigned to different attributes and the weighted sums of the entropies at each generalization level are used to measure data utility and data privacy. Utility requirements of a data user is then combined with privacy needs of a data collector in a single entity who is interested in maximizing both of these needs at the same time. The optimum value of this newly formed function is found by Lagrange Multipliers Method [Dix90]. Using economic price theory is very novel but the proposed privacy quantification metric requires detailed information about intruders' preferences over the attributes. These preferences mainly depend on adversaries' knowledge. The assumption of knowing the intruder and her back ground knowledge does not seem realistic. Moreover, at a higher level, to form the optimization function itself, weights must be assigned to the amount of tradeoff between utility and privacy. The authors do not provide any direction as to how these weights are determined. This thesis aims at addressing this problem at a macro level and analyzing how a data user and a data collector can come to an agreement on making a decision about the tradeoff of privacy/utility. In our solution the motivation behind providing a higher level of privacy protection is explained and quantified more realistically.

2.3 Privacy Policy Declaration

Another group of scholars [AHK+03, Mos05a, CDE+06, ACdVS08, BAB+09] realized that regardless of the sanitization method used, the generalized and/or perturbed data tables are still located inside the sphere of data providers' private matters. This group of researchers attempt to collect information in a manner that is aligned with well established privacy guidelines. A major reference for these privacy guidelines is explained in Hippocratic Databases

[AKSX02]. Agrawal *et al.*[AKSX02] extracted ten principles from privacy acts such as Organization of Economic Cooperation and Development (OECD) [fECOD02] and United States Privacy Act of 1974 [Rot00] and introduced them as a set of requirements for every database system which manages private information. Among the ten principles, the following four have drawn special attention from researchers in privacy policy developments:

- **Consent -** Data must be collected lawfully and with the consent of data providers.
- **Purpose Specification -** At the point of data collection the purpose of data collection and usage must be declared.
- **Limited Use -** Data practices on private data must be limited to those that fulfill the stated purpose(s).
- **Limited Retention -** Personal information must be removed from the repository once it fulfills all the purposes it has been collected for.

The principles entail declaring a policy and making sure that access to private information is permitted only if the data practices conform to the policy. Traditional access control policies are not fully equipped with the structures required for resolving privacy issues [MP05]. First, the notion of purpose and intricacies of verifying whether a subject really has the intentions specified in the purpose or not, does not exist in traditional access control methods. Secondly, privacy policies define an agreement between two or more parties and preferences of each party have to be considered. In the simplest form, these two parties are data providers and data collectors where the data collector specifies a policy and the consent of data providers must be sought before collecting their information. As a result, privacy policy languages and frameworks have been proposed which are mainly focusing on addressing privacy (not security) concerns [AHK$^+$03, Mos05a, CDE$^+$06, ACdVS08].

The main objective of each privacy policy declaration framework is to provide a uniform (and possibly machine readable) construct to explain data practices and dispute resolutions. Since these languages are not tightly tied to a specific infrastructure, implementation details are usually not included in the language description. Independent projects such as PRIME (PRivacy and Identity Management for Europe) [CLS11] or Adaptive Privacy Management System (APMS) [MP05] propose ideas to develop a system capable of implementing privacy policies. PRIME [CLS11] is a project aiming at defining the infrastructure for privacy enhancing identity management. It provides a comprehensive set of modules that allow for conducting online-transactions while protecting privacy of all parties. In APMS, Mont and Pearson [MP05] propose a system to enforce privacy policies using concepts of sticky policies and encryption. This system is designed to be implemented within PRIME's [CLS11] infrastructure. In the proposed system, at the point of data storage, each piece of information is encrypted with a symmetric key. The triple $\langle policy,\ key,\ encyptedData \rangle$ is then encrypted with the public key of a trusted third party. At the point of data retrieval, the trusted third party decrypts the triple and if the request conforms with the associated policy, decrypts *encryptedData* with the associated symmetric key and reveals the information.

In the following sections we summarize some privacy policy languages and frameworks independent of any specific implementation method.

2.3.1 eXtensible Access Control Markup Language (XACML)

XACML [Mos05a] is an XML-based language mainly developed to express access control policies at a concrete level. In XACML, description of data practices (containing access control and privacy expressions) is provided in terms of a "policy set" containing multiple policies. Each policy in a policy set can be potentially specified by a different department of an enterprise.

A policy is a set of rules and a conflict resolution algorithm (to resolve the situations where multiple rules applies to a case). Each rule in a policy explains who (subject) can

preform a specific action on a resource (piece of information) and what are the conditions and obligations (actions that must be preformed after or instead of the requested action). Recipients and resources of a rule can be specified using predicates over characteristics of recipients, resources, and the environment. These characteristics are known as attributes. For example, address of a data user and creation date of a file are attributes of recipients and resources, respectively.

In an attempt to extend XACML beyond a simple access control language, Privacy Policy Profile of XACML [Mos05b] adds the "purpose" attribute to explain the characteristics of resources and actions. The exact values of the purpose attribute are not pre-defined and each policy can have its own set of purpose values with its own interpretation. XACML allows for obligation specification. Obligations and conditions on environmental attributes can be used to define retention periods. There is not an explicit provision for defining multiple abstraction levels at which a resource is made available to the recipient. Such functionality can be achieved by defining *ad-hoc* functions.

Flexibility and multiple extension points are important features of the XACML language. However, this flexibility decreases the uniformity in policy specification vocabularies and therefore increases the complexity of a policy. A complex policy is less understandable to data providers and can be less appealing.

2.3.2 Platform for Privacy Preferences (P3P)

The Platform for Privacy Preferences (P3P) [CLM+02b, CDE+06] proposes an XML-based language to facilitate privacy policy specification and its association with web resources. The most recent version of P3P (P3P 1.1 [CDE+06]) is still under development but the previous version (P3P 1.0 [CLM+02b]) is already established as a W3C recommendation in April 2002.

A P3P policy is a set of statements along with information on assurances and dispute resolution procedures in a predefined XML format. For each piece of information collected

by a data collector, P3P requires a statement to describe privacy decisions involved in the information usage. Every statement in P3P has five mandatory elements of purpose, recipient, retention, data group, and consequence with an optional `non-identifiable` element. The data group explains to which pieces of information does the statement apply. The purpose element describes the purposes of accessing private information. Currently, P3P offers eleven predefined purposes and the 12th one simply allows the policy to specify other purposes. The recipient element specifies the legal entity to which collected information might be distributed. Six different tags are suggested for the recipient element. Every purpose and recipient can optionally include a required element to clarify whether providing the information for the corresponding purpose and/or recipient is mandatory or individuals can opt-in/opt-out. The retention element clarifies the duration for which the collected data is stored. P3P suggests five different tags to be used for retention. These tags establish connections between the retention period and purposes of data usage. All of these tags are ultimately translated into a specific time via a destruction time table. The optional `non-identifiable` keyword means either identifiable pieces of information are not used or are anonymized before being accessed.

Following such a standard protocol to specify a privacy policy offers the opportunity of implementing a user agent as a part of web browsers, web plug-ins or other data management tools to ensure that a data collector's privacy practices conform to the data provider's privacy preferences. To this aim a compatible privacy policy language called APPEL [CLM02a] is available to data providers to specify their privacy preferences. An agent can verify if privacy preferences of data providers specified in APPEL allow for providing personal information to a data collector based on her P3P privacy policy.

P3P specifications have an access element which is part of a policy construct (not a statement) and describes the type of access to all collected information. This element only specifies whether a website collects and discloses identifying information (data that can be

potentially used to identify an individual). Consequently, P3P offers a very limited set of options to describe at what level of abstraction each piece of information is provided to the recipients. Moreover, there are cases where none of the P3P vocabularies are precise enough to describe data practices. In these cases a non-standard vocabulary that most closely describes the practice must be used. Allowing for such arbitrary descriptions can increase flexibility but reduces uniformity in privacy policy specification.

Compared to XACML [Mos05a], P3P provides a higher level language which is fully independent of the underlying structure and implementation details. XACML provides almost the same features at a more concrete level and is harder to understand by regular data providers.

2.3.3 Privacy Aware Access Control System

Arganda *et al.*[ACdVS08] make a distinction between two types of privacy policies: release policies and data handling policies. A release policy expresses the regulations and preferences set by a data collector on how accesses to personal identifiable information is governed. A data handling policy is defined by the owner of personal identifiable information and explains privacy preferences of a data provider.

A release rule in a release policy has five elements: subject (or a set of subjects selected based on their characteristics), action, object (or a set of objects selected based on their characteristics), purpose, and conditions. A data handling rule has seven elements: recipients (specified by their identity, category, or attributes), action, personal identifiable information (specified as separate pieces of information or a category of personal information), purpose (a string organized in a hierarchy), conditions, provisions (actions that must be done before access is granted), and obligations. Retention can be specified as an obligation. Moreover, *ad-hoc* generic conditions can be used to express a limited number of restrictions on how precise the values of personal information are revealed.

In the proposed framework [ACdVS08], a data collector provides a template to data

providers to modify based on their preferences and allows for negotiation. This approach divides the power fairly between the data providers and data collectors. However, modifying data handling policy templates before each attempt to access an online service can be perceived as a large overhead to data providers.

2.3.4 Privacy Taxonomy

Barker *et al.*[BAB+09] propose a data privacy taxonomy as a tool to classify different perspectives on the notion of privacy. The taxonomy identifies four dimensions namely visibility, granularity, retention, and purpose. On each dimension, certain points are specified and ordered by the degree of privacy they offer.

The visibility axis addresses the same concept as P3P's recipient. However, the points identified along this axis seem more reasonable compared to P3P recipient tags. The privacy taxonomy identifies four points along the visibility axis: data owner (data provider), house (data collector), third-party, and all/world. All/world means the data would be made available (keeping other privacy aspects fixed) to anyone who asks for it.

The concept of granularity as a dimension in privacy was first introduced by Barker *et al.*[BAB+09]. Granularity specifies how specific and accurate a piece of information would appear in a query result. In the privacy taxonomy [BAB+09] three different points are identified along this axis namely existential, partial, and specific. Existential is defined as the level where queries can only check if a piece of information exists in the database or not. The partial level refers to the situations where data items or the result of the query are somehow perturbed to protect privacy of a data provider. This level generally refers to application of a sanitization mechanism (see Section 2.2) to private data. Finally, "specific" is the finest grained level at which the micro data is revealed without any data masking.

The retention dimension explains for how long a piece of information is stored in the system. The values along this axis can be exact points in date/time or can be expressed based on how many times an access to a piece of information is allowed.

Finally, the purpose dimension conveys the same concept as the purpose element in P3P but with completely different points along the axis.

Since the privacy taxonomy is only proposing a framework to define privacy (and not privacy policy), no concrete structure is offered to explain purpose or retention. Nonetheless, the captured four dimensions can serve as an appropriate skeleton to define privacy policies.

2.3.5 Does Privacy Policy Declaration Solve the Problem?

The privacy policy declaration approach is more aligned with privacy legislations. As a result, this approach seems to be more appealing in practice especially since the process of reading and agreeing to a policy can be facilitated by some software consoles (Such as P3P agent [CLM02a] and PRIME console [CLS11]). These consoles can assist data providers with the process of analyzing privacy policies and making a decision based on their preferences.

However, the privacy policy declaration approach is not sufficient on its own. This approach mainly revolves around restricting recipients and purposes of data usage with an overly optimistic hope that legitimate data recipients always have good intentions and do not disseminate/abuse private data that they have access to. In fact, once the data is revealed to a party then there is no way to prevent that party from misusing the information and the association of data and policies can be broken [MP05].

2.4 The Privacy Protection Approach Adapted in this Research

As explained in Sections 2.2 and 2.3 neither of the sanitization or privacy policy declaration approaches can solve the privacy protection problem on their own. The data sanitization approach mainly focuses on providing methods to manipulate (generalize, suppress, and/or perturb) a private data table in order to prevent misuse by authorized data users. But since none of the existing methods in this approach guarantee *absolute* privacy, the risk to privacy still exists. Therefore, without data providers' consent, a data collector who only uses a

data sanitization method to protect privacy still faces the risk of being charged by privacy lawsuits. The privacy policy declaration approach emphasizes on data providers' consent but masking data (providing access to private information at an abstract level) is mostly overlooked. As a result a malicious data recipient can abuse her privileges and use data for other purposes or distribute them to other parties.

To avoid shortcomings of the two classic approaches (data sanitization and privacy policy declaration), we believe that the essence of data providers' consent and providing private information at a less specific or perturbed level (either through using a sanitization method or a simple generalization) must co-exist in a privacy protection mechanism. Consequently, if a privacy policy declaration method is used, it needs to explicitly address the data granularity level and if a sanitization approach is used the data collector needs to have data providers' consent on the level of privacy protection (based on a privacy parameter value such as k in k-anonymity) before she publishes a private data table. In both cases a data collector must decide on some privacy settings (data retention and granularity in case of privacy policy declaration, and the value of a privacy parameter such as k in k-anonymity in case of data sanitization) and seek data providers' consent to the chosen settings before conducting any data practices. As a result, data providers' privacy preferences become an important decision factor in setting a balanced privacy protection levels.

2.5 Social Studies on Data Providers' Privacy Preferences

To analyze individuals' behaviours and concerns regarding privacy issues, social science studies have mainly followed two approaches. While some literature [KC05, HA96, SGB01, ACR99, CRA00] focuses on classifying individuals based on the level of their privacy concerns, others [SGB01, AG05, CA99, HAF05, Kob07, MG93, SMC93, Tur03] concentrate on discovering the impact of different socioeconomic and privacy related parameters on data providers' willingness to share.

2.5.1 Classification of Data Providers

Perhaps one of the most well-known classifications of individuals, based on their privacy concerns, are privacy indexes provided by Westin's studies [KC05]. Through a series of 30 privacy surveys Westin analyzed the trends in privacy concerns. In these surveys individuals are classified in three groups: privacy fundamentalists, privacy pragmatists, and privacy unconcerned. Privacy fundamentalists are described as the ones who "are generally distrustful of organizations that ask for personal information". People in this group tend to share their information just for the purpose of receiving the service they need since they are concerned about future abuse of their private data. Privacy pragmatists may tradeoff their privacy to receive some services. Finally, the unconcerned "are ready to forego privacy claims to secure consumer service benefits or public-order values". Based on a survey on Westin's studies [KC05], about 25%, 59%, and 16% of the public fall into the privacy fundamentalists, privacy pragmatists and unconcerned categories, respectively. In 2001, an observation-based empirical study [SGB01] revealed higher percentage for privacy fundamentalists (30%) and unconcerned. In this study, 21% of privacy pragmatists are further classified as identity concerned and the rest of them (26%) are referred to as profiling averse, who are not willing to disclose their profiling interests, hobbies, *etc.*. The problem with this classification is that the two newly introduced categories are not disjoint. In other words, an individual could be both identity concerned and profiling averse which makes the reported percentages unreliable. In a similar attempt Acquisti and Grossklags [AG05] subdivided the privacy pragmatists group into online identity concerned and offline identity concerned clusters. These two clusters also seem to have overlaps and therefore the reported percentages have to be seen skeptcally.

The classification approach assumes an underlying cost-benefit analysis for the privacy pragmatist group. However, due to lack of clear information about data flows which take place behind the scene [Tur03] this analysis is not deterministic and is context based. Grossklags and Acuisti [GA07] show this bounded rationality and uncertainty about the

value of privacy through a study on the gap between data providers' willingness to share and willingness to protect their personal information. These studies suggest that the degree of privacy awareness and concerns in individuals is not enough to fully determine the public's privacy behaviour. Instead, the actual behaviour needs to be analyzed based on the context and the various factors that affect data providers privacy decisions.

2.5.2 Influential Factors on Privacy Preferences

The second type of approach to study data providers' privacy behaviour is to identify the factors which can influence an individual's privacy decisions and study the correlation of these factors with her willingness to share personal information. Multiple categories of factors have been proven to have an effect on individuals' privacy preferences.

2.5.2.1 Effects of Trust and Privacy Knowledge

The more concerned a person is about privacy, the less she is willing to share her personal information. The effects of privacy concerns, knowledge about privacy and risks, and beliefs about the cost and effectiveness of the available privacy protection method are studied in a survey conducted by Acquisti and Grossklags [AG05]. The study also reveals the positive association of an individual's willingness to share and trust in the data collector. The significant impact of trust on individuals' privacy decisions is also discussed in the literature [Kob07], where 63% of the subjects of a survey announced lack of trust as the main reason why they refuse to provide personal information to a website. Culnan and Armstrong [CA99] also show the effects of the trust factor in terms of a relationship which was already established between a data provider and a data collector, and availability of promises about reliable and valid inferences from the collected information.

2.5.2.2 Effects of Socioeconomic Characteristics

Socioeconomic factors are also shown to have influence on privacy preferences. In 1993, linear and logistic regression models were used to study the effects of age, gender, educa-

tion, language, race, household characteristics and a few other factors on mail return in the 1990 U.S. census [SMC93]. Considering the correlation between privacy concerns and the probability of mailing back the census, the investigation showed that respondents who speak English as their primary language have higher privacy concerns compared to people who speak other languages. Moreover, members of the black non-Hispanic group show more concerns about privacy than Hispanic or white non-Hispanic respondents. Harris-Equifax surveys [CA99] also study privacy concerns among individuals based on demographic characteristics and life experiences. The results show that African Americans, Hispanics, Women, and less educated people are more concerned about their privacy. This result contradicts the outcomes of the census mail return survey [SMC93] regarding the Hispanic ethnicity. Acquisti and Grossklags [AG05] also verified the correlation of income level with privacy concerns: the lower the income level is the lower the privacy concerns are. Knowledge about the correlation between demographic characteristics and individuals' privacy behaviours is of high value to a data collector. If the distribution of age, race, education, income, *etc.* in potential data providers is known *a priori* the data collector can classify data providers in different groups and analyze each group's reaction to various privacy parameter settings separately. This type of analysis provides more reliable results compared to the case where none of the demographic characteristics is considered in anticipating data providers' privacy behaviour.

2.5.2.3 Effects of Data Sensitivity

Another determining factor on privacy preferences is the sensitivity of collected information. This sensitivity depends on the type of requested information and the associated value a person has for the data item. Kobsa [Kob07] provides a comprehensive overview on various studies aiming to analyze an individual's willingness to share, based on the type of requested information. According to this literature review, personal preferences, demographics, life style information are shown to be the types of information that individuals are more willing

to share. On the contrary, the same article claims that financial, purchase related, online behaviour, religion, political party identification, and occupation are more critical regarding privacy concerns. Acquisti and Grossklags [AG05] classify people based on their privacy concerns about providing different types of information. The study investigates an individual's behaviour when sharing data about online identity, offline identity, personal profile, professional profile, and sexual and political identity. The results suggest that offline identity is the most private information type and personal profile is the least. From a different perspective, Huberman *et al.*[HAF05] establish the correlation between sensitivity of the *value* of a requested information and data providers' willingness to share. This study shows that people are more reluctant to share information as the value negatively deviates from social norms. Finally, Spiekermann *et al.*[SGB01] argue that the difficulty of answering a question (providing information) also contributes to an individual's unwillingness to share the information.

2.5.2.4 Effects of Incentive

The effects of incentives offered as an exchange for collecting personal information has also been studied [Kob07, Tur03]. Regarding the cost-benefit analysis in privacy related decision making, the incentives are the benefits a person receives in return for providing a piece of information. Kobsa [Kob07] identifies two types of incentives which can affect an individual's disclosure behaviour: financial rewards and social adjustment benefits. Based on a study conducted in 2003 [Tur03], 21% of participants agree that they like to give information in exchange for offers, and 16% agreed that they will give out information if paid.

2.5.2.5 Effects of Privacy Settings

Privacy settings significantly impacts the decisions made by data providers. The most relevant study in this area is done by Milne and Gordion in a direct mail context [MG93]. The direct mail contract has four attributes: volume, targeting, compensation, and permission. The permission attribute corresponds to the purpose attribute in a privacy policy. The

probability of a customer agreeing with such a contract is determined by regressing (linear regression model) on the four attributes. Another study by Turow [Tur03] shows that 66% of people agree that the government needs to track users' online activities. This result suggests that if the purpose of information collection is specified as legal requirements, 66% of people share their personal information.

2.5.3 Challenges in Analyzing Privacy Preferences

The findings of the social science literature are heavily based on data collected from surveys. The empirical data collection methods are either inquiry-based or observation-based. On the one hand, the inquiry-based method usually suffers from the disadvantage of not portraying the actual disclosure behaviour [Kob07]. In fact, an extensive experiment [SGB01] has proven that in real life people act less concerned about their privacy than what they claim. On the other hand, it is practically impossible to investigate higher-level disclosure patterns through observation-based methods [Kob07]. Biased population and biased responses can also reduce the reliability of privacy surveys [Kob07]. Participants of privacy related surveys tend to be the ones with personal concerns about privacy which in turn results in a biased population. Moreover, Participants' responses to the survey can be biased since the subjects tend to answer and act in a socially desirable way (they don't always provide an honest answer). These factors can all contribute to skew the results of empirical studies and therefore one should be very careful in reporting and using the outcomes of such studies.

2.5.4 Privacy Preferences and Privacy Settings

To convince more data providers to provide personal information, a data collector must set privacy parameters based on privacy preferences of data providers. The knowledge offered by social studies on classes of data providers and the population of each class can guide a data collector on how to target different groups of data providers differently. Moreover, as we see in the next chapter, having access to a mathematical model (such as a regression model)

that explains the effects of privacy parameters and incentive on data providers' behaviour is an essential element to find a balanced setting for privacy parameters (and incentive).

2.6 Summary

Collection and storage of personal information raises the ever-increasing concern of risks to respondents' privacy. Data sanitization and privacy policy declaration are two common approaches to address such concerns. We briefly explained some seminal data sanitization methods such as k-anonymity, l-diversity, t-closeness, and differential privacy in Sections 2.2.1 and 2.2.2. Each of these sanitization methods is associated with some parameters that establish the tradeoff of privacy versus utility. Section 2.2.3 summarizes some of the existing privacy/utility metrics and methods to assess the tradeoff. Since absolute privacy is only achievable at the cost of zero utility, we postulate that in data sanitization methods data providers' consent must be sought before publishing a private data table. The key element in privacy policy declaration approach is data providers' consent to a policy before data collection. We discussed XACML, P3P, Privacy aware access control systems, and privacy taxonomy as existing frameworks to explain privacy policies in Section 2.3. These methods rarely address the problem of protecting data against authorized but malicious recipients. Therefore, we believe that a granularity parameter must exist in these methods to explain different levels of data revelation. The value for a granularity parameter must be chosen such that a balanced privacy/utility tradeoff is achieved. With this perspective, we recognize data providers' consent and data masking as two key ingredients of any privacy protection mechanism. Since data providers' consent must exist before data collection, their privacy behavior can influence privacy setting decisions. In Section 2.5 we provided an overview on data providers' classification based on their privacy concerns and factors that affect privacy decisions of an individual.

In this chapter we reviewed the essential components of privacy protection mechanisms

as one of the perquisites for developing a game model to address the challenge of setting balancing values for privacy parameters. In Chapter 3, we provide a brief overview on game theory and discuss some literature on "game theory applications in privacy" (since they are better understood after introducing some basic definitions in game theory). With a background on both privacy protection mechanisms and game theory, in Chapter 4, we formalize the game of setting privacy parameters (for both sanitization and privacy policy declaration approaches) and provide a generic guideline on how to solve the game.

Chapter 3

Game Theory and Privacy

Game theory is a mathematical approach to study interdependencies between individual's decisions in strategic situations (games). It has been successfully used to better understand the competing behavior of firms in business, participants in auctions and bargaining, jury members in reaching a verdict, political parties in collecting votes, animals in fighting over prey, agents in multi-agent systems, *etc.*[Osb03]. This chapter reviews essential topics in game theory and discusses related interdisciplinary applications of game theory in data privacy. Related game theory concepts and solutions are provided in Section 3.1 and a summary of game theory applications in addressing privacy issues is given in Section 3.2.

3.1 Game Theory Preliminaries

Game theory offers a formal mechanism to study strategic situations [Dut99]. In a strategic situation each decision maker is affected by the consequences of not only her own decisions but also the choices made by other decision makers [Dut99]. In this sense, if a decision maker's choice only affects herself (and no one else)[1] then the situation is not considered to be a strategic situation [Dut99].

3.1.1 Specifications of a Game

Game theory studies the interplay of a set of *rational* decision makers who make choices strategically by considering the effects of decisions that other decision makers might make [Dut99].

[1]Patric K. Dutta [Dut99] also notes that if there are too many decision makers so that it is infeasible to keep track of the consequences of each person's action then the situation cannot be explained as a game.

The description of any game must provide a clear specification for the following three components:

- Who are the **players** of the game?
- What **actions** are available to each player?
- For each player, what are her **preferences** over possible outcomes of a game?

As we shall see in Section 3.1.2, the description of a game may also include temporal information on a player's turn to make a decision. If such information is absent then by default the assumption is that players make decisions simultaneously.

Players are the decision makers who strategically interact in a game. In any given situation during the game, a set of alternative actions are available to a player to choose from. Players are assumed to be *rational*, which means every player makes decisions that benefit her the most considering the effects of other players' choices [Dut99]. Rationality of a player implies that she chooses the most beneficial action according to *her own* preferences. Therefore, a player can be rational even if her preferences do not seem to be rational [Osb03].

A combination of choices made by all players during a game represents a play of the game and corresponds to a possible outcome of the game. For each two possible outcomes of a game, each player is assumed to clearly know which one she prefers or if both outcomes are equally desirable to her. Moreover, players' preferences are assumed to be consistent in the sense that if a player prefers outcome a to outcome b and prefers outcome b to outcome c then she must prefer outcome a to c [Osb03]. A player's preferences over all possible outcomes of a game are represented by a *payoff function* (also known as a utility function). A higher payoff for an outcome represents the player's higher preference for that result. We use the notation $U_i(s)$ to denote how much player i stands to gain for an outcome that corresponds to a specific play s of the game. As we will see in the following, s is a profile of *strategies* that contains a strategy for each player.

Example 3.1.1. Prisoners' dilemma

Prisoners' dilemma is a classic example of a game analyzed in game theory. Players of this game are two prisoners Alice and Bob who are hauled for a suspected crime. The police interrogates the two prisoners separately and offers them the same deal. If one of the players "confesses" and implicates the others he/she may go free while the other goes to jail for 15 years. If both confess, then they both go in for five years, and if neither of them confesses then they are sentenced to one year in jail. In this example, both players have the same set of available actions: $AC_{Alice} = AC_{Bob} = \{Confess,\ No\ Confess\}$.

We represent each possible outcome of the game as a vector $s = \langle ac_{Alice},\ ac_{Bob}\rangle$ *where* ac_{Alice} *and* ac_{Bob} *represent actions chosen by Alice and Bob, respectively. Consequently, we can define Alice's payoff function as follows:*

$U_{Alice}(\langle Confess,\ Confess\rangle) = -5$

$U_{Alice}(\langle Confess,\ Not\ Confess\rangle) = 0$

$U_{Alice}(\langle Not\ Confess,\ Confess\rangle) = -15$

$U_{Alice}(\langle Not\ Confess,\ Not\ Confess\rangle) = -1$

Bob's payoff function can be explained similarly.

A player's strategy is a plan that explains which action she chooses at each point of decision. In other words, if a player appoints an agent to play on her behalf and gives the agent her strategy then the agent knows exactly how to "carry out her wishes, whatever actions the other players take" [Osb03]. Let AC_i^j denote the set of actions available to player i at point j of the game (where it is player i 's turn to move). If there are m different points during the game at which player i makes a decision then each strategy of player i is a vector $s_i = \langle ac_i^1, ac_i^2, ..., ac_i^m\rangle$ which is interpreted as player i taking action ac_i^k at point k in the game. The set of all available strategies to player i is:

$$S_I = AC_i^1 \times AC_i^2 \times ... \times AC_i^m$$

The set S_I contains all possible *pure* strategies of player i.

Besides pure strategies, the concept of *mixed* strategies is sometimes used to solve a game. If $S_I = \{s^1, s^2, ..., s^M\}$ is the set of pure strategies available to player i, a mixed strategy for the player is defined in terms of a probability distribution $(p^1, p^2, ...p^M)$ that assigns probability p^j to strategy s^j. The *support* of a mixed strategy is the set of strategies that have an associated probability of greater than zero. When mixed strategies are used, players' preferences over lotteries are denoted by the expected value of the specified payoff functions [Osb03].

Example 3.1.2. *In the game of Prisoners' dilemma (see 3.1.1), each player moves only once during the game. Therefore, for each player, the set of their available actions is the same as the set of their available strategies. For instance, we have:* $S_{Alice} = \{Confess,\ Not\ Confess\}$. *The prisoners can use a mixed strategy by choosing each of their pure strategies with certain probability. For example, Alice can adapt a mixed strategy* $(\frac{1}{2}, \frac{1}{2})$ *in which she chooses to* $Confess$ *or* $Not\ Confess$ *with the same probability* $\frac{1}{2}$.

The usefulness of the mixed strategy concept is not very clear to researchers. The two assumptions that (1) people randomize to make decisions and (2) players can develop a correct believe about the exact probability distribution of their opponents' strategy are not very realistic [Dut99]. However, the mixed strategy concept can be used to model those situations where players are not sure about who they are playing against [Har73, Har67]. In Section 3.1.4 this interpretation of mixed strategy is discussed.

Information about a game's components are often assumed to be a *common knowledge* of the game [Dut99]. This means that every player has the knowledge about the game's specifications (players, available actions to each player, payoffs to each player, and game rules), the players know that every player has knowledge about the game's specifications, the players know that every player knows that every player knows the game's specifications, and so on. In Section 3.1.4 we see how this assumption can be relaxed in games of incomplete information.

3.1.2 Strategic vs. Extensive Form

Every game can be represented in either *strategic* (also known as normal) or *extensive* form. The strategic representation of a game explains the players, their available actions and payoffs without clearly specifying when each player moves. The default is to assume that they all move simultaneously. If there are only two players in the game, the strategic form represents the game as a table; each row in the table corresponds to one of the actions available to player 1 and each column stands for one of the available actions to player 2. At the intersection of each row and column, the payoffs to each player for that specific play of the game is written in the cell.

Example 3.1.3. *Consider the game of Prisoners' dilemma (see 3.1.1). The following table provides a strategic representation of the game:*

Table 3.1: Strategic representation for the game of Prisoners' dilemma

Alice / Bob	*Confess*	*Not Confess*
Confess	*-5, -5*	*0, -15*
Not Confess	*-15, 0*	*-1, -1*

A strategic form is usually used to represents a "one-time simultaneous move" game [Dut99]. However, if players have more than one decision point and there are certain orders for their turn to move, then each blueprint of their strategies during the whole game can be treated as single complex action and we can assume that they pick their strategies simultaneously. Therefore, such games can still have a strategic representation.

The extensive form adds a temporal dimension to the description of a game by specifying the order in which players play. Compared to the strategic form, in an extensive form two new components must be specified [Osb03]:

- The set of all sequence of actions that can occur in the game.
- The order of players' turns to move.

The specific ordering and the details of a game are represented as a tree.

A tree starts from the root node where a single player has to make a choice from the available actions. Each branch that emerges from the root node represents a possible action that the first player can choose. The end of each branch is either a *decision node* or a *terminal node*. A decision node is a point in the tree where a single player has to make a decision (each possible decision is represented by a branch emanating from this node.). A terminal node represents an end point of the game. Subsequently, the children of every decision node can be a combination of other decision nodes and terminal nodes. A label at each node clarifies the player that must make a decision at the node and each branch has a label to specify the action it represents.

A possible sequence of actions that starts from the root and ends at a node in the tree is called a *history*. If a history ends at a terminal node then it is referred to as a *terminal history*. We use the notation $h = (a_1, a_2, ...a_n)$ to show a history where the first player plays action a_1 at the root, the next player (based on the description of the tree) chooses the action a_2 and so on. Notice that a terminal history corresponds to a play of the game. Therefore, players' preferences over each outcome of the game are shown at each terminal node. If all of the terminal histories of a game are composed of a finite number of actions and each player at each point has finitely many actions to choose from, then the game is a *finite* game [Osb03].

Example 3.1.4. Entry Game

Another classic game studied in game theory is known as the Entry game. In this game, firm A is an incumbent monopolist. Firm B is a challenger who has the opportunity to enter into the industry currently occupied by A. If B enters, then firm A (the incumbent) may either acquiesce or fight.

Figure 3.1 represents this strategic situation as a tree. At each terminal node, we list players' payoffs (player A's payoff followed by player B's payoff) for the corresponding outcome

of the game.

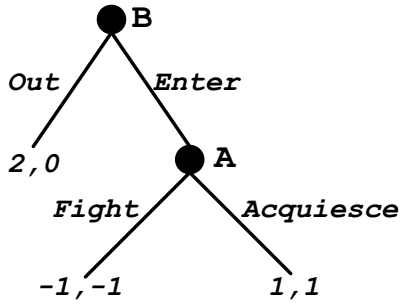


Figure 3.1: The Entry game

This tree has four possible histories: $h1 = (Out)$, $h2 = (Enter)$, $h3 = (Enter,\ Fight)$, *and* $h4 = (Enter,\ Acquiesce)$. *Except for history* $h2$, *all of the other histories are terminal histories.*

To represent simultaneous moves in a tree the concept of *Information Set* is used. An information set is a set of decision nodes that the corresponding player cannot distinguish between [Dut99]. A decision node is represented as an oval that circles around the indistinguishable decision nodes. In a game of *perfect information* at each point a player exactly knows the choices that are already made by previous players. Therefore in games of perfect information every information set has exactly one decision node (no simultaneous moves allowed). A game that has at least one information set with more than one decision node is a game of *imperfect information.*

As explained earlier, both strategic and extensive representations can be used to model a game[2]. However, it is more natural to use the strategic form for games with one-shot simultaneous moves and the extensive form for games with a certain order for players' turns to move.

Example 3.1.5. *Consider the game of Prisoners' dilemma (see 3.1.1). An extensive repre-*

[2]Each extensive form game has a unique strategic form equivalent to the original game. A game in strategic form can be represented with at least two different extensive forms [Osb03].

sentation of this game has an information set composed of two decision nodes for one of the players. This representation is illustrated in Figure 3.2.

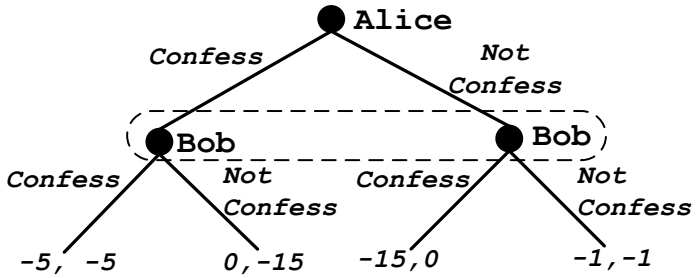


Figure 3.2: The game of Prisoners' dilemma (extensive representation)

The strategic representation of the Entry game (see 3.1.4) is shown in Table 3.2.

Table 3.2: Strategic representation for the Entry game

A / B	*Out*	*Enter*
Fight	*2, 0*	*-1, -1*
Acquiesce	*2, 0*	*1, 1*

3.1.3 Nash Equilibrium

In a game with n players, a strategy profile $s = \langle s_1, s_2, ..., s_n \rangle$ is a vector of strategies (one for each player) and determines a specific play of the game. For the rest of our discussion we use the notation $s_i \in S_I$ to denote a strategy of player i selected from the set S_I of all of her possible strategies. We also use use the symbol $s_{-i} = \langle s_1, s_2, ...s_{i-1}, s_{i+1}, ..., s_n \rangle \in S_{-I}$ to represent a combination of strategies for all players except for player i. Consequently, the notation $s = \langle s_i, s_{-i} \rangle$ is equivalent to $s = \langle s_1, s_2, ..., s_n \rangle$.

A strategy s_i for player i is her *dominant strategy* if regardless of what other players decide, s_i always provides a higher payoff to player i compared to all of her other strategies. If every player has a dominant strategy, then the combination of players' dominant strategies represent the *dominant strategy solution* to the game.

The *Iterated Elimination of Dominated Strategies (IEDS)* is another solution concept that suggests iteratively eliminating strategies that are already dominated by other strategies until only one strategy profile is left for the whole game. If such a unique strategy profile is found by this method the game is called *dominance solvable* [Dut99]. However, not all games are dominance solvable.

In solving strategic games, the most widely used solution concept is known as *Nash Equilibrium*. The key component in a Nash equilibrium is the concept of players' *best response* to a belief.

Best Response Let $s_{-i} \in S_{-I}$ denote a strategy choice of all players other than player i. If player i believes that other players choose their actions based on what s_{-i} suggests then her best response (to such a belief) is a strategy s_i such that:

$$\forall s'_i \;\; U_i(\langle s_i, s_{-i} \rangle) \geq \; U_i(\langle s'_i, s_{-i} \rangle) \tag{3.1}$$

If s_i is a mixed strategy (and not a pure strategy), every pure strategy in its support (see Section 3.1.1) must be a best response to the same belief [Dut99].

A Nash equilibrium is a strategy profile where each player is playing her best response to the other players' strategies.

Nash Equilibrium A strategy profile $s^* = \langle s^*_1, s^*_2, ..., s^*_n \rangle$ is a Nash equilibrium if:

$$\forall i \in \{1, ..., n\} \wedge \; \forall s_i \in S_I \;\; U_i(\langle s^*_i, s^*_{-i} \rangle) \geq \; U_i(\langle s_i, s^*_{-i} \rangle) \tag{3.2}$$

A Nash equilibrium is *strict* if for all $s_i \neq s^*_i$ the utility $U_i(\langle s^*_i, s^*_{-i} \rangle)$ in Equation 3.2 is strictly better than $U_i(\langle s_i, s^*_{-i} \rangle)$ (the comparison between the two utilities is based on the $>$ relation and not $\geq$).

If s^* is a mixed strategy Nash equilibrium then every pure strategy in the support of the mixed strategy s^*_i (for player i) yield the same expected payoff and no other pure strategy results in a higher expected payoff to player i (assuming that other players' strategies are

what s^* suggests). It has been shown that every strategic game in which players have finitely many strategies to choose from has at least one Nash equilibrium in mixed strategies [Osb03].

Example 3.1.6. *In the game of Prisoners' dilemma (see 3.1.1), the strategy profile* $s = \langle Confess,\ Confess\rangle$ *is a Nash equilibrium. To see the reason, notice that if Alice believes that Bob is going to* $Confess$ *then her best response to this belief is to* $Confess$ *because* $U_{Alice}(\langle Confess,\ Confess\rangle) > U_{Alice}(\langle No\ Confess,\ Confess\rangle)$. *Similarly, if Bob believes that Alice is going to* $Confess$*, his best response is to* $Confess$ *as well.*

Based on the definitions of Nash equilibrium and best response, a simple method to find a Nash equilibrium can be as follows:

1. For each player i find the set B_i of strategy profiles $s = \langle s_i, s_{-i}\rangle$ such that for each $s \in B_i$ strategy s_i represents player i's best response to the combination of other players' strategies s_{-i}.

2. The intersection of all B_i's is the set of Nash equilibria of the game. In other words:

$$Nash\ Equilibria = \bigcap_{i\in\{1,\ldots,n\}} B_i \tag{3.3}$$

Dutta [Dut99] introduces Nash equilibrium as a *stable* prescription for play in the sense that if such a strategy profile is suggested to the players then no single player has the incentive to make a different decision. He further notes that Nash equilibria are the only plays of the game on which all players can reach a common agreement if they are allowed to have pre-game communications. Nash equilibrium points have also been recognized as *steady states* of a game where each player is randomly chosen from a pre-defined population in each play of the game [Osb03]. With this interpretation, a Nash equilibrium represents a "social norm" in the sense that if every one follows the norm, no single player wishes to deviate from the norm [Osb03]. This is why through the rest of this thesis we sometime use

the terms stable or steady privacy parameter values to refer to those values found in the Nash equilibria of the game.

Nash equilibrium is a more general solution concept than dominant strategy and IEDS. Every solution achieved by the dominant strategy or IEDS method in a game is also a Nash equilibrium (the reverse is not necessarily true).

The concept of Nash equilibrium is incapable of considering the sequential structure of a game (if such a structure exists) [Osb03]. Therefore, in games with a certain order on players' turns to move a Nash equilibrium might not be robust. To model a robust steady state in extensive form games (see Section 3.1.2) a solution concept stronger than Nash equilibrium is usually used. This concept is known as *subgame perfect equilibrium*. To understand the premise of a subgame perfect equilibrium, we need to explain the concept of *subgame* in extensive form games:

Subgame In a game tree, a subgame is a subtree that starts from a *single* decision node and contains all of its successor decision and terminal nodes. If one of the successors is a decision node contained in an information set then all of the decision nodes in that information set must also be included in the subgame.

A subgame perfect equilibrium induces a Nash equilibrium in *every* subgame of the game tree. Therefore, for all those parts where players move simultaneously in the tree (games of imperfect information), the group's decision must be a Nash equilibrium. It can be easily seen that every subgame perfect equilibrium is also a Nash equilibrium of the game. It has been proven that every finite game (games that eventually terminate and players have finitely many available strategies) has at least one subgame perfect equilibrium [Osb03].

A process known as *backward induction* is used to find the subgame perfect equilibria of extensive form games with perfect information (see Section 3.1.2). If the game is not a perfect information game (simultaneous moves are allowed) then an extension of the backward induction method can be used. For simplicity, we refer to the latter as the *generic* backward

induction method. The key idea in (generic) backward induction is *sequential rationality*. Sequential rationality means "whenever a player needs to make a decision, she deduces how the subsequent players react to each of her actions and then chooses the most profitable action. Future players reason the same way" [Dut99].

The method of generic backward induction starts with finding the Nash equilibria in all of the smallest subgames of the tree (the ones that end only in terminal nodes). The next step is to back out to subgames that precede the smallest subgames and find the Nash equilibria of these subgames assuming that the outcome of the smallest subgames are the Nash equilibria that we have already found in the first step. The same procedure applies to subgames that immediately precede the subgames solved in the second step and so on. We fold back the game tree one step at a time until we reach the root node. At this point we are only left with strategy profiles that are the subgame perfect equilibria of the game. If we have a game of perfect information then at each subtree instead of finding the Nash equilibria we find the best response of the player whose turn it is to move (backward induction method). In these games, the backward induction method provides the same solution as IEDS in the strategic representation of the game [Dut99].

Example 3.1.7. *The Entry game (see 3.1.4) has two subgames. The first subgame is the whole game itself, and the second subgame is the subtree that starts after history* $h2 = (Enter)$. *To solve this game with backward induction, we start from the smallest subgame which is the subgame starting after history* $h2 = (Enter)$. *Since this is a game of perfect information, we find firm A's best response to firm B's choice of entering the industry. Based on A's payoffs in this subgame, firm A's best response is to Acquiesce. The next step is to find the best response of firm B, considering the fact that "firm A chooses to Acquiesce if firm B enters". By comparing firm B's payoffs for terminal histories* $h1 = (Out)$ *and* $h4 = (Enter,\ Acquiesce)$, *the best response of firm B is to enter into the industry and the game's subgame perfect equilibrium is* $s = \langle Enter,\ Acquiesce \rangle$.

3.1.4 Games with Incomplete Information

Games with incomplete information model those situations where players are uncertain about some characteristics of other players (such as their payoff, their available actions, their information about the game, and their beliefs) [Dut99]. In the most generic form, every player assigns a subjective joint probability distributions to all unknown variables and attempts to maximize her expected payoff based on the probability distribution that she believes is correct.

Harsanyi [Har67] showed that regardless of the nature of unknown variables, they can be expressed as players' ignorance on each others' payoff functions. Therefore in games of incomplete information uncertainty is modeled by considering multiple types t_i^1, ..., t_i^m for player i.

In a series of three consecutive articles [Har67, Har68b, Har68a], Harsanyi proposed the *Bayes-Nash equilibrium* to solve games with incomplete information. His approach is limited to *consistent games* where each player's beliefs on how other players assign conditional probability distributions to possible events, can be drawn based on a single basic probability distribution. The assumption of having a consistent game model is always the default in analyzing games with incomplete information unless there is a good reason to assume otherwise [Har68a]. To solve consistent games with incomplete information, Harsanyi [Har67, Har68b, Har68a] suggests finding the *Bayes-equivalent* of the original game. The equivalent game is a game of complete but imperfect information which incorporates "nature" as an artificial player. Nature plays at a *chance node*. The branches that emanate from a chance node have associated probabilities and represent a possible combination of types for each player in the game. The probability associated with a branch is equal to the product of the probabilities of having the corresponding types for players. For example, if there are two players a and b in the game the branch that represents the case of type t_a^2 for player a and t_b^3 for player b has probability $p_a^2 \times p_b^3$ where p_i^j denotes the probability of having

type j of player i. The information sets in the tree are designed according to how much each player knows (about other players' types). The payoffs at the terminal nodes are replaced with the expected payoffs (considering the probability of having a certain combination of types for the players). Once the game of imperfect information is created, Nash equilibria of the game represent the Bayes-Nash equilibria.

Example 3.1.8. *Consider a new version of the Entry game (see 3.1.4) where the incumbent (firm A) is "tough" with probability* $1-p$ *and is normal with probability* p*. Depending on the type of incumbent, the payoffs of the game are different. This situation is shown in Figure 3.3.*

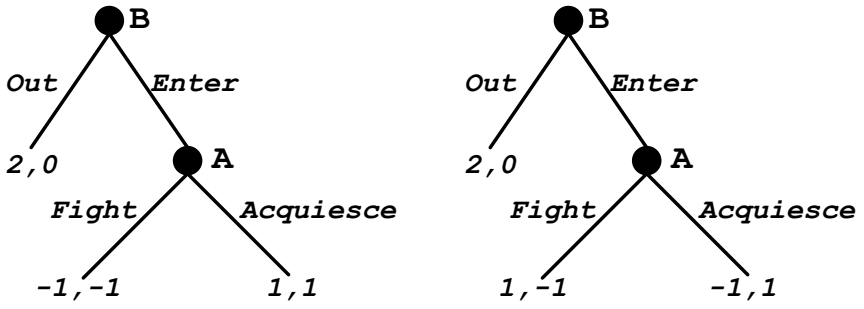


Figure 3.3: Modified version of the Entry game with (a) normal incumbent, and (b) tough incumbent.

The Bayes-equilvalent of this game has nature as the first player and is illustrated in Figure 3.4.

If player x is informed about her own type and does not have uncertainty about characteristics of other players she simply chooses her best response (different types of this player might have different best responses). The other players of the game who are not certain about player x's type realize that each type of player x plays a certain strategy in a Nash equilibrium. They are also assumed to know the probability by which player x is of each certain type. Therefore, other players observe the situation as if player x is using a mixed strategy [Har73, Har67]. In this mixed strategy, the probability weights in the mixture are

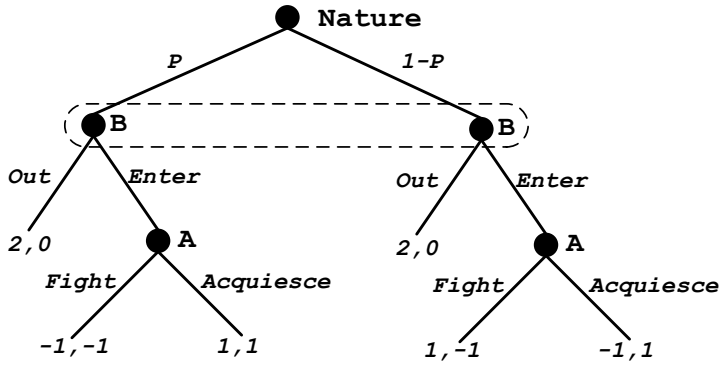


Figure 3.4: The Bayes-equilvalent of the Entry game.

the same as the probabilities associated with different types for player x. In other words, let strategy s_x^t be the best response of type t of player x and let p_x^t denote the odds of player x being of type t. In this situation, player x uses strategy s_x^t with probability p_x^t in her mixed strategy[3]. Consequently other players choose their best response to this mixed strategy. For example, suppose one of the players (mystery player) can take two different types: t^1 with probability p^1 and t^2 with probability p^2. Type t^1 has a best strategy s^1 and type t^2 has a best strategy s^2. To the other players it appears as if the mystery player is mixing between strategies s^1 and s^2 with probabilities p^1 and p^2. As we see in Section 4.2 this situation is the case in our model and we use this interpretation of incomplete information when we model the data providers' best response in Section 4.3.1.

3.2 Game Theory Applications in Data Privacy

In this project we use game theory to investigate balancing(stable) settings for privacy parameters by adopting a broader perspective on effective components. Game theory has been successfully applied to study various privacy-oriented challenges from different angles such as legal and economic aspects [And06, BK07, Var09], private multi-party computation [KDL07, MR11, MN12, Kat08], providing location privacy [GD08, FMHP09], privacy pro-

[3]If other types of player x also have the same strategy s_x^t as their best response then the odds of getting those types are added to the weight of strategy s_x^t.

tection strategies in social networks [SSP09, BP09], the role of trust in distributed privacy protection settings [KPM11], the risk involved in providing truthful information [RS10], reward allocation for private information [KPR01], providing a dynamic privacy model [Pre06], sharing private information between two firms [CP01], and establishing a definition for "green information communication technology services" [DF09]. In this section, we briefly summarize these works.

Various legal and economic intricacies of privacy have been studied using game theory and other microeconomic principles. With regard to the conflicts between privacy regulations in United States and Europe, Horace and Anderson [And06] use strategic and extensive form games to find equilibrium solutions for data transfers between firms in Europe and U.S.. They show that as long as the two players (U.S. and E.U.) are concerned about future interactions (the extensive version of the game) they choose to cooperate and this choice leads to better outcomes for both players compared to the strategic form game. Varian [Var09] also shows the interdependencies of privacy and price discrimination in economic transactions. He claims that there are contexts in which data providers voluntarily disclose private information in order to benefit from price discrimination. Therefore, he concludes that a rigid global legislation for private data practices might fail to provide flexibility to allow for such useful policies. Through another elaborate study [BK07], Böhme and Koble analyze the consequences of pseudonymous interactions between customers and service providers on price discrimination, social welfare, and profits to service providers. This study shows that in a self-regulated privacy adoption environment, strategic decisions of customers can lead to a higher increase in service providers' benefits as opposed to a government regulated environment.

Interests in combining the fields of game theory and multi-party computation has recently grown in two different directions [Kat08]:

(a) Some scholars focus on the application of multi-party computation in achieving certain

types of equilibria in games.

(b) Other studies apply game-theory to model and design more reliable protocols for multi-party computations with "self-interested" parties.

The concept of "correlated equilibria" in game theory is only achievable through the use of a trusted "mediator". The first group of literature in this area, propose "fair secure" [LMPS04] and "secure-with-abort" [DHR00] multi-party computation protocols to implement an initial cheap talk[4] phase that recommends the correlated equilibrium in absence of a mediator. The second group of literature in this field, use game theory to study the assumption of having honest parties and the incentives for parties to deviate from a protocol. Kargupta *et al.*[KDL07] use a strategic game to explain the possibility of collusion in multi-party secure sum computation. In their model, each node is a player whose payoff depends on the amount of computations, messages it receives, messages it sends, and additional information it gains by forming a collusion with k other nodes in the network. They show that in Nash equilibria of the game, nodes will have tendencies to collude and therefore the assumption of semi-honest parties is not realistic. They further design a mechanism based on the concept of *cheap talk* where honest nodes penalize a suspicious collusion by increasing the amount of computation costs. The designed mechanism is claimed to have an equilibrium where all nodes follow the protocol and behave honestly.

Miyaji and Rahman [MR11] use game theory concepts to design a protocol for privacy preserving multi-party set intersection. Their proposed protocol does not rely on assumptions of having semi-honest or malicious parties. Instead, they design a protocol such that when parties act rational (they benefit from a malicious action as long as they are not caught) the game has a strict Nash equilibrium in which all parties follow the protocol.

In a different approach, More and Naumov [MN12] use game-based reasoning to formally

[4]Communication is cheap in the sense that its cost is negligible and players can lie in their statements [Kat08].

prove privacy claims of a multi-party communication protocol. To verify privacy claims of a protocol, they suggest to design a game of perfect information with parties (or channels) as the players. Players are then split into two disjoint groups. Each group has the goal of finding a legal value (according to the protocol and independence criteria of the game) for part of the message sent over a channel. If there exists a strategy for each player to reach her goal, then the authors conclude that the protocol fulfils the claimed privacy property. They further use the informal game-based reasoning to define a formal logical system called calculus of cooperation to prove privacy properties of multi-party communication protocols.

Game theoretical concepts have also been used to study strategic situations in protecting location privacy [GD08, FMHP09]. Gianini and Daminani [GD08] model the problem of protecting location privacy by a cloaking technique as a zero-sum game between an attacker and the location anonymizer. They recognize that non-neutral landscape information can help an attacker to perform more accurate inferences over the obfuscated location data. By solving the game, they show that the anonymizer has an equilibrium strategy that cancels the effect of auxiliary information about positional constraints. Adopting a different technique for protecting location privacy, Freudiger *et al.*[FMHP09] model the problem with *mix zones* technique as strategic games of complete and incomplete information. In the mix zones approach, a mobile node can announce a zone as a mix zone by broadcasting a pseudonym change request to its neighbours. The neighbours decide whether to cooperate or not based on their current level of privacy and cost of pseudonym change. The privacy protection level of a mobile node decreases as the elapsed time since its last pseudonym change increases. In the incomplete information version of the game, players are uncertain about the time and privacy gain of other players' last pseudonym change. The authors show that in the game's equilibria, having a larger number of mobile nodes encourages players to act more selfishly and not cooperative. Moreover, as the cost of pseudonym change increases, players are shown to have more incentive to cooperate.

The widespread usage of social networks and their special structure has raised a variety of new challenges in privacy protection. Researchers [BP09, SSP09] have studied some of these issues using game theory. After a thorough analysis of hundreds of privacy features in the 45 most popular social networking websites, Bonneau and Preibusch [BP09] introduce a game to explain the observed trends in *privacy communication* behaviors of operators in such networks. The privacy communication behavior is explained as how much unnecessary data is collected, how advanced privacy control features are, and how clear the privacy policy of a social network website is explained. Their game model has two players: a network operator (data collector) and a data provider. They assume that a data provider can be either *non-fundamentalist* or *privacy fundamentalist*. In their model, the explained sets of available actions to each player are inaccurate and payoffs to players are not formally defined. Nevertheless, preferences of the network operator lead to strategies in which she attempts to attract more non-fundamentalist data providers and encourage them to use no privacy protection while discouraging privacy fundamentalists from joining the network and reducing their privacy criticism. They have considered the *privacy salience* phenomenon [JAL09] for non-fundamentalists' reaction to privacy promotions. According to this phenomenon, a non-fundamentalist is less likely to provide her personal information if the network operator talks about privacy practices (even if the operator is promoting safe privacy practices). As a result, the best response of a network operator is explained as minimizing privacy awareness of non-fundamentalists by not making the network's privacy policy readily available (partially concealing it), refusing to implement P3P [CLM+02b], and choosing the defaults of privacy settings to "no-privacy". To avoid the complains from privacy fundamentalists (which may discourage the non-fundamentalist population) the network operator relies on the fact that fundamentalists actively investigate the website for privacy features and policies. Consequently, Bonneau and Preibusch [BP09] rationalize why in most successful social networks a network operator implements complicated privacy protection features and pre-

pare a clear privacy policy statement without openly promoting these features (to minimize privacy salience).

In the context of privacy management within social network environments, Squicciarini *et al.*[SSP09] use game theory to design a mechanism that promotes co-ownership and truthfulness (in explaining privacy preferences) as the game's equilibrium. Their proposed model rewards and penalizes players by numeric credit units. A player scores credits for announcing other qualified players as co-owners of a shared resource (such as a family photo). To collectively set privacy preferences for a shared resource, every co-owner is asked to provide a numeric valuation for every possible privacy setting. The privacy setting with the highest number of votes is chosen for the shared resource and each co-owner is penalized based on how much her votes affected the final decision. The authors claim that with this system of reward/penalty assignments, the game has a Nash equilibrium where each co-owner truthfully announces her preferences. Although the model theoretically provides evidence of such a desired behavior, its applicability is not clear since a virtual numeric credit might not seem an effective incentive to members of a social network site.

The effects of trust and risk factors in privacy protection and private data provision have been analyzed via game theoretical models. Kamhoua *et al.*[KPM11] focus on strategic situations where protecting privacy or security of a shared resource requires the cooperation of multiple parties who manage the resource. More specifically, to protect a resource all (or a minimal subset) of parties in charge must decide to undergo the cost of privacy/security protection. If one of the parties (or more than the minimum number of required honest parties) decides to deviate then the resource is considered unprotected against privacy/security threats. Kamhoua *et al.* model this situation as a stag hunt game with complete information and based on the game's mixed Nash equilibrium explain the minimum trust level (players' beliefs on probability of other players' cooperation) required to have a protected resource. They further use evolutionary game theory to examine the effect of assuming rational play-

ers and show how the game approaches one of its two pure Nash equilibria depending on the original population's trust level. Rajbhandari and Snekkenes [RS10] use game theory to study privacy risks. They propose a strategic game model of complete information between a data provider and a data collector. A data provider chooses to either trust the data collector and submit genuine personal information or to provide fake information. A data collector can act honestly and follow the promised privacy policy or be malicious and sell the collected information to a third-party without the data provider's permission. The authors apply the solution concept of Nash equilibrium to determine mixed strategies of the two players and to measure the amount of privacy risk involved in the strategic situation.

Challenges and issues within the process of setting a privacy policy have also been subjects of game theoretic analyses. Prebusch [Pre06] models interactive privacy negotiation as a Bayesian game. He proposes a dynamic approach to set privacy policies using a series of offers proposed by the data collector to each data provider who is not a privacy fundamentalist (based on the classification suggested by Spiekermann [SGB01]). As the negotiation progresses, lower privacy with higher incentive is offered to a data provider. In this work, the notion of Bayesian game is used only to provide a visual representation of the sequence of actions within the negotiation and no formal analysis of the game is provided.

The issue of allocating a reward in exchange for private information has been studied by Kleinberg *et al.*[KPR01] where they describe three scenarios modeled as coalition games [Osb03] and use core and shapely values to find a "fair" reward allocation method. However, the underlying assumption in these 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.

Calzolari and Pavan [CP01] use an extensive form game with incomplete information to explore the optimum flow of customers' private information between two interested firms. The players of the game are a data provider (customer) and two firms who individually

contact the data provider. The first firm offers a price for the good that the data provider is interested in and provides the two options of keeping the terms of their contract private or publicly reveal it. If the data provider agrees to the price associated with the option of public revelation of contract information, then the first firm sends the private contract information to the second firm (with some probability). The second firm can use this information to refine her prior belief about the data provider's willingness to pay and adjust the price accordingly.

Domingo-Ferrer [DF09] recognizes axioms of functionality, security, and privacy as three different requirements in information systems. From a very broad perspective, he postulates that parties' requirements in each context can be a mixture of these three dimensions and proposes that if the situation is modeled as a game, then a balanced level between these axioms can be found in the game's equilibrium. Besides suggesting to consider a mixture of these three axioms for players' payoffs, this article does not provide any specific formulation of the game, players' strategies, player's payoff functions, or solution to it. Moreover, the interleaved connection between these axioms are never addressed. Therefore, this work can be considered as a very general perspective on applicability of game theory in providing "green ICT services".

3.3 Summary

In this chapter we discussed game theory concepts that are related to this project. Players, actions, and preferences are the main ingredients of a game. We discussed how a game can be represented as either a strategic or an extensive form. The most common solution concept used for solving games is known as the Nash equilibrium. To solve extensive form games, a concept stronger than Nash equilibrium is used which is referred to as the subgame perfect equilibrium. The method of (generic) backward induction is commonly applied to find the subgame perfect equilibria of an extensive form game. We finished this chapter by reviewing some literature that apply these game theory concepts to study various challenges in privacy.

In the next chapter we explain how we use game theory to find balanced privacy/utility tradeoffs during the process of privacy parameter setting.

Chapter 4

Our Generic Privacy Game Model

Armed with sufficient background on privacy protection mechanisms and game theory principles, this chapter introduces a new application of game theory in setting privacy parameters. In this application, the challenge of establishing balancing values for privacy parameters (within either a privacy policy specification or a data sanitization mechanism) is modeled as an extensive form game with incomplete and imperfect information. *Consensual* privacy protection levels are found as the game's subgame perfect equilibria in the sense that no player has an incentive to unilaterally deviate form the strategy profile.

Section 4.1 explains the challenge and unanswered questions in setting privacy parameters. In Section 4.2 we explain how this challenge can be modeled as a strategic situation and describe the backbone of our proposed game model. Section 4.3 provides a generic approach to solve the game. As we see in the following chapters, for each instantiation of our generic game model, an instance of the provided solution guideline is used to find the game's subgame perfect equilibria.

4.1 Privacy Parameter Setting

As we saw in Chapter 2 there are two common approaches to protect individuals' privacy in data repositories: Data sanitization and Privacy policy declaration. Data sanitization systems are not effective if data providers' consent to the level of privacy protection is not sought in advance. Moreover, privacy policy declaration methods must specify some data granularity levels at which data is provided to users of the data. Therefore, in every effective privacy protection mechanism the essence of data providers' consent and data masking (using generalization, perturbation and similar methods) must coexist. The amount of data masking

applied to private records is controlled by value(s) of one or more parameters in each privacy protection mechanism. In this work, we use the term *privacy parameter* to refer to a vector $\delta = \langle \delta_1, \cdots, \delta_z \rangle$. For each privacy protection mechanism, privacy parameter δ represents parameters embodied in the mechanism to adjust the amount of data masking. For instance, in k-anonymity $\delta = \langle k \rangle$, in l-diversity $\delta = \langle l \rangle$, in differential privacy $\delta = \langle \epsilon, \Delta Q \rangle$, and in privacy taxonomy (with g_i representing the granularity level of data item i and r representing retention) $\delta = \langle g_1, g_2, \cdots g_n, r \rangle$ are privacy parameters. Therefore, the exact meaning of δ must be interpreted according to the privacy protection mechanism chosen for the game.

Depending on the privacy protection mechanism, each component of the vector δ can be of one of the following two types:

- **privacy enhancer -** A component is a privacy enhancer if the privacy protection level increases (or stays the same) as the value of the component increases. An increase in a privacy enhancer's value leads to a decrease in data utility. For example, k in k-anonymity is a privacy enhancer.

- **utility enhancer -** A component is a utility enhancer if the utility of the masked data increases (or stays the same) as the value of the component increases. An increase in a utility enhancer's value leads to a decrease in the privacy protection level. For example, ϵ in differential privacy is a utility enhancer (as the value of ϵ increases less noise is applied to the data).

Regardless of its type, a privacy parameter component usually has opposing effects on privacy protection level and data utility. As the privacy protection level implied by δ increases more data masking (such as generalization, perturbation, and/or suppression) is applied to the records and hence the private dataset has lower utility to a data user. However, higher levels of privacy protection convince more data providers to share their personal information. Therefore, any privacy protection mechanism has the challenge of finding balanced values for its privacy parameter components. Formulating this challenge as just a function to be

optimized by the collector of personal data is an oversimplification of the problem because there are other rational agents in the environment who can make decisions and their decisions influence the outcome. Therefore, a privacy/utility tradeoff that is optimum for one party might never be achieved as long as other stakeholders of the data have some incentive to act differently and change the results towards a setting that is more profitable for them. In fact, the challenge of setting a privacy parameter is a strategic situation where data users try to get more data with higher utility, the data collector attempts to achieve the highest profit from providing a private data table to data users, and data providers prefer to provide their personal data at higher levels of privacy protection and incentive.

We model this strategic situation using an extensive form game with imperfect information.[1] The privacy parameter values found at the game's subgame perfect equilibria provide balanced prescriptions for privacy/utility tradeoffs. In this study we refer to these balancing values of the privacy parameters as "stable" privacy parameter settings (because they are found at the game's equilibria an no player has the incentive to unilaterally deviate from the strategies prescribed by the equilibria).

4.2 The Game of Setting Privacy Parameters

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

[1] Our game theoretic approach to address the privacy/utility tradeoff challenge can be considered as a realization of the "green ICT services" suggested by Domingo-Ferrer [DF09].

4.2.1 Players and Payoffs

In a game of privacy parameter setting we recognize three different groups of decision makers: a group of n data providers DP, a data collector DC, and some data users DU. [2] The description of each player and her payoff is provided in the following sections.

4.2.1.1 Data Providers

Data providers are individuals who decide whether to provide their personal information at a specific privacy protection level implied by 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.

In our game we consider a pool of n different data providers each making a decision independent of the decisions made by other data providers. Notice that assuming an independent process of decision making for each data provider is a very realistic assumption, unless social networking is considered as the incentive for collecting information. In the latter case, each data provider prefers the outcomes in which more data providers opt-in for the data collection.

As we saw in Chapter 2, the privacy/incentive preferences of each data provider is affected by several demographic and socioeconomic factors [AG05, CA99, SMC93]. Therefore, it is practically infeasible to specify how much utility is gained by a data provider for each combination of δ and incentive. The uncertainty about data providers' payoff functions can be modeled as a game of incomplete information (see Section 3.1.4). In this model, multiple types are considered for each data provider. In every instance of the game, data provider DP_i knows her own type but other players can only guess her type based on a probability

[2] Other scholars [RS10, Pre06, CP01] have also considered a subset of these players to model a privacy challenge as a game. However, these works either study the privacy challenges in a different context [RS10, CP01] or use the informal game model only to provide a visual representation of the interactions with no further analysis [Pre06].

distribution (known as a common knowledge of the game to every player).

Based on Westin's privacy indexes [KC05] we consider three main types for each data provider: privacy unconcerned with probability p_1, privacy pragmatist with probability p_2, and privacy fundamentalist with probability p_3 (see Chapter 2 for more information on these types). The last two types are further categorized into several subtypes. Characteristics and preferences of each type are as follows:

- **Privacy Unconcerned -** A data provider of this type provides her personal information for any combination of privacy/incentive. Therefore, for any combination $\langle \delta, I \rangle$, where δ is the announced privacy parameter value and I is the amount of incentive in monetary value, utility of a privacy unconcerned data provider for taking action b is defined as:

$$U_i^u(b, \langle \delta, I \rangle) = \begin{cases} I + \epsilon & if \quad b = optIn \\ 0 & if \quad b = optOut \end{cases} \tag{4.1}$$

  In the formulation we added the extra utility ϵ to represent other hidden benefits that a privacy unconcerned data provider finds in providing her personal information (and that is why she always provides her information regardless of the announced settings). As the utility function implies, for a privacy unconcerned data provider, the action of opt in dominates her action of opt out. As already mentioned, a data provider is of this type with probability p_1.

- **Privacy Pragmatist -** A privacy pragmatist data provider considers privacy tradeoffs to receive the incentive. We can think of a privacy pragmatist as a player who has some *threshold* on the combination of privacy protection level and incentive, below which she will refuse to provide personal information. In other words, for some functions $h(.)$ and $g(.)$, a pragmatist data provider compares $w_1 \cdot h(\delta) + w_2 \cdot g(I)$ to her threshold and opts in if this value is at least equal to her threshold. In this formulation, $g(I)$ is the rewarding

effect of the incentive. The function $h(\delta)$ denotes the privacy gain and is an increasing function of the privacy protection level promised by privacy parameter δ. This means, for each component δ_j in vector δ if δ_j is a privacy enhancer then $\frac{\mathrm{d}h}{\mathrm{d}\delta_j} \geq 0$ and if δ_j is a utility enhancer then $\frac{\mathrm{d}h}{\mathrm{d}\delta_j} \leq 0$. Function $g(I)$ is an increasing function of the incentive I. Parameters w_1 and w_2 represent the weights of $h(\delta)$ and $g(I)$. In fact, the formula $w_1 \cdot h(\delta) + w_2 \cdot g(I) = k, \quad k \in \{k_1, k_2, \cdots k_m\}$ represents *indifference curves* of data providers with the assumption that privacy and incentive are *perfect substitutes* [Var10]. Each curve has a constant slope of $-\frac{w_1}{w_2}$, the magnitude of which represents the marginal rate of substitution of incentive for privacy. Therefore, a more privacy aware population will have a larger ratio of $\frac{w_1}{w_2}$ since their marginal willingness to give up incentives to receive a small amount of privacy is relatively higher. With characteristics of these two functions we can safely assume that $0 \leq w_1 \cdot h(\delta) + w_2 \cdot g(I)$. We use parameter $\overline{v}$ to denote the upper bound on $w_1 \cdot h(\delta) + w_2 \cdot g(I)$ values [3]. Consequently, we have $0 \leq w_1 \cdot h(\delta) + w_2 \cdot g(I) \leq \overline{v}$.

We do not assume a single threshold for every privacy pragmatist. Instead, we consider multiple subtypes of a privacy pragmatist data provider i where each subtype $t_i^{p,j}$ (we use super script p to specify that subtype belongs to the privacy pragmatist type) has threshold $l_i^{p,j} \in [0, \overline{v}]$. For subtype $t_i^{p,j}$ of a privacy pragmatist data provider i, the utility of taking action b is defined as:

$$U_i^{p,j}(b, \langle \delta, I \rangle) = \begin{cases} w_1 \cdot h(\delta) + w_2 \cdot g(I) - l_i^{p,j} & if \quad b = optIn \\ 0 & if \quad b = optOut \end{cases} \tag{4.2}$$

[3] Assuming an upper bound on this function is not unrealistic; for each privacy protection method there is often a certain level of privacy protection after which data will lose almost all of its utility (For example, a value of k in k-anonymity greater than the number of records) and as we see later a data collector never announces incentives higher than what she is paid by the data user.

In our game analysis, we are only considering pure strategies for each type of data provider since the game already has enough randomness and in such games "players need not -indeed should not - introduce any additional randomness by using mixed strategies" [Har73]. Since the data providers are only using pure strategies, it is sufficient for the payoff functions to convey only *ordinal* information about their preferences [Osb03]. The utility function in Equation 4.2 simply states that a privacy pragmatist prefers to opt in only when $w_1 \cdot h(\delta) + w_2 \cdot g(I)$ is the same or greater than her threshold. Otherwise, her payoff for opting in would be negative and she prefers not to opt in.

In this work, we assume that the threshold of a privacy pragmatist data provider is distributed uniformly between 0 and $\overline{v}$.[4] Therefore, all subtypes for a privacy pragmatist are equally probable. More formally, values of $l_i^{p,j}$ (and hence the corresponding subtypes for a privacy pragmatist) admit the following uniform density function:

$$f^p(l_i^{p,j}) = 1/\overline{v} \tag{4.3}$$

As a result, in any instance of the game, the fraction of privacy pragmatists with threshold of at most l is:

$$F^p(l) = \int_0^l f^p(x)\mathrm{d}x = l/\overline{v} \tag{4.4}$$

As we shall see in Section 4.3.1, we use the cumulative distribution function of Equation 4.4 to find the fraction of subtypes of a privacy pragmatist data provider who would opt in for a combination of δ and I. A data provider is a privacy pragmatist with probability p_2.

[4] Assuming a different distribution might lead to a different privacy behavior model for data providers (see Section 4.3.1). In this case, the process of analyzing the game remains the same but the game solution could be different. Without further empirical studies on data providers' behavior we see no reason to consider a more complicated distribution function.

- **Privacy Fundamentalist -** Westin [KC05] defines privacy fundamentalists as those who never opt in if their data is collected to be stored. In a data collection environment, this definition for privacy fundamentalists leads to a subset of the population who simply never opt in and we can simply leave this subset of the population out of the game model since they do not affect any player's payoff. Instead, we use a different interpretation for privacy fundamentalists.: " A privacy fundamentalist is a data provider who provides her personal information only if she feels that her privacy is properly protected". With this definition, a privacy fundamentalist only cares about the level of privacy protection. But what level of δ is considered a secure privacy protection level? Similar to the privacy pragmatist type, each data provider of type privacy fundamentalist might have a different *threshold* below which she feels her privacy is not protected. Let $h(\delta)$ be a function that represents the amount of privacy protection provided by privacy parameter δ. We assume that $h(\cdot)$ is the same function specified for privacy pragmatists' tradeoffs and is bounded by $0 \leq h(\delta) \leq \overline{v}'$.

  We do not assume a single threshold for every privacy fundamentalist. Instead, we consider multiple subtypes of a privacy fundamentalist data provider i where each subtype $t_i^{f,j}$ (we use super script f to specify that the subtype belongs to privacy fundamentalist type) has threshold $l_i^{f,j} \in [0, \overline{v}']$. For subtype $t_i^{f,j}$ of a privacy fundamentalist data provider i, utility of taking action b is defined as:

$$U_i^{f,j}(b, \langle \delta, I \rangle) = \begin{cases} h(\delta) - l_i^{f,j} & if \quad b = optIn \\ 0 & if \quad b = optOut \end{cases} \tag{4.5}$$

  In this work, we assume that threshold of a privacy fundamentalist data provider is distributed uniformly between 0 and $\overline{v}'$.[5] Therefore, all subtypes

[5]Assuming a different distribution might lead to a different privacy behavior model for data providers

for a privacy fundamentalist are equally probable. More formally, values of $l_i^{f,j}$ (and hence the corresponding subtypes for a privacy fundamentalist) follow the following uniform probability density function:

$$f^f(l_i^{f,j}) = 1/\overline{v}' \tag{4.6}$$

As a result, in any instance of the game, the fraction of privacy fundamentalists with threshold of at most l is:

$$F^f(l) = \int_0^l f^f(x)\mathrm{d}x = l/\overline{v}' \tag{4.7}$$

As we shall see in Section 4.3.1, we use the cumulative distribution function of Equation 4.7 to find the fraction of subtypes of a privacy fundamentalist data provider who would opt in for a privacy setting defined by δ. As mentioned earlier, a data provider is a privacy fundamentalist with probability p_3.

4.2.1.2 Data Users

A data user is an entity interested in accessing personal information for some data analysis purposes. A data user prefers a dataset with higher quality (more accurate query results) and higher cardinality (results with higher statistical significance). The privacy parameter δ affects these requirements in positive and negative ways. Therefore a data user asks for a value δ that balances these needs and initiates the game by offering some value for privacy parameter δ and some price, p, for each data record.

We generally consider some bounds Min and Max as part of the game description to define the minimum and maximum number of data records necessary for data analysis. In other words, if the dataset has less than Min number of records, the collected dataset is not useful to the data user and increasing the number of collected records over the Max limit

(see Section 4.3.1). In this case, the process of analyzing the game remains the same but the game solution could be different. Without further empirical studies on data providers' behavior we see no reason to consider a more complicated distribution function.

does not add extra value to the data user's profit. Let N denote the number of collected records in a private dataset.[6] From the data user's perspective, the *effective cardinality* of the dataset EN is:

$$EN(N) = \begin{cases} Max & if \quad Max < N \\ N & if \quad Min \leq N \leq Max \\ 0 & if \quad N < Min \end{cases} \tag{4.8}$$

As a result, the data user only pays price p for each record based on the effective cardinality of the dataset and not its actual cardinality. The data user's expenditure is:

$$expenditure_{DU} = p \cdot EN(N) \tag{4.9}$$

Let b denote the economic value of each record to the data user, *i.e.*, b represents the net revenue of a data record if the data user gets the record for free. If the number of data records collected from individuals is denoted by N, the effective cardinality of the dataset will be $EN(N)$ (see Equation 4.8) and we can initially define the data user's income as $b \cdot EN(N)$. However, after applying a privacy protection method the utility of data drops due to the imprecision introduced to the results of the queries. We use the parameter $0 \leq Precision(\delta, N) \leq 1$ 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:

$$income_{DU} = b \cdot \ EN(N) \cdot \ Precision(\delta, N) \tag{4.10}$$

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 protection method and the algorithm used, database schema, privacy parameter value 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 common knowledge of the game). Therefore, $Precision$ is defined as a function of two variables δ and N.

[6] As we see later in this chapter, N is a function of δ and I.

To sum up, payoff to a data user can be defined as:

$$U_{DU} = b \cdot EN(N) \cdot Precision(\delta, N) - p \cdot EN(N) \tag{4.11}$$

We give the detailed analysis for games with a single data user. Please refer to Section 4.2.3 for a discussion on how to model multiple data users and data reuse.

4.2.1.3 Data Collector

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

A data collector prefers to receive more money from the data user and spend less money on the amount of incentive she pays to the data providers. Consequently, the cardinality of the dataset (number of data providers who opt in) affects the payoff to the data collector.

Let p denote the amount of money a data user pays to the data collector for each data record based on the effective cardinality of the dataset EN. As explained in Section 4.2.1.2, if the total number of data providers who opt in is N then based on the values of Max and Min, the effective cardinality of the dataset $EN(N)$ is a value from 0 to Max (see Equation 4.8). Consequently, the data collector's income is:

$$income_{DC} = p \cdot EN(N) \tag{4.12}$$

The data collection procedure, applying a privacy protection method, and storing the dataset are costly and we denote these costs by C. In fact, the cost C represents the amount of money that the data collector requires as a compensation for all of her data collection, storage and privacy protection efforts. Moreover, the data collector has to pay some incentive,

I, to each data provider. As a result, the expenses to the data collector can be defined as:

$$expenditure_{DC} = I \cdot N + C \tag{4.13}$$

For simplicity of analysis we have assumed a fixed cost C for the data collector. We drop this assumption for one of the game's instantiations in Chapter 5 and define the cost as a function of the privacy level δ. The payoff to the data collector is therefore defined as:

$$U_{DC} = income_{DC} - expenditure_{DC} = p \cdot EN(N) - I \cdot N - C \tag{4.14}$$

4.2.2 Rules of the Game

We model interactions between the data user, the data collector and the data providers as a sequential game with *imperfect* (some players are not fully aware of the actions taken by previous players) and *incomplete* (some uncertainty exists about data providers' preferences) information. More specifically, when a data provider makes a decision, she is unaware of the decisions made by other data providers (but she knows the decisions made by the data user and the data collector). Each data provider is aware of her own preferences but other players are uncertain about her type and only know a probability distribution over possible types she can take. There is no uncertainty about the data user and the data collector (they are assumed to know each other's preferences).

The game starts with an offer from the data user to the data collector. In the offer, the required value for privacy parameter δ and price p (per each record) must be specified. We denote an offer by $Of = \langle \delta, p \rangle$. Once the data collector receives the offer she 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 she needs to announce an incentive in exchange for collecting personal information. Here, we assume that I represents the monetary value of the incentive and its domain is $\mathbb{R}_{\geq 0}$. When the data collector announces an incentive, n data providers see the combination (Of, I) and simultaneously decide whether they want to opt in or not. By the term "simultaneously"

we do not mean they all decide at the same time. Rather, the simultaneous move implies that when each data provider makes a decision, she is unaware about the decision made by other data providers. We denote these situations by information sets (see Section 3.1.2) in the game tree.

Let $dpv \in \{optIn, optOut\}^n$ be an $n-$length vector where element dpv_i in the vector represents the decision made by data provider i. The terminal histories of this game are either of the form $(Of, Reject)$ or (Of, I, dpv).

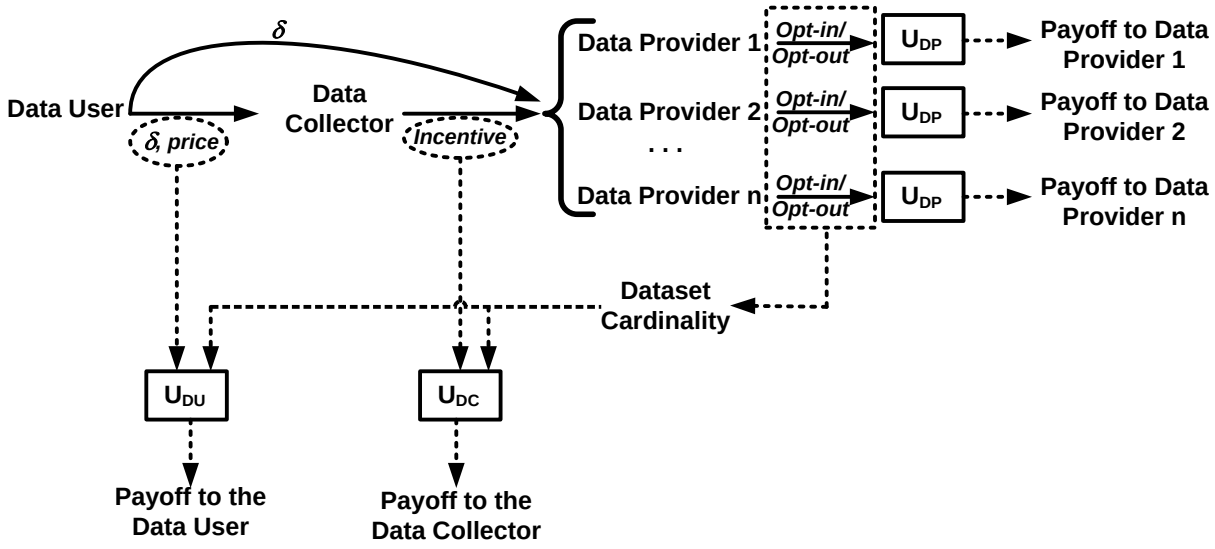


Figure 4.1: The dynamics of setting a stable value for a privacy parameter

The interactions and mutual effects of players' decisions are captured in Figure 4.1. Based on the game's dynamics, Figure 4.2 illustrates the game tree, where figure triangles represent ranges of possible offers and incentives and from each range a sample branch is drawn.

Example 4.2.1. *To provide a concrete example of such interactions, consider a publisher (the data user) who is evaluating the market for certain types of books. To conduct a more reliable analysis and target the right population, she requires a database that describes some characteristics as well as shopping habits of potential customers. Consequently, she may contact Amazon Inc (the data collector) and offer to pay price p for each record that Amazon*

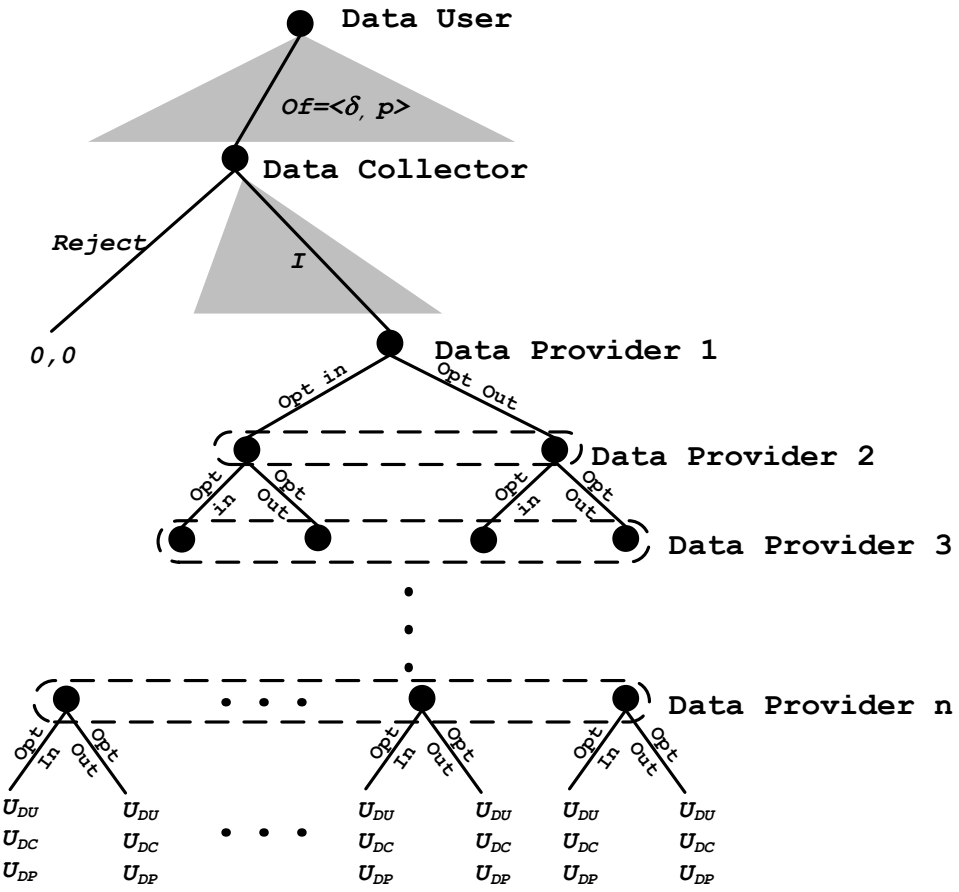


Figure 4.2: The visualization of Privacy game tree

provides her with. Moreover, she understands that Amazon is going to employ method x to protect privacy of its customers (data provider). Considering the effects of method x on precision of her data analysis results, she also suggests a privacy protection level that can potentially maximize her revenue. If Amazon finds such an offer profitable, it announces some incentive (for example some discount on customers' shopping cart) to those customers who agree to the level of privacy and provide the required personal information. Finally, the data providers (customers) individually assess the combination of privacy level and incentive and make their decision.

The proposed extensive-form game models a "one shot negotiation" between the data user and the data collector. In practice, the negotiation can continue in multiple rounds (each time a data user makes an offer, the data collector can accept, reject, or suggest a

different offer). Since we are not considering the time factor and both the data collector and the data user are perfectly informed about the decisions made by players before them, adding multiple rounds of negotiation does not change the results as long as it ends with an offer from the data user followed by a response from the data collector. We believe that a negotiation which is initiated by a data user (and not the data collector) is more natural and realistic. However, should the data collector initiate the negotiation, the analysis of the game and the outcome will potentially be different.

4.2.3 Data Reuse and Multiple Data Users

Consider a situation where a private data table is already collected for purpose *pur* and with privacy parameter δ (and provided to a data user). Let $\overline{\Delta}$ denote the set of all possible privacy parameter values that guarantee a privacy protection at least as high as δ and let $\underline{\Delta}$ represent the set of all possible privacy parameter values that provide a lower privacy protection level compared to δ. If a second data user (or the same data user with a different purpose) asks to access the same data table with privacy parameter δ', depending on the requirements of the secondary data user, two situations can happen:

(a) The new data user wants the data table for some purpose allowed by *pur* and chooses to offer with a privacy parameter $\delta' \in \overline{\Delta}$. 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 $\delta' = \delta$ because cardinality of the dataset is fixed in this case and asking for any $\delta' \in \overline{\Delta} - \{\delta\}$ just causes lower data utility and lower payoff. In this equilibrium, the data user offers price $p = C'$ where C' is the new cost representing the amount that the data collector requires to be financially compensated for all of her efforts in providing the existing data table to a secondary data user. The cost C' can be either higher or lower than the original cost C. The accumulative cost of data maintenance for a future secondary data reuse must

be accounted for when C' is calculated. However, depending on the privacy protection method, the data collector might not need to repeat the data masking procedure and therefore the cost C' can be lower than C. Notice that a secondary data user can associate a different value to data records. For example, if a data user has access to a private data table, her competitors might become more interested in the same data table to keep up with their rival.

(b) The new data user wants the data table for some purpose not allowed by *pur* or chooses to offer with a privacy parameter $\delta' \in \underline{\Delta}$ (δ' implies lower privacy protection than the level of protection provided by δ). Since the new data usage has a different purpose or a lower privacy protection is requested, it is the data collector's responsibility to ask the data providers again for a data reuse with a different purpose or lower privacy. This case becomes a new instance of the game with different problem settings. As explained in scenario (a), the new cost of data collection, storage, and protection C' and the new data user's valuation for each record b' can be either higher or lower than the original cost C and original valuation b, respectively. Moreover, the data providers might react differently to a secondary request for their data. In some cases, a data provider might require the combination of privacy/incentive to meet a higher threshold since she might associate a higher value to her data. In other cases, a data provider might think that she has already lost control over her information by providing it to the data collector the first time. In this case, a combination of privacy/incentive that meets a lower threshold can still convince her to provide her consent to a secondary usage of her information.

Data sanitization methods do not explicitly support the notion of purpose. Therefore, whenever the privacy protection method is a sanitization mechanism, the purpose is implicitly considered to be *any* which is assumed to allow data usage for any purpose that a data user might have.

Knowing these two cases, a new data user (or the same data user with a different purpose) can find the equilibria of the game in each case and after comparing her expected payoffs, choose the case that provides her with a higher overall benefit. A more comprehensive approach, must model this strategic situation as a single repeated game with multiple data users. Using a repeated game model, one can incorporate the induction about subsequent data users' reactions into the first data user's decision. Moreover, the data collector might reject some (low profit) offers to establish a reputation. Other aspects of the negotiation such as requesting an exclusive right to the dataset can also be considered in a repeated game model. Consequently, a repeated game model can clarify how the new values of C', b', and other problem settings change from one iteration to the next one. Studying this model is in our plan for future extension to this work.

4.3 General Approach to Solve the Game

The game model we have explained is an extensive game with imperfect information. As explained in Section 3.1.3 we can find the subgame perfect equilibria of the game by using the generic backward induction method. This method suggests to start with the smallest subgames of the tree. Based on the definition of subgame and description of our game, the smallest subgames are the ones that start after all histories of the form $h = (Of, I)$ where Of is an offer chosen by the data user and I is the amount of incentive that the data collector announces for such an offer. For simplicity, we refer to all such subgames as "data collection subgames". Each of these subgames is a game of simultaneous move by n data providers all knowing the chosen combination of δ and I. Once the Nash equilibria of each of these subgames are determined, we fold back the game tree one step and find the data collector's best response to each offer received from the data user assuming that the outcome of the "data collection subgames" are the ones we found earlier. Finally, in the last step, we examine the data user's best response taking the best responses of data providers and the

data collector as given. Each of these three steps are explained in the next three subsections.

4.3.1 Data Providers' Best Response

The first step in solving an instance of the game is to find the Nash equilibria of the smallest subgames referred to as "data collection subgames". At each of these subgames, the data user has already chosen to request data with a specific privacy parameter value δ and the data collector has accepted the offer and announced the incentive I. The n data providers are aware of the chosen values δ and I and *individually* decide whether to opt in or opt out from the data collection procedure.

As mentioned in Section 4.2.1.1, the decision made by a data provider i is assumed to not affect any other data provider (but it affects the data collector and the data user). More formally, let s_i be a strategy of player i and let s_{-i} and s'_{-i} denote two different strategies for all other players (a sequence of $n-1$ choices of $OptIn$ or $OptOut$), we have the following for all data providers $i \in \{1, \cdots, n\}$:

$$\begin{gathered} \forall s_i \in \{OptIn, OptOut\} \ \wedge \ \forall s_{-i}, s'_{-i} \in \{OptIn, OptOut\}^{n-1} : \\ U_i(s_i, s_{-i}) = U_i(s_i, s'_{-i}) \end{gathered} \tag{4.15}$$

Based on Equation 4.15, we can find the best response of each data provider in isolation without considering the effects of other data providers' decision. Once the best response BR_{DP_i} of data provider i is found, the best response of every data provider would be the same as BR_{DP_i}. This is due to the fact that data providers all have the same strategy set $\{OptIn, OptOut\}$ and players of each type have identical payoffs. Therefore, each data collection subgame is a *symmetric* game [Dut99]. Since a data provider's decision has no bearing on other data providers' decisions and the subgame is symmetric, the process of finding the best response of data provider i, is identical to the process of finding the best response of any other data provider i'.

At each instance of the data collection subgame, a data provider i knows δ, I, and her own

type. Based on the payoff functions specified for each type and subtype of a data provider (see Section 4.2.1.1), we can summarize a data provider's best response as follows:

- If the data provider is of type "privacy unconcerned" then her payoff function is what Equation 4.1 suggests. In this function we have:

$$\forall \langle \delta, I \rangle \ : \ U_i^u(OptIn, \langle \delta, I \rangle) > U_i^u(OptOut, \langle \delta, I \rangle) \tag{4.16}$$

  Therefore, opting out is always dominated by strategy $OptIn$ for this type of player i. The best response of this type of data provider is $OptIn$.

- If the data provider is of type "privacy pragmatist" with threshold l then depending on how l compares to the combination of privacy protection level and incentive, her best response might be either to $OptIn$ or $OptOut$. Consider functions $h(\cdot)$ and $g(\cdot)$ with the specifications explained in Section 4.2.1.1, based on Equation 4.2 we have the following three inequalities:

$$\forall \ l < w_1 \cdot h(\delta) + w_2 \cdot g(I) \ : \ U_i^p(OptIn, \langle \delta, I \rangle) > U_i^p(OptOut, \langle \delta, I \rangle) \tag{4.17}$$

$$\forall \ l = w_1 \cdot h(\delta) + w_2 \cdot g(I) \ : \ U_i^p(OptIn, \langle \delta, I \rangle) = U_i^p(OptOut, \langle \delta, I \rangle) \tag{4.18}$$

$$\forall \ l > w_1 \cdot h(\delta) + w_2 \cdot g(I) \ : \ U_i^p(OptOut, \langle \delta, I \rangle) > U_i^p(OptIn, \langle \delta, I \rangle) \tag{4.19}$$

  These three equations show that for a subtype with threshold less than $w_1 \cdot h(\delta) + w_2 \cdot g(I)$ strategy $OptIn$ dominates strategy $OptOut$. On the contrary, for a subtype with threshold higher than $w_1 \cdot h(\delta) + w_2 \cdot g(I)$ the strategy of $OptOut$ is strictly preferred to the strategy of $OptIn$. If the subtype has a threshold equal to $w_1 \cdot h(\delta) + w_2 \cdot g(I)$ both strategies $OptIn$ and $OptOut$ are best responses.

- If the data provider is of type "privacy fundamentalist" with threshold l then depending on how l compares to the level of privacy protection guaranteed by δ, her best response might be either to $OptIn$ or $OptOut$. Consider function $h(\delta)$ as an increasing function of privacy protection levels (see Section 4.2.1.1), based on Equation 4.5 we have the following three inequalities:

$$\forall\ l < h(\delta)\ :\ U_i^f(OptIn, \langle \delta, I \rangle) > U_i^f(OptOut, \langle \delta, I \rangle) \tag{4.20}$$

$$\forall\ l = h(\delta)\ :\ U_i^f(OptIn, \langle \delta, I \rangle) = U_i^f(OptOut, \langle \delta, I \rangle) \tag{4.21}$$

$$\forall\ l > h(\delta)\ :\ U_i^f(OptOut, \langle \delta, I \rangle) > U_i^f(OptIn, \langle \delta, I \rangle) \tag{4.22}$$

  These three equations show that for a subtype with threshold less than $h(\delta)$ strategy $OptIn$ dominates strategy $OptOut$. On the contrary, for a subtype with threshold higher than $h(\delta)$ the strategy of $OptOut$ is strictly preferred to the strategy of $OptIn$. If the subtype has a threshold equal to $h(\delta)$ both strategies $OptIn$ and $OptOut$ are best responses.

It has been proven that to find the Nash equilibria in a game of incomplete information where fraction p of the population representing player i choose a specific action a we can equivalently consider a situation where player i plays with a mixed strategy s such that s assigns weight p to action a [Osb03]. Consequently, knowing how likely it is to have a data provider of each type and the best response of each type, we can find the Nash equilibria of the game by interpreting the collection of pure strategies (one for each type) as a single mixed strategy that explains the behavior of a data provider randomly chosen from a large population (with different types).

In each data collection subgame with a combination of privacy parameter δ and incentive I, a privacy unconcerned opts in with probability 1, a privacy pragmatist opts in if her threshold is at most $w_1 \cdot h(\delta) + w_2 \cdot g(I)$, and a privacy fundamentalist opts in if her

threshold is at most $h(\delta)$. More specifically, the probability of a privacy pragmatist opting in, $Prob(OptIn \mid type = pragmatist)$, is the probability of a privacy pragmatist having a threshold of at most $w_1 \cdot h(\delta) + w_2 \cdot g(I)$. Similarly, the probability of a privacy fundamentalist opting in, $Prob(OptIn \mid type = fundamentalist)$, is the probability of a privacy fundamentalist having a threshold of at most $h(\delta)$. With the assumption of uniform distributions for the thresholds, functions $F^p(\cdot)$ from Equation 4.4 and $F^f(\cdot)$ from Equation 4.7 can be used to find the probabilities of the events in which a privacy pragmatist and a privacy fundamentalist opt in for a combination of δ and I. These probabilities are as follows:

$$\begin{aligned} Prob(OptIn \mid type = pragmatist) &= \int_0^{w_1 \cdot h(\delta) + w_2 \cdot g(I)} f^p(x)\mathrm{d}x = [w_1 \cdot h(\delta) + w_2 \cdot g(I)]/\overline{v} \\ Prob(OptIn \mid type = fundamentalist) &= \int_0^{h(\delta)} f^f(x)\mathrm{d}x = h(\delta)/\overline{v}' \end{aligned} \tag{4.23}$$

In these equations, $\overline{v}$ is the upper bound on $w_1 \cdot h(\delta) + w_2 \cdot g(I)$ and $\overline{v}'$ is the upper bound on $h(\delta)$.

Let p_1, p_2 and p_3 denote the probability of a data provider being of types privacy unconcerned, privacy pragmatist, and privacy fundamentalist, respectively. Based on Equation 4.23, the probability of a data provider (randomly chosen from a population of mixed types) opting in, $Prob(OptIn)$, is:

$$Prob(OptIn) = p_1 \cdot 1 + p_2 \cdot \frac{w_1 \cdot h(\delta) + w_2 \cdot g(I)}{\overline{v}} + p_3 \cdot \frac{h(\delta)}{\overline{v}'} \tag{4.24}$$

Let $\beta_0 = p_1$, $\beta_1 = [p_2 \cdot w_1]/\overline{v} + p_3/\overline{v}'$, and $\beta_2 = [p_2 \cdot w_2]/\overline{v}$. We can rewrite Equation 4.24 as the following regression model:

$$Prob(OptIn) = \beta_0 + \beta_1 \cdot h(\delta) + \beta_2 \cdot g(I) \tag{4.25}$$

Equation 4.25 represents the best response of a data provider (chosen randomly from a population of mixed types) as a mixed strategy. As explained earlier, because the subgame is symmetric and data providers do not affect each other's decisions we conclude that Equation

4.25 is the mixed strategy adopted by *every* data provider in a Nash equilibrium. As a result, in the Nash equilibrium of each data collection subgame n data providers optimally respond to a combination of privacy parameter δ and incentive I with a mixed strategy $BR_{DP_i} = \{(Prob(OptIn), 1 - Prob(OptIn))\}$. In this mixed strategy, the first element represents the weight of choosing to opt in and the second is the weight of choosing to opt out.

As an alternative interpretation, we can view Equation 4.25 as a regression model produced based on some observation which is available to the data user and the data collector. 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 [AG05, SMC93, MG93]. A regression model that explains the effects of δ and the incentive seems to be a natural extension to those studies.

In each instance of the game (for each privacy protection mechanism), the procedure of finding Nash equilibria of data collection subgames is identical and leads to the opt-in probability formula specified in Equation 4.25. However, within Equation 4.25 an appropriate definition for functions $h(\cdot)$ must be adopted such that $h(\delta)$ is a reasonable metric for the amount of privacy protection. In all of our game analyses we have chosen the identity function for the description of function $g(\cdot)$ (*i.e.*, we choose $g(I) = I$). If after some experimental studies $g(\cdot)$ is proven to be a different function then the procedure of solving the games remains the same but the outcome of the games might be different.

4.3.2 Data Collector's Best Response

Once the best response of data providers has been determined in data collection subgames, we fold back the game tree one step and solve the subgames that start after histories of the form $h = (Of)$. At the root of each of these subgames the data user has already made an offer $Of = \langle \delta, p \rangle$ and the data collector must make a decision. Based on the offer Of

received from the data user, the data collector chooses the most profitable action assuming that the outcome of data collection subgames are the ones we found in Section 4.3.1.

The data collector can estimate the expected cardinality of the dataset, N, for each δ and I based on Equation 4.25 as follows:

$$N(\delta, I) = n \cdot [\beta_0 + \beta_1 \cdot h(\delta) + \beta_2 \cdot g(I)] \tag{4.26}$$

If we plug Equation 4.26 into the U_{DC} formula from Equation 4.14, the data collector's payoff after accepting $Of = \langle \delta, p \rangle$ will be:

$$U_{DC} = p \cdot EN(\beta_0 + \beta_1 \cdot h(\delta) + \beta_2 \cdot g(I)) - I \cdot [\beta_0 + \beta_1 \cdot h(\delta) + \beta_2 \cdot g(I)] - C \tag{4.27}$$

In this function $EN(N)$ is a function of the expected cardinality and denotes the effective cardinality of the dataset from the data user's perspective (after considering the Max and Min limits).

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

$$\hat{I} = \hat{i}(\delta, p) = \arg\max_I U_{DC} = \arg\max_I \; p \cdot EN(N(\delta, I)) - I \cdot N(\delta, I) - C \tag{4.28}$$

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 or rejecting and in the case where $\hat{U}_{DC} < 0$ the data collector rejects. Therefore, the data collector's best response, BR_{DC} is a function that associates with each offer $Of = \langle \delta, p \rangle$ the following action:

$$BR_{DC}(\langle \delta, p \rangle) = \begin{cases} \{Reject\} & if \; p \cdot EN(N(\hat{I})) - \hat{I} \cdot N(\hat{I}) - C < 0 \\ \{Reject, Accept\} & if \; p \cdot EN(N(\hat{I})) - \hat{I} \cdot N(\hat{I}) - C = 0 \\ \{Accept\} \; with \; \hat{I} & if \; p \cdot EN(N(\hat{I})) - \hat{I} \cdot N(\hat{I}) - C > 0 \end{cases} \tag{4.29}$$

In this equation $N(\hat{I})$ is the expected cardinality of the dataset when values δ and $\hat{I}$ are plugged into the Equation 4.26.

The optimum incentive $\hat{I}$ must only be calculated when the data collector accepts the offer. This means $\hat{I} \leq p$, otherwise $\hat{U}_{DC} < 0$ because $EN(N) \leq N$ and $0 \leq C$. Since U_{DC} is continuous in the closed and bounded interval $[0, p]$ (the domain of I), according to the Extreme Value Theorem [SH95], U_{DC} reaches its maximum at least once and therefore $\hat{I}$ is guaranteed to exist.

4.3.3 Data User's Best Response

The last step to find the subgame perfect equilibria is to find the most profitable action of the data user; knowing the data collector and data providers' best responses (see Sections 4.3.1 and 4.3.2) to each $Of = \langle \delta, p \rangle$, what combination of δ and p maximizes the data user's payoff?

For each offer Of if the data collector's maximum payoff is zero she is indifferent between accepting the offer (and choosing the optimum incentive $\hat{I}$) and rejecting it. Both of these actions are data collector's best response to offer Of. To find *all* subgame perfect equilibria of the game, we need to trace back the implications of every possible combination of data collector's best responses (to all possible offers) separately. In other words, the data user must find a best response to every optimum strategy of the data collector (composed of one of her best responses to each possible offer). Since the number of possible offers $|OF|$ can be very large, the data collector can have significantly many optimum strategies (for example, if we assume two best responses for each offer, then there will be $2^{|OF|}$ different optimum strategies for the data collector.). In this thesis, we only analyze those equilibria that correspond to the data user's most profitable outcomes. In these equilibria, the data user is optimistic and believes that whenever the data collector is indifferent between accepting and rejecting an offer, she always accepts. Nevertheless, the game can have many other equilibria in which the data collector sometimes chooses to reject when she is indifferent.

The implications of these equilibria of the game are not studied in this thesis and can be considered as a future extension.

When the data collector accepts an offer $Of = \langle \delta, p \rangle$, she chooses the optimum incentive $\hat{I}$. Depending on the exact function definitions used in Equation 4.26 and Equation 4.28, if $\hat{I}$ is unique for every combination of δ and p, then $\hat{I}$ can be defined as a function of δ and p (*i.e.*, $\hat{I} = \hat{i}(\delta, p)$). Without loss of generality, we assume that this is the case. If multiple values of I maximize U_{DC}, then data collector can potentially have a different best response to each combination of data user's best strategy (with each maximizing $\hat{I}$) and each of these resulting strategy profiles is a subgame perfect equilibrium of the game.

According to Section 4.3.2, if the data collector accepts the offer she starts collecting personal information with the privacy parameter δ and incentive $\hat{I} = \hat{i}(\delta, p)$. Otherwise, no dataset will be provided to the data user. As a result, the expected number of records N can be determined as:

$$N = z(\delta, \hat{I}) = \begin{cases} n \cdot [\beta_0 + \beta_1 \cdot h(\delta) + \beta_2 \cdot g(\hat{I})] & if\ \hat{U}_{DC} \geq 0 \\ \\ 0 & Otherwise \end{cases} \tag{4.30}$$

Plugging the function definition of $\hat{I} = \hat{i}(\delta, p)$ into Equation 4.30, $N = z_2(\delta, p)$ becomes a function of δ and p as well. Recall (from Section 4.2.1.2) that $Precision(\delta, N)$ is a function of δ and N. Since N is a function of δ and p, we can define $Precision(\delta, N) = prec(\delta, p)$ and $EN(N) = en(\delta, p)$ as functions of δ and p as well. After substituting N, EN, and $Precision$ with $z_2(\delta, p)$, $en(\delta, p)$, and $prec(\delta, p)$, the U_{DU} function from Equation 4.11 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 her payoff:

$$\langle \hat{\delta}, \hat{p} \rangle = \arg\max_{\delta,p} U_{DU} = \arg\max_{\delta,p} \left[b \cdot \ prec_2(\delta, p) - p \right] \ \cdot \ en(\delta, p) \tag{4.31}$$

By definition, the lower bound on p is zero, *i.e.*, $p \geq 0$. Moreover, since $Precision \leq 1$ then $b \cdot \ prec(\delta, p) \leq b$. Choosing a value $p > b$ leads to a negative payoff to the data user and she

can always do better by choosing $p = 0$ (which leads to payoff zero)[7]. Therefore, the upper bound for p is b. Parameter δ is not necessarily bounded (although in some privacy protection mechanisms there are implicit bounds on the privacy parameter values). 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 form $\langle(\hat{\delta}, \hat{p}), br_{DC}, br_{DP_1}, br_{DP_2}, \cdots, br_{DP_n}\rangle$. In this strategy profile, br_{DC} is a function that returns one of the data collector's best response actions to each offer (δ, p). Strategy br_{DP_i} is the function that returns the i^{th} data provider's mixed strategy in response to each combination of (δ, I).

If no equilibrium $s^* = \langle(\hat{\delta}, \hat{p}), br_{DC}, br_{DP_1}, br_{DP_2}, \cdot \cdot \cdot, br_{DP_n}\rangle$ can be found such that $br_{DC}(\hat{\delta}, \hat{p}) \neq Reject$ (*i.e.*, the best strategy of the data user is always an offer that is going to be rejected) then the negotiation of the game is considered unsuccessful. Such games can distinguish *impractical* privacy protection mechanisms given the problem settings. If the cost of implementing a privacy protection mechanism is too high and data providers' trust in the method is not high enough, the game might become an instance of unsuccessful negotiations and we have a case of an impractical privacy protection method.

4.4 The Game's Abstractions and Assumptions

With the proposed model, we are aiming at capturing the essence of the strategic situation without directly incorporating those details that seem less relevant. To reach this aim, we introduced parameters and assumptions to our game model. In the following, we discuss the implications of these parameters and assumptions.

Parameters such as the economic value of each record to the data user (denoted by b),

[7]In any instance of the game, if the data user makes an offer with $p = 0$, regardless of other parameter settings, the best response of the data collector would be *Reject* because choosing any value of $I \geq 0$ leads to a payoff less than zero to the data collector. A a result, in any equilibrium of the game the data collector and the data user's payoffs are at least zero.

the data collection, storage and protection cost C, and the distribution of different types of data providers encapsulate our game model from those external factors that add unnecessary details to our model. In fact, the values of parameters b and C and the characterizations of different data provider types can be functions of other external parameters. Once these values and the behavioural trends of data providers are determined they can be plugged into the game's description to define an instance of the game.

To reach the probability model in Equation 4.25, we have assumed a uniform distribution for thresholds of privacy pragmatists and privacy fundamentalists. This assumption leads to a probability model with only a few parameters to explain data providers' privacy behavior. However, any other distribution that leads to a regression model linear in $h(\delta)$ and $g(I)$ is also compatible with our model and our game analysis does not change if different values are found for coefficients in data providers' privacy behavior.

We have also assumed that the probability distribution over possible types of data providers is a common knowledge of the game. More specifically, we assume that both the data user and the data collector assign the same probability distribution to different types for a data provider. In analyzing games with incomplete information this is not an unusual assumption. As Harsanyi [Har68a] shows, the process of analyzing games with incomplete information (by finding the Bayes-equivalent of the game) promotes such similar beliefs. This is due to the fact that each player must create a universal probability distribution (for all possible type assignments to the players) from the point of view of an outside observer and independent of his own prejudice.

The games where the data user and the data collector have different beliefs about the probability distributions over data providers' possible types can be an instance of an inconsistent game (see Section 3.1.4). For instance, if the conditional probability distribution that the data user assigns to types of data providers is different from the conditional probability distribution that *she believes* the data collector is assigning to different types of data

providers, then the game is inconsistent and a Bayes equivalent model cannot capture all the intricacies of the game. Solving such games are beyond the scope of this work and should be considered as a possible extension to this project.

4.5 Summary

In this chapter, we defined the problem of balancing the tradeoff of privacy and utility as finding stable values for privacy parameters. To find the stable privacy parameter values, a generic game model is proposed. Players of the game are a data user, a data collector, and n data providers who interact in an extensive form game with incomplete and imperfect information. We consider three base types for the data providers. Two of these types have many subtypes each with a different minimum required level for privacy/incentive gain.

The game's solution is explained at an abstract level based on the generic backward induction method. The process of solving the game starts by finding each data provider's best response as a mixed strategy. Their mixed strategies explain the probability of the event that a random data provider opts in for a combination of privacy parameter value and incentive. Based on the best response of the data providers, the data collector's best response can be found. This best response is the data collector's optimum choice of incentive for each offer she receives from the data user. Finally, the data user's optimum offers are found according to the best responses of the data collector and the data providers.

In the next two chapters, we show how the generic game model can be instantiated for various privacy protection methods and provide a concrete game analysis in each case.

Chapter 5

Game Model Instantiation for Privacy Policy Declaration

The generic game model (explained in Chapter 4) can be used to find stable privacy settings in both privacy policy declaration and data sanitization approaches (see Chapter 2). In this chapter, we target the privacy policy declaration approach and describe a game to find stable privacy policy settings when data is collected for answering aggregate queries. We explain the game's subgame perfect equilibria as a choice among a subset of fifteen options (based on the problem setting). Then we consider a simplified scenario to further analyze these options and partition the problem space into 22×22 disjoint classes. The equilibria of the game are explained separately for each class. Finally, with the help of two case studies we illustrate the usage and implications of the game's equilibria.

5.1 Privacy Policy Declaration Overview

A common approach to address privacy concerns involved in collecting personal data is to declare a privacy policy and seek for data providers' consent before collecting their information. The specification of a privacy policy clarifies the scope and limitations of data usage with a structured format. As explained in Chapter 2, various privacy policy languages have been proposed so far. In this thesis, we adopt a language close to what P3P [CLM+02b] and at least one privacy taxonomy [BAB+09] suggest. The main idea behind these two languages is that for each piece of collected information (data field), purpose, visibility (recipient), retention period, and granularity levels must be specified. Considering a private data table $T(A_1, \cdots, A_m)$, we refer to each attribute A_i as a data field. Purpose and visibility are

represented as strings to describe the purpose of data usage and the data user, respectively. Retention period is a number which specifies for how long the data would be stored in the system. The granularity level is adopted from the privacy taxonomy [BAB+09]. It determines how specific and accurate the value of a data field would be when it is revealed to a data user[1]. The set of available options for granularity levels must be expressive and easy to understand at the same time. Here, we suggest seven different levels of granularity and continue the discussion based on these levels. However, as long as the levels are specified clear enough so that data providers and data users can develop some sense of preference over each level, any other hierarchy of granularity levels can be used in our game model. Languages with more options for granularity levels induce more cases to be considered in the course of solving the game. But the procedure of finding the game's solution remains the same.

The hierarchy of our granularity levels (from the least specific to the most specific level) are explained as follows:

(0)-None: No information on the data field is provided.

(1)-Unlinkable, Partial: The value of a data field cannot be linked to values of other data fields provided by the same individual (Unlinkable) and the value of the data field is generalized or perturbed with some noise.

(2)-Unlinkable, Exact: The data field is not linkable but the value of the field is revealed in the exact form.

(3)-Linkable, non-identifiable, Partial: the data field is linkable to all other linkable data fields. A sanitization method is used to anonymize data. The exact value of the data is not revealed.

(4)-Linkable, non-identifiable, Exact: It is the same as as level (3), except that the exact value of the data field is revealed.

[1]To provide an option for hiding unnecessary information from an *authorized* data user and prevent data misuse, the notion of data granularity must be somehow addressed in every privacy policy language.

(5)-Linkable, identifiable, Partial: It is the same as level (3) with no sanitization.
(6)-Linkable, identifiable, Exact: It is the same as level (4) with no sanitization.

Let $DF = \{A_1, \cdots, A_m\}$, Pur, V, R, and $Gr = \{0, 1, 2, 3, 4, 5, 6\}$ denote the sets of all possible data fields, purposes, visibilities, retentions, and granularity levels. A privacy statement ps can be defined as follows:

$$ps \in DF \times Pur \times V \times Gr \times R \tag{5.1}$$

Consequently a privacy policy PP can be defined as a set of privacy statements:

$$PP \subset DF \times Pur \times V \times Gr \times R \tag{5.2}$$

The set of privacy statements is usually chosen by the data collector according to data requirements of data users interested in the dataset. Once the data collector publishes a privacy policy PP, the data providers have the choice of opting in or out for the statements. To provide a semantically consistent functionality of opt-in/opt-out options, we consider *statement groups* in the sense that if a group of privacy statements share the same purpose and visibility, either all or none of the statements in the group must be accepted. Some privacy policy languages such as P3P [CDE+06] already support this idea via the "consent attribute" in the statement group construct. Since a statement group represents a collection of statements necessary to accomplish a certain task by a data user, it is reasonable to expect that all data fields requested in the same statement group expire at the same time. Therefore, we assume that all statements in a statement group have the same value for their retention parameter, R. This assumption is similar to the retention tag ⟨`stated-purpose`⟩ in P3P.

5.2 Game Model for Privacy Policy Settings

Privacy statements chosen for a privacy policy determine the level of privacy protection. The statements affect data quality and decisions of the data providers, differently. Therefore, in

privacy policy declaration approach, the whole privacy policy, PP, can be considered as the privacy parameter. However, as mentioned in Chapter 4, we only solve the games with a single data user. Therefore, in each instance of the game every privacy statement $ps \in PP$ has the same value for the visibility parameter. Moreover, we assume that the data collector responds independently to data requests (of the data user) with different purposes. With this assumption, we only consider a single purpose for the data user since any multi-purpose case can be described as an aggregation of single purpose independent cases[2].

Since we are modelling the game with a single data user/single purpose, in each instance of the game, the challenge of privacy policy setting is reduced to choosing a *single* statement group (see Section 5.1). This statement group is a set of privacy statements with pre-defined values for parameters Pur and V. All statements in this statement group share the same value for the retention parameter, R, but are defined over different data fields each with an arbitrary granularity level. Therefore, in each instance of the game, granularity levels for each data field and the retention period are the privacy components on which the players need to decide. Consequently, in a data table $T(A_1, \cdots, A_m)$ we can define the privacy parameter δ as $\delta = \langle g_1, g_2, \cdots, g_m, r \rangle$, where g_i is the granularity level of data field (*i.e.*, attribute) A_i and r is the retention period (in years) of each data record. Even though values of purpose and visibility can affect the decisions made by players, we do not include them in the definition of δ since they are fixed in each instance of the game. We assume that these fixed values are common knowledge to all players.

We tailor the generic game model to the specific needs of the privacy policy declaration approach. The game's ingredients and rules of the game are very similar to what we explained

[2]Note that if the same data user requests the same data for two different purposes with different granularity levels, there is a potential of inferring additional information by combining the two versions of data. This situation can be modeled as two separate games and in the second game, the data has a higher economic value to the data user. A more realistic and complicated approach would be modelling such cases with a single repeated game. Solving the problem using the latter approach is a future work we are interested in.

in Chapter 4. But, some of the function definitions are specific to each particular instantiation of the game. In this section, we mainly focus on providing those details that must be especially defined when privacy policy declaration is chosen as the method to protect privacy.

5.2.1 Players and Payoffs

Players of the game are n data providers DP, a data user DU, and a data collector DC.

5.2.1.1 Data Providers

Data providers are the potential donors of private information. Each data provider reads the privacy policy PP (which is a single statement group in each instance of the game) and if she agrees to the policy, provides her information and receives the promised reward (incentive).

If a data provider is privacy unconcerned, she always provides the required information regardless of the privacy policy. But a privacy pragmatist and a privacy fundamentalist (see Chapter 2) take the guaranteed level of privacy protection and the incentive into consideration when they are making a decision. To model this trade-off analysis, the two functions of $h(\delta)$ and $g(I)$ (see Section 4.2.1.1) must be defined in the context of privacy policy declaration. As explained in Section 4.3.1, in this thesis we always assume that the rewarding effects of incentive is explained by the identity function. Therefore, we assume $g(I) = I$. The privacy parameter $\delta = \langle g_1, g_2, \cdots, g_m, r \rangle$ has $m + 1$ components. Each of these components is a utility enhancer because as its value increases, privacy decreases (see Section 4.1). As a result, the privacy gain provided by granularity of each data field (or the retention component) must be defined as a strictly decreasing function of the granularity levels (or the retention component). Let function $h_g(\cdot)$ explain the privacy gain provided by the value of a granularity component and function $h_r(\cdot)$ capture the privacy gain caused by the value chosen for the retention component. With these two functions, we define the privacy gain offered by δ to be the weighted sum of the privacy gains provided by each component of the

privacy parameter vector. In other words, we define $h(\delta)$ as:

$$h(\delta) = \lambda_1 \cdot h_g(g_1) + \lambda_2 \cdot h_g(g_2) + \cdots + \lambda_m \cdot h_g(g_m) + \lambda_{m+1} \cdot h_r(r) \tag{5.3}$$

The weights $\lambda_1, \cdots, \lambda_{m+1}$ $(0 < \lambda_i)$ reflect the general sensitivity of each data field (attribute) and the retention component. As we see in Section 5.3.1, functions $h(\delta)$ and $g(I)$ will be used to find the best response of each type of data provider and calculate the expected cardinality of the data table.

5.2.1.2 Data User

A data user is the recipient of the private data table. She pays for each record and practices data analysis (for a specific purpose) according to the privacy policy. As explained in Chapter 4, the data user initiates the game with an offer $Of = \langle \delta, p \rangle = \langle \langle g_1, \cdots, g_m, r \rangle, p \rangle$.

Considering the maximum and minimum number of required records (Min and Max)[3], a collected data table with cardinality N has the effective cardinality of $EN(N)$ to the data user (see Equation 4.8 from Chapter 4). Consequently, if the data user pays price p for each record, she spends $EN(N) \cdot p$ to access a data table of cardinality N.

We consider situations where aggregate queries with the following `COUNT`-query format are used for data analysis:

`SELECT COUNT(*) FROM T WHERE Pred(`A_j`)`

In this query `Pred` is a predicate defined on the data field A_j. With this query, the data user only needs to know values of attribute A_j.[4] If these values are provided at one of the *exact* granularity levels, then the result set of the query would be 100% accurate. If one of the *partial* granularity levels is applied to attribute A_j, then depending on the amount of

[3]Parameters Min and Max are part of the problem definition.

[4]For this query the only data granularity factor that affects the data user's payoff function is whether the value of the requested data field, A_j, is provided at the exact or partial level. Therefore, if a simpler model for the granularity axis is adapted with only partial and exact levels, the game analysis and results remain the same.

distortion and how aligned the values are with the `COUNT`-query predicate, the result have an accuracy of ξ, where $0 \leq \xi < 1$. Therefore, we simply define the *Precision* function as:

$$Precision(\delta) = \begin{cases} 0 & if\ g_j = 0 \\ \xi & if\ g_j \in \{1,3,5\} \\ 1 & if\ g_j \in \{2,4,6\} \end{cases} \tag{5.4}$$

Let b denote the economic value of each record to the data user for each year that the record is stored in the system. The data user's income can be defined as $EN(N) \cdot b \cdot Precision(\delta) \cdot r$. Subtracting data user's expenditure from her income, we can define her payoff as:

$$U_{DU} = \begin{cases} 0 - EN(N) \cdot p & if\ g_j = 0 \\ EN(N) \cdot [b \cdot \xi \cdot r - p] & if\ g_j \in \{1,3,5\} \\ EN(N) \cdot [b \cdot r - p] & if\ g_j \in \{2,4,6\} \end{cases} \tag{5.5}$$

Let a represent the economic value of a record when values of data field A_j are provided at a partial granularity level (*i.e.*, $a = b \cdot \xi$). Using this notation and the definition of function $EN(N)$ (from Equation 4.8), we re-write the U_{DU} function as:

$$U_{DU} = \begin{cases} 0 & if\ N < Min & \\ -N \cdot p & if\ Min \leq N \leq Max & \wedge g_j = 0 \\ N \cdot [a \cdot r - p] & if\ Min \leq N \leq Max & \wedge g_j \in \{1,3,5\} \\ N \cdot [b \cdot r - p] & if\ Min \leq N \leq Max & \wedge g_j \in \{2,4,6\} \\ -Max \cdot p & if\ Max < N & \wedge g_j = 0 \\ Max \cdot [a \cdot r - p] & if\ Max < N & \wedge g_j \in \{1,3,5\} \\ Max \cdot [b \cdot r - p] & if\ Max < N & \wedge g_j \in \{2,4,6\} \end{cases} \tag{5.6}$$

5.2.1.3 Data Collector

A data collector is the trusted party who collects personal information from data providers and provides it to the data user under the terms and conditions of the privacy policy. During the game, the data collector must choose between accepting or rejecting data user's offer. If she decides to accept, she must choose some incentive I to reward the data providers with.

As explained in Section 4.2.1.3, the data collector receives payment from the data user (for each collected record) but needs to pay the data providers and allocate some budget, C, for storing the data table, applying data manipulation procedures, and enforcing the privacy policy. As a result, the data collector's payoff (in case of an acceptance) can be defined as $U_{DC} = EN(N) \cdot p - N \cdot I - C$.

We can break down the cost C into smaller components based on the granularity level of each data field. Let G be the cost of applying a generalization (or perturbation) method to make the value of a data field partial and A represent the cost of anonymization. The cost of providing data field A_i to the data user at granularity level g_i can be defined as:

$$CG(g_i) = \begin{cases} 0 & if\ g_i = 0\ \vee\ g_i = 2\ \vee\ g_i = 6 \\ G & if\ g_i = 1\ \vee\ g_i = 5 \\ A + G & if\ g_i = 3 \\ A & if\ g_i = 4 \end{cases} \tag{5.7}$$

Besides the costs of generalization and anonymization we need to consider a base cost B for the expense of storing the data table and enforcing the privacy policy. The cost of providing the data user with the collected data table in response to offer o can be calculated as:

$$C_o = \sum_{i=1}^{m} CG(g_i) + B \tag{5.8}$$

Using the definitions of C_o and function $EN(N)$ (from Equation 4.8) we can rewrite the payoff to the data collector in the *acceptance case* as follows:

$$U_{DC} = \begin{cases} 0 - N \cdot I - C_o & if\ N < Min \\ N \cdot [p - I] - C_o & if\ Min \leq N \leq Max \\ Max \cdot p - N \cdot I - C_o & if\ Max < N \end{cases} \tag{5.9}$$

5.3 Subgame Perfect Equilibria in Privacy Policy Settings

To find the game's subgame perfect equilibria, we follow the principle of backward induction. As discussed in Section 4.3, we first start by finding data providers' best response. Based

on their best response we can calculate the expected cardinality of the data table for each combination of privacy policy and incentive. We then find the optimal actions of the data collector, given data providers' behavior in the rest of the game. The last step is to find the data user's most profitable offers considering the best responses of the data collector and data providers.

5.3.1 Data Providers' Best Response

Data collection subgames are the smallest subgames where all data providers simultaneously and independently make decisions based on the promised privacy policy and incentive. While some types of data providers (privacy unconcerned) choose their action regardless of the privacy parameter δ, other types (privacy pragmatists and privacy fundamentalists) consider the effects of privacy parameter δ and incentive I before making their decision. After considering all of these types and their trade-offs, in Section 4.3.1 we showed how to find the best response of a randomly chosen data provider as the probability of opting in for a combination of δ and I. According to Equation 4.24 this probability is as follows:

$$Prob(OptIn) = p_1 + p_2 \cdot \frac{w_1 \cdot h(\delta) + w_2 \cdot g(I)}{\overline{v}} + p_3 \cdot \frac{h(\delta)}{\overline{v}'} \tag{5.10}$$

In this formula, p_1, p_2, and p_3 are the probabilities of the events that a data provider is a privacy unconcerned, privacy pragmatist, and privacy fundamentalist. parameters w_1 and w_2 represent the weights of privacy gain and incentive on a privacy pragmatist's tradeoff decision. $\overline{v} = \max_{\delta,I}\{w_1 \cdot h(\delta) + w_2 \cdot g(I)\}$ and $\overline{v}' = \max_{\delta}\{h(\delta)\}$.

Let $\beta_0 = p_1$, $\beta_1 = [p_2 \cdot w_1]/\overline{v} + p_3/\overline{v}'$ and $\beta_2 = [p_2 \cdot w_2]/\overline{v}$. When privacy policy declaration is chosen as the privacy protection method, privacy gain $h(\delta)$ is defined by Equation 5.3 and the rewarding effect of incentive is assumed to be $g(I) = I$. Therefore, we have:

$$Prob(OptIn) = \beta_0 + \beta_1 \cdot [\lambda_1 \cdot h_g(g_1) + \cdots + \lambda_m \cdot h_g(g_m) + \lambda_{m+1} \cdot h_r(r)] + \beta_2 \cdot I \tag{5.11}$$

Using the transformations $\tau_0 = \beta_0$, $\tau_i = \beta_1 \cdot \lambda_i$ (for $i \in \{1, \cdot\cdot, m\}$), $\theta = \beta_1 \cdot \lambda_{m+1}$, and

$\gamma = \beta_2$, we can re-write Equation 5.11 as:

$$Prob(OptIn) = \tau_0 + \tau_1 \cdot h_g(g_1) + \tau_2 \cdot h_g(g_2) + \cdots + \tau_m \cdot h_g(g_m) + \theta \cdot h_r(r) + \gamma \cdot I \quad (5.12)$$

Since we are assuming that data providers make their decisions independently, the expected cardinality of a dataset with n data providers can be defined as $N = n \cdot Prob(OptIn)$.

5.3.2 Data Collector's Best Response

The data collector's best response is her optimum reaction to each offer $o = \langle\langle g_1, \cdots, g_m, r\rangle, p\rangle$ given data providers' behavior in the data collection subgames. In response to the data user's offer, she can either reject or accept and announce some incentive.

As explained in Section 4.3.2, the data collector must find the optimum value of I that maximizes her payoff U_{DC} from Equation 5.9. To simplify the notation, let α_o denote the probability of a data provider opting in for offer o with *zero incentive*. In other words:

$$\alpha_o = \tau_0 + \tau_1 \cdot h_g(g_1) + \tau_2 \cdot h_g(g_2) + \cdots + \tau_m \cdot h_g(g_m) + \theta \cdot h_r(r) \quad (5.13)$$

The values of g_i and r are plugged in from the specifications of offer o. Since we are assuming n potential data providers, the expected number of data providers who will opt-in for a privacy policy according to offer o and incentive I can be calculated as:

$$N = n \cdot [\alpha_o + \gamma \cdot I] \quad (5.14)$$

When N is plugged into Equation 5.9 and the conditions are re-arranged to be defined based on I, we can restate U_{DC} as follows:

$$U_{DC} = \begin{cases} 0 - n \cdot [\alpha_o + \gamma \cdot I] \cdot I - C_o & if\ I < \frac{Min - n \cdot \alpha_o}{n \cdot \gamma} \\ n \cdot [\alpha_o + \gamma \cdot I] \cdot [p - I] - C_o & if\ \frac{Min - n \cdot \alpha_o}{n \cdot \gamma} \leq I \leq \frac{Max - n \cdot \alpha_o}{n \cdot \gamma} \\ Max \cdot p - n \cdot [\alpha_o + \gamma \cdot I] \cdot I - C_o & if\ \frac{Max - n \cdot \alpha_o}{n \cdot \gamma} < I \end{cases} \quad (5.15)$$

Since we can safely assume that $\alpha_o > 0$ the first piece of the U_{DC} function always yields a negative payoff and therefore the choice of an incentive $I < \frac{Min - n \cdot \alpha_o}{n \cdot \gamma}$ is always dominated

Table 5.1: Data collector's best response in **Case1:** $n \cdot \alpha_o < Min$

	Condition	Best incentive	Best action†	Maximum payoff	Dataset Cardinality
a	$\frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma} < \frac{Min - n \cdot \alpha_o}{n \cdot \gamma}$	$I = \frac{Min - n \cdot \alpha_o}{n \cdot \gamma}$	Accept if $\hat{U}^{1a}_{DC} > 0$ Reject if $\hat{U}^{1a}_{DC} < 0$ Indifferent otherwise	$\max\{\hat{U}^{1a}_{DC}, 0\}$	Min if Accept 0 if Reject
b	$\frac{Min - n \cdot \alpha_o}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma} \leq \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}$	$I = \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}$	Accept if $\hat{U}^{1b}_{DC} > 0$ Reject if $\hat{U}^{1b}_{DC} < 0$ Indifferent otherwise	$\max\{\hat{U}^{1b}_{DC}, 0\}$	$n \cdot \left\lfloor \frac{\alpha_o + \gamma \cdot p}{2} \right\rfloor$ if Accept 0 if Reject
c	$\frac{Max - n \cdot \alpha_o}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}$	$I = \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}$	Accept if $\hat{U}^{1c}_{DC} > 0$ Reject if $\hat{U}^{1c}_{DC} < 0$ Indifferent otherwise	$\max\{\hat{U}^{1c}_{DC}, 0\}$	Max if Accept 0 if Reject

† Utilities are defined as:
$\hat{U}^{1a}_{DC} = Min \cdot \left[p - \frac{Min - n \cdot \alpha_o}{n \cdot \gamma}\right] - C_o,\ \hat{U}^{1b}_{DC} = \frac{n}{\gamma} \cdot \left[\frac{\alpha_o + \gamma \cdot p}{2}\right]^2 - C_o,\ \hat{U}^{1c}_{DC} = Max \cdot \left[p - \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}\right] - C_o.$

by at least the reject action. To find the maximum of the second and the third pieces of the function we find the maximizing I in each piece by setting the derivative of the piece to zero. The maximizing values of I must be within the upper and lower bound criteria of the piece. Moreover, we must verify that maximizing incentives are not less than zero.

To simplify the comparisons and organize the results, we consider three cases: case 1 happens when the lower bound of I in the second piece is greater than zero, case 2 happens when this lower bound is less than zero but the upper bound of I in the second piece is greater than zero, and case 3 occurs when the upper bound of I in the second piece is less than zero. When the data collector receives an offer o, regardless of her choice for incentive value I, only one of the three cases would apply. The applicable case can be distinguished based on contents of offer o and expected cardinality of the data table when no incentives promised.[5]

5.3.2.1 Case 1: $n \cdot \alpha_o < Min$

This case happens when the anticipated number of data providers who opt-in for the policy *with no incentive* is less than Min.

If the maxima of either the second or the third piece of the U_{DC} function in Equation 5.15 is greater than zero then the data collector decides to accept and the number of data

[5]These three cases cover the space of all possible values for upper and lower bounds on values of I in the second piece of Equation 5.15.

Table 5.2: Data collector's best response in **Case 2:** $Min \leq n \cdot \alpha_o \leq Max$

	Condition	Best incentive	Best action†	Maximum payoff	Dataset Cardinality
a	$\frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma} < 0$	$I = 0$	Accept if $\hat{U}^{2a}_{DC} > 0$ Reject if $\hat{U}^{2a}_{DC} < 0$ Indifferent otherwise	$\max\{\hat{U}^{2a}_{DC}, 0\}$	$n \cdot \alpha_o$ if Accept 0 if Reject
b	$0 \leq \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma} \leq \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}$	$I = \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}$	Accept if $\hat{U}^{2b}_{DC} > 0$ Reject if $\hat{U}^{2b}_{DC} < 0$ Indifferent otherwise	$\max\{\hat{U}^{2b}_{DC}, 0\}$	$n \cdot \left[\frac{\alpha_o + \gamma \cdot p}{2}\right]$ if Accept 0 if Reject
c	$\frac{Max - n \cdot \alpha_o}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}$	$I = \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}$	Accept if $\hat{U}^{2c}_{DC} > 0$ Reject if $\hat{U}^{2c}_{DC} < 0$ Indifferent otherwise	$\max\{\hat{U}^{2c}_{DC}, 0\}$	Max if Accept 0 if Reject

† Utilities are defined as:
$\hat{U}^{2a}_{DC} = n \cdot \alpha_o \cdot p - C_o$, $\hat{U}^{2b}_{DC} = \frac{n}{\gamma} \cdot \left[\frac{\alpha_o + \gamma \cdot p}{2}\right]^2 - C_o$, $\hat{U}^{2c}_{DC} = Max \cdot \left[p - \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}\right] - C_o$.

Table 5.3: Data collector's best response in **Case 3:** $Max < n \cdot \alpha_o$

Condition	Best incentive	Best action†	Maximum payoff	Dataset Cardinality
none	$I = 0$	Accept if $\hat{U}^{3}_{DC} > 0$ Reject if $\hat{U}^{3}_{DC} < 0$ Indifferent otherwise	$\max\{\hat{U}^{3}_{DC}, 0\}$	$n \cdot \alpha_o$ if Accept 0 if Reject

† Utility is defined as: $\hat{U}^{3}_{DC} = Max \cdot p - C_o$

records can be calculated based on the maximizing incentive. If the maximum of these two pieces is equal to zero then both accepting and rejecting are optimum choices for the data collector. As explained in Chapter 4, in this thesis we only study those equilibria in which the data collector accepts when she is indifferent.

To find the maximum of the second piece of U_{DC} function, its derivate (with respect to I) must be set to zero:

$$\frac{dU_{DC}}{dI} = -n \cdot \alpha_o - n \cdot \gamma \cdot I + n \cdot \gamma \cdot p - n \cdot \gamma \cdot I = 0 \Rightarrow \hat{I} = \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma} \tag{5.16}$$

Since the second derivative of the function is negative, the function is concave down and thus $\hat{I}$ represents a local maximum. Depending on the value of $\hat{I}$ the following three sub-cases can occur:

(a) If $\hat{I} < \frac{Min - n \cdot \alpha_o}{n \cdot \gamma}$, the second piece of function reaches its maximum at an incentive which is less than the lower bound.Therefore, the maximum happens at the beginning of the interval (*i.e.*, $I = \frac{Min - n \cdot \alpha_o}{n \cdot \gamma}$). In this case the maximum payoff to the data collector (if

she accepts), denoted by $\hat{U}_{DC}^{1a}$, is calculated as:

$$\hat{U}_{DC}^{1a} = Min \cdot \left[p - \frac{Min - n \cdot \alpha_o}{n \cdot \gamma} \right] - C_o \tag{5.17}$$

(b) If $\frac{Min - n \cdot \alpha_o}{n \cdot \gamma} \leq \hat{I} \leq \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}$, then the maximum happens at $\hat{I}$. In this case the maximum payoff to the data collector (if she accepts), denoted by $\hat{U}_{DC}^{1b}$, is calculated as:

$$\hat{U}_{DC}^{1b} = \frac{n}{\gamma} \cdot \left[\frac{\alpha_o + \gamma \cdot p}{2} \right]^2 - C_o \tag{5.18}$$

(c) If $\frac{Max - n \cdot \alpha_o}{n \cdot \gamma} \leq \hat{I}$, the second piece of function reaches its maximum at an incentive which is greater than the upper bound. Therefore, the maximizing value (that is also within the boundaries) is the end of the interval (*i.e.*, $I = \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}$). In this case the maximum payoff to the data collector (if he accepts), denoted by $\hat{U}_{DC}^{1c}$, is calculated as:

$$\hat{U}_{DC}^{1c} = Max \cdot \left[p - \frac{Max - n \cdot \alpha_o}{n \cdot \gamma} \right] - C_o \tag{5.19}$$

The maximum of the third piece of the function can be determined by finding its derivative with respect to I and setting it to zero:

$$\frac{dU_{DC}}{dI} = -n \cdot \alpha_o - n \cdot \gamma \cdot I - n \cdot \gamma \cdot I = 0 \Rightarrow \hat{I} = -\frac{\alpha_o}{2 \cdot \gamma} < 0 < \frac{Max - n \cdot \alpha_o}{n \cdot \gamma} \tag{5.20}$$

Since the second derivative is negative, $\hat{I}$ is a local maximum. However, as Equation 5.20 shows, $\hat{I}$ is less than the beginning of the interval and the maximizing incentive will be $I = \frac{Max - n \cdot \alpha_o}{n \cdot \gamma}$. Plugging this incentive in the third piece of the payoff function we receive the same maximum payoff as U_{DC}^{1c} in Equation 5.19. The best strategies in case 1 are summarized in Table 5.1.

5.3.2.2 Case 2: $Min \leq n \cdot \alpha_o \leq Max$

This case happens when the anticipated number of data providers who opt-in for the policy *with no incentive* is more than (or equal to) Min but less than or equal to Max.

A reasoning similar to Section 5.3.2.1 can be used to find the local optima of the payoff function U_{DC}. The only difference between the two cases happens when $\hat{I}$ (from Equation

5.16) is negative or less than $\frac{Min - n \cdot \alpha_o}{n \cdot \gamma}$. Contrary to case 1, in this case the maximizing incentive cannot be set to the beginning of the interval since $\frac{Min - n \cdot \alpha_o}{n \cdot \gamma} < 0$. Instead, the incentive should be set to zero. The best strategies in case 2 are summarized in Table 5.2.

5.3.2.3 Case 3: $Max < n \cdot \alpha_o$

In the last case, the number of data providers who are willing to share their information *with no incentive* is already greater than Max. Since the data table is expected to have cardinality of Max or above with zero incentive, increasing the incentive above zero will entice more data providers without adding anything to the data collector's payoff. In this case, the best action of the data collector is to either accept with $\hat{I} = 0$ or reject. With zero incentive, the payoff to the data collector would be $\hat{U}^3_{DC} = Max \cdot p - C_o$. This result is shown in Table 5.3.

5.3.3 Data User's Best Response

The last step in finding the game's subgame perfect equilibria, is to find the data user's optimum choice, taking the data collector and data providers' best responses (see Sections 5.3.1 and 5.3.2) as given. The following three theorems help us to reduce the search space for the data user's best responses.

Theorem 5.3.1. *Let* $o1 = \langle\langle g_1, \cdots, g_j, \cdots, g_m, r\rangle, p\rangle$ *be an offer such that at least one of the* g_i*'s with* $i \neq j$ *is set to a level higher than zero. Recall that* A_j *is the data field over which the predicate of the* `COUNT`*-query is defined. The data user can do at least as good as* $o1$ *by making an offer* $o2 = \langle\langle 0, 0, \cdots, g_j, \cdots, 0, r\rangle, p\rangle$ *or* $o2' = \langle\langle 0, 0, \cdots, g_j, \cdots, 0, r\rangle 0\rangle$.

Proof. (Sketch) Consider the description of α_o given in Equation 5.13. Since all parameters τ_0, ..., τ_m are greater than zero, function $h_g(\cdot)$ is a strictly decreasing function of granularity level, and there is at least one g_i that is zero in $o2$ but more than zero in $o1$, we have $\alpha_{o1} < \alpha_{o2}$. Moreover, as one of the g_i's in $o2$ changes from zero to another granularity level in $o1$, $CG(g_i)$ from Equation 5.7 either increases or stays the same. Thus, the inequality

$C_{o2} \leq C_{o1}$ holds. With these two facts we show that for all meaningful combinations of cases (from Section 5.3.2):

- **part 1 -** If the data collector accepts $o1$ she will also accept $o2$.
- **part 2 -** Cardinality of the data table, N, after accepting offer $o2$ is expected to be at least as large as accepting offer $o1$.
- **part 3 -** By offering $o2$ or $o2'$, the data user's profits are at least as large as her profit when she offers $o1$.

To prove parts 1 and 2, all possible combinations of cases for offers $o1$ and $o2$ must be considered.[6] For each combination, we first explain how accepting both offers $o1$ and $o2$ with the optimum incentive, affect the data collector's maximum payoff and cardinality of the data table. Then we use facts about conditions that apply to the offers and the shape of payoff functions to show that the data collector's payoff after accepting $o2$, is at least as large as her payoff if she accepts $o1$. Here, we only prove parts 1 and 2 for one of the most complicated combinations where Case 1a applies to $o1$ and Case 1b applies to $o2$. The complete proof of all combinations can be found in Appendix A.

Case 1a for *o1* vs. case 1b for *o2* - If the data collector accepts offer $o1$ in case 1a, then her payoff $\hat{U}_{DC}^{1a}$ for offer $o1$ is greater than or equal to zero[7]. In other words:

$$0 \leq Min\left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} = U_{DC}[o1] \tag{5.21}$$

Based on Equation 5.16, the absolute maximum of U_{DC} in case 1 happens at $\hat{I} = \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}$ and this maximum yields $\hat{U}_{DC}^{1b}$. Therefore for every offer, $\hat{U}_{DC}^{1a} \leq \hat{U}_{DC}^{1b}$. We have:

$$Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right]^2 - C_{o1} \tag{5.22}$$

[6]Since $\alpha_{o1} < \alpha_{o2}$ not all combinations of cases apply to offers $o1$ and $o2$. A table of all applicable combination of cases can be found in in the full proof provided in Appendix A.

[7]In the rest of the proof, we denote the data collector's payoff for offer o as $U_{DC}[o]$ and don't include the superscripts. The superscripts can be deduced based on the case that applies to the offer.

Since $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, we have:

$$0 \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right]^2 - C_{o1} \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o2} + \gamma \cdot p}{2}\right]^2 - C_{o2} = \ U_{DC}[o2] \tag{5.23}$$

Therefore, $o2$ will be accepted.

According to Table 5.1 if the data collector accepts an offer in case 1a cardinality of the dataset will be Min and if she accepts an offer in case 1b the expected cardinality would be $n \cdot \left[\frac{\alpha_{o2}+\gamma \cdot p}{2}\right]$. Since Case 1b applies to offer $o2$, the condition for this case is satisfied:

$$\frac{Min - n \cdot \alpha_{o2}}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma} \Rightarrow Min \leq n \cdot \left[\frac{\gamma \cdot p + \alpha_{o2}}{2}\right] \tag{5.24}$$

Consequently, the expected cardinality of the data table after accepting offer $o2$ is at least as large as accepting offer $o1$.

Once parts 1 and 2 are proven for every possible combination of cases that apply to offers $o1$ and $o2$, it is possible to show that offering $o2$ or $o2'$ (an offer with price 0) provides the data user with a payoff at least as large as her payoff for offering $o1$. We prove part 3 by considering three different scenarios as follows:

Scenario 1 - *Both offers get accepted by the data collector:* In part 2 of the proof we showed that if both offers are accepted then cardinality of the data table after accepting offer $o2$ is at least as large as its cardinality after accepting offer $o1$. Since both offers $o1$ and $o2$ have the same values for g_j (granularity level of the attribute A_j) and price, p, the payoff to the data user for offering $o2$ is at least as large as her payoff for offering $o1$.

Scenario 2 - *Offer $o1$ does not get accepted but offer $o2$ gets accepted by the data collector:* In this scenario the data user's payoff for offer $o1$ is zero. The data user can achieve a higher (or the same amount of) profit by choosing to offer $o2$ if such offer provides her with a non-negative payoff or choosing to offer $o2'$ (with a guaranteed payoff of zero), otherwise.

Table 5.4: Potentially optimal payoffs to the data user if $g_j = 1$

case	Maximum payoff†	Subject to
$p1$	$\hat{U}_{DU}^{p1} = Min \cdot [a \cdot \hat{r} - \hat{p}]$	**[c-1]:** $n \cdot \alpha_{op} < Min$ **[c-1a]:** $\alpha_{op} + \gamma \cdot p < \frac{2 \cdot Min}{n}$ **[c-p1]:** $\frac{\gamma \cdot C_{op}}{Min} + \frac{Min}{n} \leq \alpha_{op} + \gamma \cdot p$
$p2$	$\hat{U}_{DU}^{p2} = \frac{n}{2} \cdot [\alpha_{op} + \gamma \cdot \hat{p}] \cdot [a \cdot \hat{r} - \hat{p}]$	**[c-1]:** $n \cdot \alpha_{op} < Min$ **[c-1b]:** $\frac{2 \cdot Min}{n} \leq \alpha_{op} + \gamma \cdot p \leq \frac{2 \cdot Max}{n}$ **[c-p2]:** $2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} \leq \alpha_{op} + \gamma \cdot p$
$p3$	$\hat{U}_{DU}^{p3} = Max \cdot [a \cdot \hat{r} - \hat{p}]$	**[c-1]:** $n \cdot \alpha_{op} < Min$ **[c-1c]:** $\frac{2 \cdot Max}{n} \leq \alpha_{op} + \gamma \cdot p$ **[c-p3]:** $\frac{\gamma \cdot C_{op}}{Max} + \frac{Max}{n} \leq \alpha_{op} + \gamma \cdot p$
$p4$	$\hat{U}_{DU}^{p4} = n \cdot \alpha_{op} \cdot [a\hat{r} - \hat{p}]$	**[c-2]:** $Min \cdot \leq n \cdot \alpha_{op} \leq Max$ **[c-2a]:** $\gamma \cdot p - \alpha_{op} < 0$ **[c-p4]:** $C_{op} \leq n \cdot \alpha_{op} \cdot p$
$p5$	$\hat{U}_{DU}^{p5} = \frac{n}{2} \cdot [\alpha_{op} + \gamma \cdot \hat{p}] [a \cdot \hat{r} - \hat{p}]$	**[c-2]:** $Min \leq n \cdot \alpha_{op} \leq Max$ **[c-2b]:** $0 \leq \gamma \cdot p - \alpha_{op}$ and $\alpha_{op} + \gamma \cdot p \leq \frac{2 \cdot Max}{n}$ **[c-p5]:** $2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} \leq \alpha_{op} + \gamma \cdot p$
$p6$	$\hat{U}_{DU}^{p6} = Max \cdot [a \cdot \hat{r} - \hat{p}]$	**[c-2]:** $Min \leq n \cdot \alpha_{op} \leq Max$ **[c-2c]:** $\frac{2 \cdot Max}{n} \leq \alpha_{op} + \gamma \cdot p$ **[c-p6]:** $\frac{\gamma \cdot C_{op}}{Max} + \frac{Max}{n} \leq \alpha_{op} + \gamma \cdot p$
$p7$	$\hat{U}_{DU}^{p7} = Max \cdot [a \cdot \hat{r} - \hat{p}]$	**[c-3]:** $Max < n \cdot \alpha_{op}$ **[c-p7]:** $C_{op} \leq Max \cdot p$

† $\hat{r}$ and $\hat{p}$ are the values that maximize the payoff in the row subject to the constraints. Parameters α_{op} and C_{op} are defined as:
$\alpha_{op} = \tau_0 + \tau_1 \cdot h_g(0) + \cdots + \tau_j \cdot h_g(1) + \tau_{j+1} \cdot h_g(0) + \cdots + \tau_m \cdot h_g(0) + \theta \cdot h_r(\hat{r})$,
$C_{op} = G + B$.

Scenario 3 - *None of the offers get accepted by the data collector:* In this scenario the data user's payoff for offering $o1$ and $o2$ are both zero. □

Theorem 5.3.2. *Let* $o1 = \langle\langle 0, 0, \cdots, g_j, \cdots, 0, r\rangle, p\rangle$ *be an offer such that* $g_j \in \{3, 5\}$ *where* A_j *is the data field over which the predicate of* `COUNT`*-query is defined. The data user can do at least as good as* $o1$ *by making an offer* $o2 = \langle\langle 0, 0, \cdots, 1, \cdots, 0, r\rangle, p\rangle$ *or* $o2' = \langle\langle 0, 0, \cdots, 1, \cdots, 0, r\rangle 0\rangle$.

Proof. (Sketch) Consider the description of α_o given in Equation 5.13. Since parameter τ_j is greater than zero and $h_g(g_j)$ (privacy gain provided by the granularity level chosen for data field A_j) is defined to be a strictly decreasing function of g_j, we have $\alpha_{o1} < \alpha_{o2}$. Moreover, based on Equation 5.7 we have $CG(1) = CG(5) < CG(3)$. Therefore we can conclude that $C_{o2} \leq C_{o1}$. The rest of the proof is the same as the proof of Theorem 5.3.1. [8] □

[8] Unlike the proof for the first scenario in Theorem 5.3.1, here, offers $o1$ and $o2$ have different values for

Theorem 5.3.3. *Let* $o1 = \langle\langle 0, 0, \cdots, g_j, \cdots, 0, r\rangle, p\rangle$ *be an offer such that* $g_j \in \{4, 6\}$ *where* A_j *is the data field over which the predicate of* `COUNT`*-query is defined. The data user can do at least as good as* $o1$ *by making an offer* $o2 = \langle\langle 0, 0, \cdots, 2, \cdots, 0, r\rangle, p\rangle$ *or* $o2' = \langle\langle 0, 0, \cdots, 2, \cdots, 0, r\rangle 0\rangle$.

Proof. (Sketch) Similar to Theorem 5.3.2, we can show that $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$. The rest of the proof is the same as the proof in Theorem 5.3.1. □

According to the three theorems, we can safely narrow down our attention to "partial", *op*, and "exact", *oe*, offers such that [9]:

$op = \langle\langle 0, 0, \cdots, 1, \cdots, \ 0, r\rangle, p\rangle$ and $oe = \langle\langle 0, 0, \cdots, 2, \cdots, 0, r\rangle, p\rangle$.

To find the data user's best strategy we consider the two offers of "partial"(*op*) and "exact" (*oe*) granularity levels, separately. For each of the two possible types of offers any of the cases mentioned in Section 5.3.2 can happen. The best responses of the data collector determines an expected cardinality, N, (explained in the last column of Tables 5.1, 5.2, and 5.3) for the data table in each sub-case. Plugging these values of N into the corresponding piece of U_{DU} function in Equation 5.6 and finding the maximizing combination of parameters r and p complete the procedure of finding the game's subgame perfect equilibria.

Tables 5.4 and 5.5 summarize the potentially optimal actions of the data user in successful negotiations if she chooses $g_j = 1$ (partial granularity level) and $g_j = 2$ (exact granularity level), respectively. In these tables the notation $\hat{U}_{DU}^{\langle case\rangle}$ represents the data user's maximum payoff by making an *acceptable* offer in case $\langle case\rangle$. The cases for partial offers are numbered as $p1$ to $p7$ (see Table 5.4) and the cases for the exact offers are numbered as $e1$ to $e7$ (see Table 5.5). Each of these cases exactly correspond to one of the seven sub-cases that could apply to an offer (*i.e.*, sub-cases 1a, 1b, 1c, 2a, 2b, 2c, and 3 from Tables 5.1, 5.2, and 5.3).

g_j. However, as can be seen in Equation 5.6, for a fixed dataset carnality N, the data user's payoff is the same for all $g_j \in \{1, 3, 5\}$.Therefore, we prove the first scenario in the same way explained in Theorem 5.3.1.

[9]Note that an offer with $g_j = 0$ provides the data user with a payoff of at most zero. Therefore, the data user can do at least as good as such offer by offering *op* or *oe* with price zero, instead.

Each row in Tables 5.4 and 5.5 lists a two-variable function to be maximized subject to three conditions (or two conditions in cases $p7$ and $e7$). The first condition in a row is the table condition (*i.e.*, the condition specified above either of the Tables 5.1, 5.2, and 5.3) and the second condition (if exists) refers to the case condition specified in the corresponding row in one of the three Tables 5.1, 5.2, or 5.3.[10] The last condition in each row ensures a successful negotiation. More specifically, the last condition guarantees that the proposed offer provides the data collector with a payoff greater than or equal to zero (otherwise, the offer will not be accepted). If the data user sets $p = 0$ her payoff would be 0. We specify this final case with $\hat{U}^0_{DU} = 0$.

The tables must be considered as a semi-lookup table; For any instance of the problem, the specific values for τ_0, τ_1, .., τ_m, θ, γ, Min, and Max can be determined based on the problem definition. Consequently, the maximizing values, $\hat{r}$ and $\hat{p}$, of each applicable row can be easily calculated subject to the conditions specified. The row which yields the maximum payoff to the data user (across both Tables 5.4 and 5.5) is the winning row. If the winning row does not provide the data user with a payoff of at least zero, then her best action is to make an offer with price zero (or equivalently not make an offer) to avoid turning a loss. In this case (case 0), any offer with price zero is in the game's subgame perfect equilibria and the game finishes with an unsuccessful negotiation. On the contrary, if the winning row provides the data user with a payoff of at least zero, then the values of $\hat{r}$, $\hat{p}$, and g_j in that row specify the data user's equilibrium strategy. The granularity levels of other data fields could be set to values greater than zero if and only if doing so yields the same payoff for the data user.

[10]The order of parameters in the conditions are re-organized to provide a simpler representation of the inequalities.

Table 5.5: Potentially optimal payoffs to the data user if $g_j = 2$

case	Maximum payoff†	Subject to
$e1$	$\hat{U}^{e1}_{DU} = Min \cdot [b \cdot \hat{r} - \hat{p}]$	**[c-1]:** $n \cdot \alpha_{oe} < Min$ **[c-1a]:** $\alpha_{oe} + \gamma \cdot p < \frac{2 \cdot Min}{n}$ **[c-e1]:** $\frac{\gamma \cdot C_{oe}}{Min} + \frac{Min}{n} \leq \alpha_{oe} + \gamma \cdot p$
$e2$	$\hat{U}^{e2}_{DU} = \frac{n}{2} \cdot [\alpha_{oe} + \gamma \cdot \hat{p}] \cdot [b \cdot \hat{r} - \hat{p}]$	**[c-1]:** $n \cdot \alpha_{oe} < Min$ **[c-1b]:** $\frac{2 \cdot Min}{n} \leq \alpha_{oe} + \gamma \cdot p \leq \frac{2 \cdot Max}{n}$ **[c-e2]:** $2 \cdot \sqrt{\frac{\gamma \cdot C_{oe}}{n}} \leq \alpha_{oe} + \gamma \cdot p$
$e3$	$\hat{U}^{e3}_{DU} = Max \cdot [b \cdot \hat{r} - \hat{p}]$	**[c-1]:** $n \cdot \alpha_{oe} < Min$ **[c-1c]:** $\frac{2 \cdot Max}{n} \leq \alpha_{oe} + \gamma \cdot p$ **[c-e3]:** $\frac{\gamma \cdot C_{oe}}{Max} + \frac{Max}{n} \leq \alpha_{oe} + \gamma \cdot p$
$e4$	$\hat{U}^{e4}_{DU} = n \cdot \alpha_{oe} \cdot [b \cdot \hat{r} - \hat{p}]$	**[c-2]:** $Min \leq n \cdot \alpha_{oe} \leq Max$ **[c-2a]:** $\gamma \cdot p - \alpha_{oe} < 0$ **[c-e4]:** $C_{oe} \leq n \cdot \alpha_{oe} \cdot p$
$e5$	$\hat{U}^{e5}_{DU} = \frac{n}{2} \cdot [\alpha_{oe} + \gamma \cdot \hat{p}] \cdot [b \cdot \hat{r} - \hat{p}]$	**[c-2]:** $Min \leq n \cdot \alpha_{oe} \leq Max$ **[c-2b]:** $0 \leq \gamma \cdot p - \alpha_{oe}$ and $\alpha_{oe} + \gamma \cdot p \leq \frac{2 \cdot Max}{n}$ **[c-e5]:** $2 \cdot \sqrt{\frac{\gamma \cdot C_{oe}}{n}} \leq \alpha_{oe} + \gamma \cdot p$
$e6$	$\hat{U}^{e6}_{DU} = Max \cdot [b \cdot \hat{r} - \hat{p}]$	**[c-2]:** $Min \leq n \cdot \alpha_{oe} \leq Max$ **[c-2c]:** $\frac{2 \cdot Max}{n} \leq \alpha_{oe} + \gamma \cdot p$ **[c-e6]:** $\frac{\gamma \cdot C_{oe}}{Max} + \frac{Max}{n} \leq \alpha_{oe} + \gamma \cdot p$
$e7$	$\hat{U}^{e7}_{DU} = Max \cdot [b \cdot \hat{r} - \hat{p}]$	**[c-3]:** $Max < n \cdot \alpha_{oe}$ **[c-e7]:** $C_{oe} \leq Max \cdot p$

† In all formulas $\hat{r}$ and $\hat{p}$ are the values that maximize the payoff in the row subject to the constraints. Parameters α_{oe} and C_{oe} are defined as:
$\alpha_{oe} = \tau_0 + \tau_1 \cdot h_g(0) + \cdots + \tau_j \cdot h_g(2) + \tau_{j+1} \cdot h_g(0) + \cdots + \tau_m \cdot h_g(0) + \theta \cdot h_r(\hat{r})$,
$C_{oe} = B$.

5.4 Results in a Simplified Scenario

To move one step further from the explained 15 cases (14 cases in Tables 5.4 and 5.5 and $\hat{U}^0_{DU}$), we make another assumption and solve the game for those situations where the data user requires the data field A_j for only one year. Similar to Theorems 5.3.2 and 5.3.3 it can be shown that among all offers with $p > 0$ those offers that ask the data for a retention period greater than a year ($r > 1$) can only provide the data user with a payoff of at most the same as the offers with $r = 1$. Therefore, it is enough to only analyze offers, *op* and *oe* with parameter $r = 1$:

$$op = \langle\langle 0,\ 0, \cdots,\ 1, \cdots,\ 0,\ 1\rangle,\ p\rangle \text{ and } oe = \langle\langle 0,\ 0, \cdots,\ 2, \cdots,\ 0,\ 1\rangle,\ p\rangle$$

In Section 5.4.1, for each case we find the optimizing value of price p that maximizes the data user's payoff, subject to the constraints of the case (See the third column of Tables 5.4 and 5.5). Within this analysis, we sometimes need to consider sub-cases and find the

maximizing price in each sub-case separately.

A maximizing price found in this way can potentially be the data user's best offer if it provides her with a payoff of at least zero.[11] In fact, for each game setting, among the best prices found for all relevant cases (and sub-cases), the one that provides the data user with the highest payoff, would be her best response.

During the process of finding the optimum pice for each case, new constraints are recognized that are independent of the chosen price. we refer to these constraints as *environmental conditions* since they can be determined based on the problem definition and their evaluations are independent of the decisions made by the data collector and the data user in the game. These constraints are products of comparing lower/upper bounds of the cases' constraints, splitting a case into sub-cases, and verifying that the data user's maximum payoff is non-negative. With the help of inequalities found as environmental conditions, we partition the problem space into disjoint classes and find the most profitable case (to the data user) among the cases that apply to each class (see Section 5.4.2).

In all calculations, when partial cases $p1$, $p2$, ..., $p7$ (from Table 5.4) are studied, parameters $r = 1$, $\alpha_{op} = \tau_0 + \ \tau_1 \cdot h_g(0) + \cdots + \tau_j \cdot h_g(1) \ + \tau_{j+1} \cdot h_g(0) + \cdots + \tau_m \cdot h_g(0) + \theta \cdot h_r(1)$ and $C_{op} = G + B$ are used. For offers of type oe (exact cases $e1$, $e2$, ..., $e7$ from Table 5.5), parameters $r = 1$, $\alpha_{oe} = \tau_0 + \tau_1 \cdot h_g(0) + \cdots + \tau_j \cdot h_g(2) \ + \tau_{j+1} \cdot h_g(0) + \cdots + \tau_m \cdot h_g(0) + \theta \cdot h_r(1)$ and $C_{oe} = B$ must be used.

5.4.1 Case-based Analysis of Data User's Best Choices

With the assumption of $r = 1$, we are basically moving from optimizing two-variable utility functions to single variable ones (r is not considered as a variable anymore). Here, we explain how to find the maximizing price for each of the cases $p1$, $p2$, ..., $p7$ from Table 5.4. The same procedure can be followed to find the maximum of the utility functions for an exact offer (cases listed in Table 5.5).

[11] Otherwise, offering with price zero is a more profitable choice.

The maximizing price and the maximum payoff to the data user in each case are denoted by $\hat{p}^{\langle case \rangle}$ and $\hat{U}_{DU}^{\langle case \rangle}$ where $\langle case \rangle$ is the ID of the specific case (followed by sub-cases) to which the price and payoff belong.

Best action in case *p1* - The constraints in this case can be summarized as:

- [c-1]: $n \cdot \alpha_{op} < Min$
- [c-1a]: $\alpha_{op} + \gamma \cdot p < \frac{2 \cdot Min}{n}$
- [c-p1]: $\frac{\gamma \cdot C_{op}}{Min} + \frac{Min}{n} \leq \alpha_{op} + \gamma \cdot p$

Since the lower bound for $\alpha_{op} + \gamma \cdot p$ must be less than it's upper bound we need the following inequality to hold:

$$\frac{\gamma \cdot C_{op}}{Min} + \frac{Min}{n} < \frac{2 \cdot Min}{n} \quad \Rightarrow \quad n \cdot \gamma \cdot C_{op} < Min^2 \tag{5.25}$$

Payoff to the data user is the maximum of the $U_{DU}^{p1} = Min \cdot [a - p]$ (the first row of Table 5.4 with $\hat{r}$ being substituted by 1). U_{DU}^{p1} is a decreasing function of price and hence the minimum value for p maximizes the function. In other words:

$$\hat{p}^{p1} = \frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \tag{5.26}$$

This value is greater than zero (since $n \cdot \alpha_{op} < Min$). If we plug this value in the definition of U_{DU}^{p1}, we get the following payoff:

$$\hat{U}_{DU}^{p1} = Min \cdot \left[a - \frac{C_{op}}{Min} - \frac{Min}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] \tag{5.27}$$

As a result, the inequality:

$$\frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \tag{5.28}$$

guarantees that the optimum price, $\hat{p}^{p1}$, provides the data user with a utility of at least zero. Otherwise, the data user is better off by setting the price equal to zero as in case $\hat{U}_{DU}^{0}$. The optimum price, $\hat{p}^{p1}$ results in $\hat{U}_{DC}^{1a} = 0$. Therefore, the data collector will be indifferent

between accepting and rejecting.[12] If $\hat{U}_{DU}^{p1}$ is the maximum among all other relevant $\hat{U}_{DU}$'s, then in the game's sub game perfect Equilibrium, the data user makes an offer *op* with price equal to $\hat{p}^{p1}$ and the data collector sets the incentive to $\frac{Min - n \cdot \alpha_{op}}{n \cdot \gamma}$.

Best action in case *p2* - The constraints in this case can be summarized as:

- [c-1]: $n \cdot \alpha_{op} < Min$
- [c-1b]: $\frac{2 \cdot Min}{n} \leq \alpha_{op} + \gamma \cdot p \leq \frac{2 \cdot Max}{n}$
- [c-p2]: $2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} \leq \alpha_{op} + \gamma \cdot p$

Since both of the lower bounds for $\gamma \cdot p$ must be less than or equal to its upper bound we need the following inequality to hold:

$$2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} \leq \frac{2 \cdot Max}{n} \quad \Rightarrow \quad n \cdot \gamma \cdot C_{op} \leq Max^2 \tag{5.29}$$

Payoff to the data user is the maximum of the $U_{DU}^{p2} = \frac{n}{2} \cdot [\alpha_{op} + \gamma \cdot p] \cdot [a - p]$ (the second row of Table 5.4 with $\hat{r}$ being substituted by 1). To find the maximum, we find the derivative of U_{DU}^{p2} with respect to p and set it to zero:

$$\frac{d(U_{DU}^{p2})}{dp} = \frac{n}{2} \cdot [\gamma \cdot [a - p] - [\alpha_{op} + \gamma \cdot p]] = 0 \Rightarrow \hat{p}^{p2} = \frac{\gamma \cdot a - \alpha_{op}}{2 \cdot \gamma} \tag{5.30}$$

To verify that the maximum of U_{DU}^{p2} occurs within the boundaries, we consider two situations:

(a) $n \cdot \gamma \cdot C_{op} < Min^2$: In this situation, $2\sqrt{\frac{\gamma \cdot C_{op}}{n}} < \frac{2 \cdot Min}{n}$. The value of $\hat{p}^{p2}$ is within boundaries if:

$$\begin{aligned} &\tfrac{2 \cdot Min}{n} \leq \gamma \cdot \hat{p}^{p2} + \alpha_{op} \leq \tfrac{2 \cdot Max}{n} && \Rightarrow \\ &\tfrac{2 \cdot Min}{n} - \alpha_{op} \leq \tfrac{\gamma \cdot a - \alpha_{op}}{2} \leq \tfrac{2 \cdot Max}{n} - \alpha_{op} && \Rightarrow \\ &\tfrac{4Min}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma} \leq a \leq \tfrac{4Max}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma} \end{aligned} \tag{5.31}$$

[12] As explained in Chapter 4, we only analyze those equilibria in which the data collector chooses the action accept when she is indifferent between accepting and rejecting.

I. If Equation 5.31 holds, $\hat{p}^{p2(a)I} = \hat{p}^{p2} = \frac{\gamma \cdot a - \alpha_{op}}{2 \cdot \gamma}$ is the optimum price and we have the following payoffs for the data user and the data collector:

$$\hat{U}_{DU}^{p2(a)I} = \frac{n}{8 \cdot \gamma} \cdot [\alpha_{op} + \gamma \cdot a]^2 \tag{5.32}$$

$$\hat{U}_{DC}^{1b} = \frac{n}{16 \cdot \gamma} [\alpha_{op} + \gamma \cdot a]^2 - C_{op} \tag{5.33}$$

II. if $a < \frac{4Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$, then the optimum price is the beginning of the boundary. Therefore, $\hat{p}^{p2(a)II} = \frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$.

The payoff to the data user would be:

$$\hat{U}_{DU}^{p2(a)II} = Min \cdot \left[a - \frac{2 \cdot Min}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] \tag{5.34}$$

The following condition guarantees a non-negative payoff to the data user:

$$\frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \tag{5.35}$$

If all of the constraints hold and $\hat{U}_{DU}^{p2(a)II}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer op with $p = \frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ and the data collector accepts with $I = \frac{Min - n \cdot \alpha_{op}}{n \cdot \gamma}$. The payoff to the data collector will be:

$$\hat{U}_{DC}^{1b} = \frac{Min^2}{n \cdot \gamma} - C_{op} \tag{5.36}$$

III. If $\frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$, then the maximizing price is $\hat{p}^{p2(a)III} = \frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$. In this case the maximum payoff to the data user is:

$$\hat{U}_{DU}^{p2(a)III} = Max \cdot \left[a - \frac{2 \cdot Max}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] \tag{5.37}$$

Notice that in this situation $\hat{U}_{DU}^{p2(a)III} > 0$. The payoff to the data collector is:

$$\hat{U}_{DC}^{1b} = \frac{Max^2}{n \cdot \gamma} - C_{op} \tag{5.38}$$

If the settings of the problem is aligned with the conditions of this situation and $\hat{U}_{DU}^{p2(a)III}$ is greater than all other relevant $\hat{U}_{DU}$'s, then in the subgame perfect equilibrium, the data user makes an offer of the form *op* with $p = \frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ and the data collector accepts with an incentive $I = \frac{Max - n \cdot \alpha_{op}}{n \cdot \gamma}$.

(b) $Min^2 \leq n \cdot \gamma \cdot C_{op}$: In this situation, $\frac{2 \cdot Min}{n} \leq 2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}}$. The value of $\hat{p}^{p2}$ is within boundaries if:

$$\begin{aligned} & 2 \cdot \sqrt{\tfrac{\gamma \cdot C_{op}}{n}} \leq \gamma \cdot \hat{p}^{p2} + \alpha_{op} \leq \tfrac{2 \cdot Max}{n} && \Rightarrow \\ & 2 \cdot \sqrt{\tfrac{\gamma \cdot C_{op}}{n}} - \alpha_{op} \leq \tfrac{\gamma \cdot a - \alpha_{op}}{2} \leq \tfrac{2 \cdot Max}{n} - \alpha_{op} && \Rightarrow \\ & 4 \cdot \sqrt{\tfrac{C_{op}}{n \cdot \gamma}} - \tfrac{\alpha_{op}}{\gamma} \leq a \leq \tfrac{4 \cdot Max}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma} \end{aligned} \tag{5.39}$$

I. If Equation 5.39 holds, $\hat{p}^{p2(b)I} = \hat{p}^{p2} = \frac{\gamma \cdot a - \alpha_{op}}{2 \cdot \gamma}$ is the optimum price and the situation is the same as p2(a)I.

II. if $a < 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$, then the optimum price is the beginning of the boundary. Therefore, $\hat{p}^{p2(b)II} = 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$ and the data user's payoff would be:

$$\hat{U}_{DU}^{p2(b)II} = \sqrt{n \cdot \gamma \cdot C_{op}} \cdot \left[a - 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} + \frac{\alpha_{op}}{\gamma} \right] \tag{5.40}$$

The following constraint guarantees a non-negative payoff to the data user:

$$2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \tag{5.41}$$

If all of the constraints hold and $\hat{U}_{DU}^{p2(b)II}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer *op* with $p = 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$ and the data collector accepts with $I = \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$. The payoff to the data collector will be:

$$\hat{U}_{DC}^{1b} = \frac{n}{\gamma} \cdot \left[\gamma \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} \right]^2 - C_{op} = 0 \tag{5.42}$$

Here the data collector would be indifferent between accepting or rejecting. The analyzed equilibria in this thesis are based on her acceptance choice.

III. If $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$, then the case is the same as p2(a)III.

Best action in case *p3* - The constraints in this case can be summarized as:

- [c-1]: $n \cdot \alpha_{op} < Min$
- [c-1c]: $\frac{2 \cdot Max}{n} \leq \alpha_{op} + \gamma \cdot p$
- [c-p3]: $\frac{\gamma \cdot C_{op}}{Max} + \frac{Max}{n} \leq \alpha_{op} + \gamma \cdot p$

The data user's payoff is the maximum of $U_{DU}^{p3} = Max \cdot [a - p]$ (the third row of Table 5.4 with $\hat{r}$ being substituted by 1). This function is maximized when p is minimized. Based on the relationship between the two lower bounds, the following two situations can happen:

(a) $Max^2 < n \cdot \gamma \cdot C_{op}$: In this situation, $\frac{2 \cdot Max}{n} < \frac{\gamma \cdot C_{op}}{Max} + \frac{Max}{n}$. The minimum price p would be $\hat{p}^{p3(a)} = \frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ and the data user's maximum payoff would be:

$$\hat{U}_{DU}^{p3(a)} = Max \cdot \left[a - \frac{C_{op}}{Max} - \frac{Max}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] \tag{5.43}$$

The following constraints must hold to guarantee a non-negative payoff to the data user:

$$\frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \tag{5.44}$$

If the constraints hold and $\hat{U}_{DU}^{p3(a)}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer *op* with $p = \frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$. The maximizing incentive after receiving such an offer is $I = \frac{Max - n \cdot \alpha_{op}}{n \cdot \gamma}$ and the payoff to the data collector would be:

$$\hat{U}_{DC}^{1c} = Max \cdot \left[\frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} - \frac{Max}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] - C_{op} = 0 \tag{5.45}$$

Here the data collector would be indifferent between accepting or rejecting but we only consider those equilibria in which the data collector chooses to accept.

(b) $n \cdot \gamma \cdot C_{op} \leq Max^2$: In this situation, $\frac{\gamma \cdot C_{op}}{Max} + \frac{Max}{n} \leq \frac{2 \cdot Max}{n}$. The minimum price p would be $\hat{p}^{p3(b)} = \frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ and data user's the maximum payoff would be:

$$\hat{U}_{DU}^{p3(b)} = Max \cdot \left[a - \frac{2 \cdot Max}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] \tag{5.46}$$

With the following constraint, the data user is guaranteed a non-negative payoff:

$$\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \tag{5.47}$$

If the constraints hold and $\hat{U}_{DU}^{p3(b)}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer op with $p = \frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$. The maximizing incentive after receiving such an offer is $I = \frac{Max - n \cdot \alpha_{op}}{n \cdot \gamma}$ and the payoff to the data collator would be:

$$\hat{U}_{DC}^{1c} = Max \cdot \left[\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} - \frac{Max}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] - C_{op} = \frac{Max^2}{n \cdot \gamma} - C_{op} > 0 \tag{5.48}$$

Best action in case *p4* - The constraints in this case can be summarized as:

- [c-2]: $Min \leq n \cdot \alpha_{op} \leq Max$
- [c-2a]: $\gamma \cdot p - \alpha_{op} < 0$
- [c-p4]: $C_{op} \leq n \cdot \alpha_{op} \cdot p$

Since the lower bound for p must be less than or equal to its upper bound we need the following inequality to hold:

$$\frac{C_{op}}{n \cdot \alpha_{op}} < \frac{\alpha_{op}}{\gamma} \Rightarrow n \cdot \gamma \cdot C_{op} < [n \cdot \alpha_{op}]^2 \tag{5.49}$$

The data user's payoff is the maximum of $U_{DU}^{p4} = n \cdot \alpha_{op} \cdot [a - p]$ (the fourth row of Table 5.4 with $\hat{r}$ being substituted by 1). This function is maximized when p is minimized. According to condition [c-p4], the minimum price value is $\hat{p}^{p4} = \frac{C_{op}}{n \cdot \alpha_{op}}$. Consequently, the maximum payoff to the data user would be:

$$\hat{U}_{DU}^{p4} = n \cdot \alpha_{op} \cdot \left[a - \frac{C_{op}}{n \cdot \alpha_{op}} \right] = n \cdot \alpha_{op} \cdot a - C_{op} \tag{5.50}$$

To guarantee a payoff of at least zero to the data user, the following constraint must hold:

$$\frac{C_{op}}{n \cdot \alpha_{op}} \leq a \tag{5.51}$$

If the constraints hold and $\hat{U}_{DU}^{p4}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer op with $p = \frac{C_{op}}{n \cdot \alpha_{op}}$. The maximizing incentive after receiving such an offer is $I = 0$ and the payoff to the data collator would be:

$$\hat{U}_{DC}^{2a} = n \cdot \alpha_{op} \cdot \left[\frac{C_{op}}{n \cdot \alpha_{op}}\right] - C_{op} = 0 \tag{5.52}$$

Here the data collector would be indifferent between accepting or rejecting. In the studied equilibria, we only consider the acceptance case.

Best action in case *p5* - The constraints in this case can be summarized as:

- [c- 2]: $Min \leq n \cdot \alpha_{op} \leq Max$
- [c-1b]:$0 \leq \gamma \cdot p - \alpha_{op}$ and $\alpha_{op} + \gamma \cdot p \leq \frac{2 \cdot Max}{n}$
- [c-p5]: $2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} \leq \alpha_{op} + \gamma \cdot p$

Since the lower bounds for $\gamma \cdot p + \alpha_{op}$ must be less than or equal to its upper bound we need the following inequality to hold:

$$2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} \leq \frac{2 \cdot Max}{n} \Rightarrow n \cdot \gamma \cdot C_{op} \leq Max^2 \tag{5.53}$$

The data user's payoff is the maximum of the $U_{DU}^{p5} = \frac{n}{2} \cdot [\alpha_{op} + \gamma \cdot p] \cdot [a - p]$ (the fifth row of Table 5.4 with $\hat{r}$ being substituted by 1). To find the maximum, we find the derivative of U_{DU}^{p5} with respect to p and set it to zero:

$$\frac{d(U_{DU}^{p5})}{dp} = \frac{n}{2} \cdot [\gamma \cdot [a - p] - [\alpha_{op} + \gamma \cdot p]] = 0 \Rightarrow \hat{p}^{p5} = \frac{\gamma \cdot a - \alpha_{op}}{2 \cdot \gamma} \tag{5.54}$$

$\hat{p}^{p5}$ maximizes U_{DU}^{p5} but we need to make sure it is within the boundaries. Two situations can happen:

(a) $[n \cdot \alpha_{op}]^2 \leq n \cdot \gamma \cdot C_{op}$: In this situation, $\frac{\alpha_{op}}{\gamma} \leq \frac{2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} - \alpha_{op}}{\gamma}$. The value of $\hat{p}^{p5}$ is within

boundaries if:

$$\begin{aligned}
&2 \cdot \sqrt{\tfrac{\gamma \cdot C_{op}}{n}} \leq \gamma \cdot \hat{p}^{p5} + \alpha_{op} \leq \tfrac{2 \cdot Max}{n} && \Rightarrow \\
&2 \cdot \sqrt{\tfrac{\gamma \cdot C_{op}}{n}} \leq \tfrac{\gamma \cdot a}{2} + \tfrac{\alpha_{op}}{2} \leq \tfrac{2 \cdot Max}{n} && \Rightarrow \\
&4 \cdot \sqrt{\tfrac{\gamma \cdot C_{op}}{n}} - \alpha_{op} \leq \gamma \cdot a \leq \tfrac{4 \cdot Max}{n} - \alpha_{op} && \Rightarrow \\
&4 \cdot \sqrt{\tfrac{C_{op}}{n \cdot \gamma}} - \tfrac{\alpha_{op}}{\gamma} \leq a \leq \tfrac{4 \cdot Max}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma}
\end{aligned} \tag{5.55}$$

I. If Equation 5.55 holds, $\hat{p}^{p5(a)I} = \hat{p}^{p5} = \frac{\gamma \cdot a - \alpha_{op}}{2 \cdot \gamma}$ is the optimum price and we have:

$$\hat{U}_{DU}^{p5(a)I} = \frac{n}{8 \cdot \gamma} \cdot [\alpha_{op} + \gamma \cdot a]^2 \tag{5.56}$$

In this case, the payoff to the data collector is:

$$\hat{U}_{DC}^{2b} = \frac{n}{16 \cdot \gamma} \cdot [\alpha_{op} + \gamma \cdot a]^2 - C_{op} \tag{5.57}$$

II. if $a < 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$, then the optimum price is the beginning of the boundary. Therefore, $\hat{p}^{p5(a)II} = 2\sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$. The payoff to the data user would be:

$$\begin{aligned}
\hat{U}_{DU}^{p5(a)II} &= \tfrac{n}{2} \cdot \left[\alpha_{op} + \gamma \cdot \hat{p}^{p5(a)II}\right] \cdot \left[a - \hat{p}^{p5(a)II}\right] \\
&= \tfrac{n}{2} \cdot \left[\alpha_{op} + 2 \cdot \sqrt{\tfrac{\gamma \cdot C_{op}}{n}} - \alpha_{op}\right] \cdot \left[a - 2 \cdot \sqrt{\tfrac{C_{op}}{n \cdot \gamma}} + \tfrac{\alpha_{op}}{\gamma}\right] \\
&= \sqrt{n \cdot \gamma \cdot C_{op}} \cdot \left[a - 2 \cdot \sqrt{\tfrac{C_{op}}{n \cdot \gamma}} + \tfrac{\alpha_{op}}{\gamma}\right]
\end{aligned} \tag{5.58}$$

The following condition guarantees a non-negative payoff to the data user:

$$2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \tag{5.59}$$

If all of the constraints hold and $\hat{U}_{DU}^{p5(a)II}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer op with $p = 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$ and the data collector accepts with $I = \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$. The payoff to the data collector will be:

$$\begin{aligned}
\hat{U}_{DC}^{2b} &= \tfrac{n}{\gamma} \cdot \left[\tfrac{\alpha_{op} + \gamma \cdot \hat{p}^{p5(a)II}}{2}\right]^2 - C_{op} \\
&= \tfrac{n}{\gamma} \cdot \left[\tfrac{\alpha_{op} + 2 \cdot \sqrt{\frac{\gamma \cdot C_{op}}{n}} - \alpha_{op}}{2}\right]^2 - C_{op} = 0
\end{aligned} \tag{5.60}$$

Here the data collector would be indifferent between accepting or rejecting.

III. If $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$, then the maximizing price is $\hat{p}^{p5(a)III} = \frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$. In this case the maximum payoff to the data user is:

$$\begin{aligned} \hat{U}_{DU}^{p5(a)III} &= \tfrac{n}{2} \cdot \left[\alpha_{op} + \gamma \cdot \hat{p}^{p5(a)III}\right] \cdot \left[a - \hat{p}^{p5(a)III}\right] \\ &= Max \cdot \left[a - \tfrac{2 \cdot Max}{n \cdot \gamma} + \tfrac{\alpha_{op}}{\gamma}\right] \end{aligned} \tag{5.61}$$

Notice that in this situation $\hat{U}_{DU}^{p5(a)III} > 0$. The payoff to the data collector is:

$$\hat{U}_{DC}^{2b} = \frac{n}{\gamma} \cdot \left[\frac{\alpha_{op} + \gamma \cdot \hat{p}^{p5(a)III}}{2}\right]^2 - C_{op} = \frac{Max^2}{n \cdot \gamma} - C_{op} \tag{5.62}$$

If the settings of the problem is aligned with the conditions of this situation and $\hat{U}_{DU}^{p5(a)III}$ is greater than all other relevant $\hat{U}_{DU}$'s, then in the subgame perfect equilibrium the data user makes an offer of the form op with $p = \frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ and the data collector accepts with an incentive $I = \frac{Max - n \cdot \alpha_{op}}{n \cdot \gamma}$.

(b) $n \cdot \gamma \cdot C_{op} < [n \cdot \alpha_{op}]^2$: In this situation, $\frac{2\sqrt{\frac{\gamma \cdot C_{op}}{n}} - \alpha_{op}}{\gamma} < \frac{\alpha_{op}}{\gamma}$. The value of $\hat{p}^{p5}$ is within boundaries if:

$$\begin{aligned} &\alpha_{op} \leq \gamma \cdot \hat{p}^{p5} \leq \tfrac{2 \cdot Max}{n} - \alpha_{op} \quad \Rightarrow \\ &\alpha_{op} \leq \tfrac{\gamma \cdot a - \alpha_{op}}{2} \leq \tfrac{2 \cdot Max}{n} - \alpha_{op} \quad \Rightarrow \\ &3 \cdot \tfrac{\alpha_{op}}{\gamma} \leq a \leq \tfrac{4 \cdot Max}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma} \end{aligned} \tag{5.63}$$

I. If Equation 5.63 holds, $\hat{p}^{p5(b)I} = \hat{p}^{p5} = \frac{\gamma \cdot a - \alpha_{op}}{2 \cdot \gamma}$ is the optimum price and the situation is the same as P5(a)I.

II. if $a < 3 \cdot \frac{\alpha_{op}}{\gamma}$, then the optimum price is the beginning of the boundary. Therefore, $\hat{p}^{p5(b)II} = \frac{\alpha_{op}}{\gamma}$. The payoff to the data user would be:

$$\hat{U}_{DU}^{p5(b)II} = n \cdot \alpha_{op} \cdot \left[a - \frac{\alpha_{op}}{\gamma}\right] \tag{5.64}$$

To achieve a non-negative payoff for the data user, the following condition must hold:

$$\frac{\alpha_{op}}{\gamma} \leq a \tag{5.65}$$

If all of the constraints hold and $\hat{U}_{DU}^{p5(b)II}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer op with $p = \frac{\alpha_{op}}{\gamma}$ and the data collector accepts with $I = \frac{\alpha_{op} - \alpha_{op}}{2 \cdot \gamma} = 0$. The payoff to the data collector will be:

$$\hat{U}_{DC}^{2b} = \frac{n}{\gamma} \cdot \left[\frac{\alpha_{op} + \alpha_{op}}{2} \right]^2 - C_{op} = \frac{n \cdot \alpha_{op}^2}{\gamma} - C_{op} > 0 \tag{5.66}$$

III. If $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$, then the case is the same as p5(a)III.

Best action in case *p6* - The constraints in this case can be summarized as:

- [c-2]: $Min \leq n \cdot \alpha_{op} \leq Max$
- [c-2c]: $\frac{2 \cdot Max}{n} \leq \alpha_{op} + \gamma \cdot p$
- [c-p6]: $\frac{\gamma \cdot C_{op}}{Max} + \frac{Max}{n} \leq \alpha_{op} + \gamma \cdot p$

The payoff to the data user is the maximum of $U_{DU}^{p6} = Max \cdot [a - p]$ (the sixth row of Table 5.4 with $\hat{r}$ being substituted by 1). This function is maximized when p is minimized. Based on the relationship between the two lower bounds, the following two situations can happen:

(a) $Max^2 < n \cdot \gamma \cdot C_{op}$: The constraints and payoff functions in this case are identical to case p3(a).

(b) $n \cdot \gamma \cdot C_{op} \leq Max^2$: The constraints and payoff functions in this case are identical to case p3(b).

Best action in case *p7* - The constraints in this case can be summarized as:

- [c-3]: $Max < n \cdot \alpha_{op}$
- [c-p7]: $C_{op} \leq Max \cdot p$

The payoff to the data user is the maximum of $U_{DU}^{p7} = Max \cdot [a - p]$ (the seventh row of Table 5.4 with $\hat{r}$ being substituted by 1). This function is maximized when p is minimized.

According to condition [c-p7], the minimum price value is $\hat{p}^{p7} = \frac{C_{op}}{Max}$. Consequently, the maximum payoff to the data user would be:

$$\hat{U}_{DU}^{p7} = Max \cdot \left[a - \frac{C_{op}}{Max} \right] \tag{5.67}$$

The data user's maximum payoff is non-negative if the following condition holds:

$$\frac{C_{op}}{Max} \leq a \tag{5.68}$$

If the constraints hold and $\hat{U}_{DU}^{p7}$ is the maximum among all other relevant $\hat{U}_{DU}$'s then in the equilibrium, the data user makes offer *op* with $p = \frac{C_{op}}{Max}$. The maximizing incentive after receiving such an offer is $I = 0$ and the payoff to the data collator would be:

$$\hat{U}_{DC}^{3} = Max \cdot \frac{C_{op}}{Max} - C_{op} = 0 \tag{5.69}$$

Here, the data collector would be indifferent between accepting or rejecting.

We showed how to find the maximum of each payoff function in Table 5.4. The procedure of finding the maximum payoffs in Table 5.5 is identical to the last seven cases (explained for partial granularity offers) and leads to the same results except that all occurrences of α_{op}, C_{op}, and a are substituted by α_{oe}, C_{oe}, and b.

5.4.2 Partitioning the Problem Space

We use the *environmental conditions* to partition the problem space into disjoint classes and find the cases that maximize the data user's payoff in each class. We illustrate the procedure for offers that request the data field A_j at the partial granularity level (offer *op*). An identical problem space partitioning parallel to the one provided for offer *op* can be defined for offer *oe* in which all occurrence of α_{op} and C_{op} are replaced by α_{oe} and C_{oe}.

Class 1 and subclasses- The environmental conditions in this class are:

1. $n \cdot \alpha_{op} < Min$

2. $n \cdot \gamma \cdot C_{op} < Min^2$

The first condition in this class applies to cases $p1$, $p2$, and $p3$. Within these cases, the sub-cases $p1$, $p2(a)$, and $p3(b)$ also require the second condition to be true (for case $p3(b)$ the second condition is a sufficient condition but not necessary).

The relevant cases can be further organized based on bounds on the parameter a. Equation 5.28, Equation 5.31, Equation 5.35, and Equation 5.47 specify different boundaries on values of a. Notice that the lower bound $\frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ (from Equation 5.28) is smaller than the lower bound $\frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ (from Equation 5.35) since the second condition of this class requires: $n \cdot \gamma \cdot C_{op} < Min^2$. When we put these lower/upper bounds of a into an order, we get the following subclasses:

I. $a < \frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$

II. $\frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$

III. $\frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$

IV. $\frac{4 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$

V. $\frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$

For each of these subclasses, we enumerate the applicable cases. Moreover, if possible, we cross out some of the cases since there are other cases in the subclass that provide the data user with a higher payoff.

I. $a < \frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$:

In this subclass, none of the inequalities in Equation 5.28, Equation 5.31, Equation 5.35, Equation 5.47 and $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$ hold. Therefore, the data user's maximum payoffs in all of the cases $p1$, $p2(a)$, and $p3(b)$ are less than zero (in successful negotiations). When the game setting belongs to this subclass, the subgame perfect equilibrium is the play in which the data user does not make an offer (an offer with price zero) or requests

the data field at the exact granularity level (if it provides the data user with a payoff greater than zero).

II. $\frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$:

The only case that applies to this subclass is $p1$ (see Equation 5.28). The subgame perfect equilibrium of the game is the set of strategies mentioned in case $p1$ or requesting the data field at the exact granularity level (if it provides the data user with a higher payoff).

III. $\frac{2 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$:

Based on Equation 5.28 and Equation 5.35, cases $p1$ and $p2(a)II$ apply to this subclass. Moreover, if $Max \leq 2 \cdot Min$ then we have $\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{4Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ and case $p3(b)$ may also apply to this subclass.

According to Equation 5.30 and Equation 5.31, if the conditions of this subclass hold then $\hat{U}_{DU}^{p2(a)II}$ is the local maximum of the $U_{DU}^{p2} = \frac{n}{2} \cdot [\alpha_{op} + \gamma \cdot p] \cdot [a - p]$. This implies:

$$\hat{U}_{DU}^{p2(a)II} \geq \hat{U}_{DU}^{p2(a)III} \tag{5.70}$$

Moreover, based on Equation 5.37 and Equation 5.46, we have $\hat{U}_{DU}^{p2(a)III} = \hat{U}_{DU}^{p3(b)}$. Consequently, for this class, $\hat{U}_{DU}^{p2(a)II} \geq \hat{U}_{DU}^{p3(b)}$ and $p2(a)II$ is a more profitable choice for the data user than $p3(b)$. The data user's payoff in case $p2(a)II$ never gets higher than her payoff in case $p1$. To see the reason, notice that the second condition of this class is $n \cdot \gamma \cdot C_{op} < Min^2$. This condition requires that $\frac{C_{op}}{Min} < \frac{Min}{n \cdot \gamma}$. Therefore, we have:

$$\begin{aligned} &\tfrac{C_{op}}{Min} + \tfrac{Min}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma} < \tfrac{2 \cdot Min}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma} && \Rightarrow \\ &Min \cdot \left[a - \left[\tfrac{C_{op}}{Min} + \tfrac{Min}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma}\right]\right] > Min \cdot \left[a - \left[\tfrac{2 \cdot Min}{n \cdot \gamma} - \tfrac{\alpha_{op}}{\gamma}\right]\right] && \Rightarrow \\ &\hat{U}_{DU}^{p1} > \hat{U}_{DU}^{p2(a)II} \end{aligned} \tag{5.71}$$

As a result, case $p1$ provides a higher profit to the data user than $p2(a)II$ (for the details of the payoff functions please see Equation 5.27 and Equation 5.34). To sum up,

case $p1$ is the subgame perfect equilibrium if the game setting belongs to this subclass (unless requesting the data field at the exact granularity level provides the data user with a higher payoff).

IV. $\frac{4 \cdot Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$:

Based on Equation 5.28 and Equation 5.31, cases $p1$ and $p2(a)I$ apply to this subclass. Moreover, since $\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ case $p3(b)$ may also apply to this subclass.

data user's payoff in case $p3(b)$ never gets higher than her payoff in case $p2(a)I$. To see the reason, notice that $\hat{U}_{DU}^{p3(b)}$ from Equation 5.46 is the same as $\hat{U}_{DU}^{p2(a)III}$ from Equation 5.37. Equation 5.30 proves that $\hat{U}_{DU}^{p2(a)I}$ from Equation 5.32 is the maximum of function U_{DU}^{p2}. Therefore, we know:

$$\hat{U}_{DU}^{p2(a)I} \geq \hat{U}_{DU}^{p2(a)III} = \hat{U}_{DU}^{p3(b)} \tag{5.72}$$

Consequently, for this subclass, $p2(a)I$ is a more profitable choice than $p3(b)$.

In this subclass, between the cases $p1$, $p2(a)I$, and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

V. $\frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$:

Based on Equation 5.28, Equation 5.31, and Equation 5.47, cases $p1$ and $p2(a)III$ and $p3(b)$ apply to this subclass. By comparing Equation 5.37 with Equation 5.46, and Equation 5.38 with Equation 5.48 we see that case $p2(a)III$ and $p3(b)$ offer the same maximum payoffs to the data user and the data collector. Therefore, they represent the same strategy profile.

In this subclass, between the cases $p1$, $p2(a)III$ (or $p3(b)$), and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

Class 2 and subclasses- The environmental conditions in this class are:

1. $n \cdot \alpha_{op} < Min$
2. $Min^2 \leq n \cdot \gamma \cdot C_{op} \leq Max^2$

The first condition in this class only applies to cases $p1$, $p2$, and $p3$. The second condition does not conform to the required inequality of case $p1$ (see Equation 5.25). Within the remaining cases $p2$ and $p3$, the sub cases $p2(b)$, and $p3(b)$ also have the second condition.

The relevant cases can be further organized based on bounds on a. Equation 5.39, Equation 5.41, and Equation 5.47 specify different boundaries on values of a. The bound $4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$ is at most equal to $\frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ because the second constraint of this class requires $n \cdot \gamma \cdot C_{op} \leq Max^2$. When we put these lower/upper bounds of a into an order, we get the following subclasses:

I. $a < 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$

II. $2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$

III. $4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$

IV. $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$

In the following, for each of these subclasses, we discuss the applicable cases and compare the amount of profit they provide to the data user.

I. $a < 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$:

In this subclass, none of the conditions in Equation 5.39, (5.41), (5.47), and $\frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$ (the necessary condition for case $p2(b)III$) hold and consequently no offer provides the data user with a payoff of at least zero (in a successful negotiation). When the game setting matches this subclass, the subgame perfect equilibrium is the play in which the data user is not making an offer or requesting the data field at the exact granularity level (if it provides the data user with a payoff greater than zero).

II. $2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$:

According to Equation 5.41 and since $a \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$ case $p2(b)II$ applies to this subclass. Moreover, if $Max^2 \leq 4 \cdot n \cdot \gamma \cdot C_{op}$ then we have $\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$ and case $p3(b)$ may also apply to this subclass.

With a reasoning similar to subclass III of class 1, it is possible to show that the data user's payoff in case $p3(b)$ never gets higher than her payoff in case $p2(b)II$.

In this subclass, between the case $p2(b)II$ and requesting the data field at the exact granularity level, the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

III. $4\sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$:

Based on Equation 5.39, case $p2(b)I$ apply to this subclass. Moreover, since $\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ case $p3(b)$ may also apply to this subclass.

The data user's payoff in case $p3(b)$ never gets higher than her payoff in case $p2(b)I$. The reasoning is similar to our argument in subclass IV of class 1. In this subclass, between the case $p2(b)I$ and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

IV. $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$:

The subclass condition matches the condition for case $p2(b)III$. Moreover, since $\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ case $p3(b)$ also applies to this subclass.

It can be seen that cases $p2(b)III$ and $p3(b)$ represent the same strategy profile. In this subclass, between the case $p2(b)III$ (or $p3(b)$) and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

Class 3 and subclasses- The environmental conditions in this class are:

1. $n \cdot \alpha_{op} < Min$
2. $Max^2 < n \cdot \gamma \cdot C_{op}$

The first condition in this class applies to cases $p1$, $p2$, and $p3$. The second condition does not conform to the required inequalities of cases $p1$ and $p2$ (see Equation 5.25 and Equation 5.29). Within the remaining case $p3$, the sub-case $p3(a)$ is the only one conforming to the second condition.

Based on the required condition of sub-case $p3(a)$ (see Equation 5.44), we can distinguish two subclasses for this class:

I. $a < \frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$:

In this subclass, the condition in Equation 5.44 doesn't hold and consequently among the offers that the data collector considers acceptable, no offer provides the data user with a payoff of at least zero. The subgame perfect equilibrium is the play in which the data user is not making an offer or requesting the data field at the exact granularity level (if it provides the data user with a payoff greater than zero).

II. $\frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a$:

This condition conforms to Equation 5.44. Between the case $p3(a)$ and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the subgame perfect equilibrium of the subclass.

Class 4 and subclasses- The environmental conditions in this class are:

1. $Min \leq n \cdot \alpha_{op} \leq Max$
2. $n \cdot \gamma \cdot C_{op} < [n \cdot \alpha_{op}]^2 < Max^2$

The first condition in this class applies to cases $p4$, $p5$, and $p6$. Within these cases, the sub-cases $p4$, $p5(b)$, and $p6(b)$ also require the second condition to be true (for case $p6(b)$ the second condition is a sufficient condition but not necessary).

The relevant cases can be further organized based on bounds on parameter a. Equation 5.51, Equation 5.63, Equation 5.65, and Equation 5.47 specify different boundaries on value of a. Notice that the lower bound $\frac{C_{op}}{n \cdot \alpha_{op}}$ (from Equation 5.51) is smaller than the lower bound $\frac{\alpha_{op}}{\gamma}$ (from Equation 5.65) since the second condition of this class requires: $\gamma \cdot C_{op} < n \cdot [\alpha_{op}]^2$. When we put these lower/upper bounds of a into an order, we get the following subclasses:

I. $a < \frac{C_{op}}{n \cdot \alpha_{op}}$

II. $\frac{C_{op}}{n \cdot \alpha_{op}} \leq a \leq \frac{\alpha_{op}}{\gamma}$

III. $\frac{\alpha_{op}}{\gamma} \leq a \leq \frac{3 \cdot \alpha_{op}}{\gamma}$

IV. $\frac{3 \cdot \alpha_{op}}{\gamma} \leq a \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$

V. $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$

For each of these subclasses, we enumerate the applicable cases. Moreover, if possible, we cross out some of the cases since there are other cases in the subclass that provide the data user with a higher payoff.

I. $a < \frac{C_{op}}{n \cdot \alpha_{op}}$:

In this subclass, none of the inequalities in Equation 5.51, Equation 5.63, Equation 5.65, Equation 5.47 and $\frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$ hold and consequently no offer provides the data user with a payoff of at least zero (among the offers that the data collector considers acceptable). In this subclass, the subgame perfect equilibrium is the play in which the data user is not making an offer (we denote such situation by an offer with price zero) or requesting the data field at the exact granularity level (if it provides the data user with a payoff greater than zero).

II. $\frac{C_{op}}{n \cdot \alpha_{op}} \leq a \leq \frac{\alpha_{op}}{\gamma}$:

The only case that applies to this subclass is $p4$ (see Equation 5.51). The subgame perfect equilibrium of the game is the set of strategies mentioned in case $p4$ or requesting the data field at the exact granularity level (if it provides the data user with a higher payoff).

III. $\frac{\alpha_{op}}{\gamma} \leq a \leq \frac{3\alpha_{op}}{\gamma}$:

Based on Equation 5.51 and Equation 5.65, cases $p4$ and $p5(b)II$ apply to this subclass. Moreover, if $Max \leq 2 \cdot [n \cdot \alpha_{op}]$ then we have:

$$\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{3 \cdot \alpha_{op}}{\gamma} \tag{5.73}$$

and case $p6(b)$ may also apply to this subclass.

By looking at the details of the case $p5(b)$, we see that according to Equation 5.54 and Equation 5.63, if the conditions of this subclass hold then $\hat{U}_{DU}^{p5(b)II}$ is the local maximum of the $U_{DU}^{p5} = \frac{n}{2} \cdot [\alpha_{op} + \gamma \cdot p]\,[a - p]$. This implies:

$$\hat{U}_{DU}^{p5(b)II} \geq \hat{U}_{DU}^{p5(b)III} \tag{5.74}$$

Moreover, data user's payoffs in cases $p6(b)$, $p3(b)$, $p5(a)III$, and $p5(b)III$ are the same. We have:

$$\hat{U}_{DU}^{p5(b)III} = \hat{U}_{DU}^{p6(b)} \tag{5.75}$$

Consequently, for this subclass $\hat{U}_{DU}^{p5(b)II} \geq \hat{U}_{DU}^{p6(b)}$ and $p5(b)II$ is a more profitable choice for the data user than $p6(b)$.

The data user's payoff in case $p5(b)II$ never gets higher than her payoff in case $p4$. To see the reason, notice that the second condition of this class is $n \cdot \gamma \cdot C_{op} < [n \cdot \alpha_{op}]^2$. This condition requires that $C_{op} < \frac{n \cdot [\alpha_{op}]^2}{\gamma}$. Therefore, we have:

$$-C_{op} > -\frac{n \cdot [\alpha_{op}]^2}{\gamma} \Rightarrow n \cdot \alpha_{op} \cdot a - C_{op} > n \cdot \alpha_{op} \cdot a - \frac{n \cdot [\alpha_{op}]^2}{\gamma} \Rightarrow \hat{U}_{DU}^{p4} > \hat{U}_{DU}^{p5(b)II} \tag{5.76}$$

As a result, case $p4$ provides a higher profit to the data user than $p5(b)II$(for the details of the payoff functions please see Equation 5.50 and Equation 5.64). To sum up, case $p4$ is the subgame perfect equilibrium of this subclass (unless requesting the data field at the exact granularity level provides the data user with a higher payoff).

IV. $\frac{3\cdot\alpha_{op}}{\gamma} \leq a \leq \frac{4\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$:

Based on Equation 5.51 and Equation 5.63, cases $p4$ and $p5(b)I$ apply to this subclass. Moreover, since $\frac{2\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{4Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$ case $p6(b)$ may also apply to this subclass.

The data user's payoff in case $p6(b)$ never gets higher than her payoff in case $p5(b)I$. To see the reason, notice that the data user's payoffs in cases $p6(b)$, $p3(b)$, $p5(a)III$, and $p5(b)III$ are the same. We have:

$$\hat{U}_{DU}^{p5(b)III} = \hat{U}_{DU}^{p6(b)} \tag{5.77}$$

Equation 5.54 proves that $\hat{U}_{DU}^{p5(b)I}$ (equal to $\hat{U}_{DU}^{p5(a)I}$ from Equation 5.56) is the maximum of the function $\hat{U}_{DU}^{p5} = \frac{n}{2}\cdot[\alpha_{op} + \gamma\cdot p]\cdot[a-p]$. Therefore, we know that:

$$\hat{U}_{DU}^{p5(b)I} \geq \hat{U}_{DU}^{p5(b)III} = \hat{U}_{DU}^{p6(b)} \tag{5.78}$$

Consequently, for this subclass, case $p5(b)I$ is a more profitable option for the data user than $p6(b)$.

In this subclass, between the cases $p4$, $p5(b)I$, and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

V. $\frac{4\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} < a$:

Based on Equation 5.51, Equation 5.63, and Equation 5.47, cases $p4$ and $p5(b)III$, and $p6(b)$ apply to this subclass.

The data user offers the same price in cases $p6(b)$, $p3(b)$, $p5(a)III$, and $p5(b)III$ and her payoffs are the same in all of these four cases. Therefore, $p5(b)III$ and $p6(b)$ represent the same strategy profile.

In this subclass, between the cases $p4$, $p5(b)III$ (or $p6(b)$), and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

Class 5 and subclasses- The environmental conditions in this class are:

1. $Min \leq n \cdot \alpha_{op} \leq Max$
2. $[n \cdot \alpha_{op}]^2 \leq n \cdot \gamma \cdot C_{op} \leq Max^2$

The first condition in this class applies to cases $p4$, $p5$, and $p6$. The second condition does not conform to the required inequality of case $p4$ (see Equation 5.49). Within the remaining cases $p5$ and $p6$, the sub cases $p5(a)$, and $p6(b)$ also have the second condition.

The relevant cases can be further organized based on bounds on the parameter a. Equation 5.55, Equation 5.59, and Equation 5.47 specify different boundaries on value of a. When we put these lower/upper bounds of a into an order, we get the following subclasses:

I. $a < 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$

II. $2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$

III. $4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$

IV. $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$

In the following, for each of these subclasses, we discuss the applicable cases and compare the amount of profit they provide to the data user.

I. $a < 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$:

In this subclass, none of the conditions in Equation 5.55, (5.59), (5.47), and $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$ (the necessary condition for case $p5(a)III$) hold and consequently no offer provides the data user with a payoff of at least zero (in a successful negotiation). In this subclass, the subgame perfect equilibrium is the play in which the data user is not making an offer (we denote this situation by an offer with price zero) or requesting the data field at the exact granularity level (if it provides the data user with a payoff greater than zero).

II. $2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$:

According to Equation 5.59 and since $a \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$ case $p5(a)II$ applies to this subclass. Moreover, if $Max^2 \leq 4 \cdot n \cdot \gamma \cdot C_{op}$ then we have:

$$Max \leq 2 \cdot \sqrt{n \cdot \gamma \cdot C_{op}} \Rightarrow \frac{2 \cdot Max}{n \cdot \gamma} \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} \Rightarrow \frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \tag{5.79}$$

and case $p6(b)$ may also apply to this subclass.

The data user's payoff in case $p6(b)$ never gets higher than her payoff in case $p5(a)II$. To see the reason, notice that data user's payoffs in cases $p6(b)$, $p3(b)$, and $p5(a)III$ are the same. We have:

$$\hat{U}_{DU}^{p5(a)III} = \hat{U}_{DU}^{p6(b)} \tag{5.80}$$

When a is within the limits defined for this subclass, $\hat{U}_{DU}^{p5(a)II}$ (see Equation 5.58) is the maximum of the $U_{DU}^{p5} = \frac{n}{2} \cdot [\alpha_{op} + \gamma \cdot p] \cdot [a - p]$ function.Therefore, we know that:

$$\hat{U}_{DU}^{p5(a)II} \geq \hat{U}_{DU}^{p5(a)III} = \hat{U}_{DU}^{p6(b)} \tag{5.81}$$

Consequently, for this subclass, case $p5(a)II$ is a more profitable choice than $p6(b)$.

In this subclass, between the case $p5(a)II$ and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

III. $4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$:

Based on Equation 5.55, case $p5(a)I$ apply to this subclass. Moreover, since $\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{4Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ case $p6(b)$ may also apply to this subclass.

The data user's payoff in case $p6(b)$ never gets higher than her payoff in case $p5(a)I$. The reason is similar to our argument in subclass IV of class 4. In this subclass, between the case $p5(a)I$ and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

IV. $\frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} < a$:

The subclass condition matches the condition for case $p5(a)III$. Moreover, since $\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$ case $p6(b)$ also applies to this subclass.

Cases $p6(b)$ and $p3(b)$ provide the same payoffs to the data user and data collector. Moreover, by comparing Equation 5.61 with Equation 5.46, and Equation 5.62 with Equation 5.48 we see that cases $p5(a)III$ and $p3(b)$ offer the same maximum payoffs to the data user and the data collector since the data user offers the same price in both cases. Therefore, $p5(a)III$ and $p6(b)$ represent the same strategy profile.

In this subclass, between the case $p5(a)III$ (or $p6(b)$) and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

Class 6 and subclasses- The environmental conditions in this class are:

1. $Min \leq n \cdot \alpha_{op} \leq Max$
2. $Max^2 < n \cdot \gamma \cdot C_{op}$

The first condition in this class applies to cases $p4$, $p5$, and $p6$. The second condition does not conform to the required inequalities of case $p4$ and $p5$ (see Equation 5.49 and Equation

5.53). Within the remaining case $p6$, the sub-case $p6(a)$ is the only one conforming to the second condition. Since the strategy in $p6(a)$ is the same as $p3(a)$, the subclasses and their analysis for this class are the same as Class 3.

Class 7 and subclasses- The environmental condition in this class is:

- $Max < n \cdot \alpha_{op}$

This condition only matches with the case $p7$. Based on the required condition of case $p7$ (see Equation 5.68), we can distinguish two subclasses for this class:

I. $a < \frac{C_{op}}{Max}$:

In this subclass, the condition in Equation 5.68 doesn't hold and no offer provides the data user with a payoff of at least zero (in a successful negotiation). The subgame perfect equilibrium is the play in which the data user is not making an offer (we denote such situation by an offer with price zero) or requesting the data field at the exact granularity level (if it provides the data user with a payoff greater than zero).

II. $\frac{C_{op}}{Max} \leq a$:

This condition conforms to Equation 5.68. Between the case $p7$ and requesting the data field at the exact granularity level the one that provides a higher payoff to the data user determines the game's subgame perfect equilibrium.

Tables 5.6 to 5.12 summarize subgame perfect equilibria of the game in different situations if requesting the data field at the partial granularity level is more beneficial to the data user than the exact one. Similar to Tables 5.6 to 5.12, seven more tables can be produced by changing every instance of parameter a to b and using α_{oe} and C_{oe} instead of α_{op} and C_{op}. Consequently, each real instance of the problem matches to one row from Tables 5.6 to 5.12 and one row in the other series of seven tables for offer oe. In fact, these tables partition the problem space into 22×22 different spaces and demonstrate the outcome and subgame perfect equilibria of each case.

Table 5.6: Subgame perfect equilibria strategies for Class 1†

	Condition	Best price	$\hat{U}_{DU}$	$\hat{U}_{DC}$
I	$a < \frac{C_{op}}{Min} + \frac{Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} = l_1$	0	0	0
II/III	$l_1 \le a < \frac{4\cdot Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} = l_2$	$\frac{C_{op}}{Min} + \frac{Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$	$Min \cdot \left[a - \left[\frac{C_{op}}{Min} + \frac{Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]$	$\hat{U}^{1a}_{DC} = 0$
IV	$l_2 \le a \le \frac{4\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} = l_3$	$\frac{C_{op}}{Min} + \frac{Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$ OR $\frac{\gamma\cdot a - \alpha_{op}}{2\cdot\gamma}$	$\max\{Min \cdot \left[a - \left[\frac{C_{op}}{Min} + \frac{Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]$, $\frac{n}{8\cdot\gamma}\cdot[\alpha_{op} + \gamma\cdot a]^2\}$	$\hat{U}^{1a}_{DC} = 0$ OR $\hat{U}^{1b}_{DC} = \frac{n}{16\cdot\gamma}\cdot[\alpha_{op} + \gamma\cdot a]^2 - C_{op}$
V	$l_3 < a$	$\frac{C_{op}}{Min} + \frac{Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$ OR $\frac{2\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$	$\max\{Min \cdot \left[a - \left[\frac{C_{op}}{Min} + \frac{Min}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]$, $Max \cdot \left[a - \left[\frac{2\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]\}$	$\hat{U}^{1a}_{DC} = 0$ OR $\hat{U}^{1b}_{DC} = \frac{Max^2}{n\cdot\gamma} - C_{op}$

† In this class, data field A_j is requested at the partial granularity level, $n\cdot\alpha_{op} < Min$, and $n\cdot\gamma\cdot C_{op} < Min^2$

Table 5.7: Subgame perfect equilibria strategies for Class 2†

	Condition	Best price	$\hat{U}_{DU}$	$\hat{U}_{DC}$
I	$a < 2\cdot\sqrt{\frac{C_{op}}{n\cdot\gamma}} - \frac{\alpha_{op}}{\gamma} = l_1$	0	0	0
II	$l_1 \le a < 4\cdot\sqrt{\frac{C_{op}}{n\cdot\gamma}} - \frac{\alpha_{op}}{\gamma} = l_2$	$2\cdot\sqrt{\frac{C_{op}}{n\cdot\gamma}} - \frac{\alpha_{op}}{\gamma}$	$\sqrt{n\cdot\gamma\cdot C_{op}}\cdot\left[a - 2\cdot\sqrt{\frac{C_{op}}{n\cdot\gamma}} + \frac{\alpha_{op}}{\gamma}\right]$	$\hat{U}^{1b}_{DC} = 0$
III	$l_2 \le a \le \frac{4\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} = l_3$	$\frac{\gamma\cdot a - \alpha_{op}}{2\cdot\gamma}$	$\frac{n}{8\cdot\gamma}\cdot[\alpha_{op} + \gamma\cdot a]^2$	$\hat{U}^{1b}_{DC} = \frac{n}{16\cdot\gamma}\cdot[\alpha_{op} + \gamma\cdot a]^2 - C_{op}$
IV	$l_3 \le a$	$\frac{2\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$	$Max \cdot \left[a - \left[\frac{2\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]$	$\hat{U}^{1b}_{DC} = \frac{Max^2}{n\cdot\gamma} - C_{op}$

† In this class, data field A_j is requested at the partial granularity level, $n\cdot\alpha_{op} < Min$, and $Min^2 \le n\cdot\gamma\cdot C_{op} \le Max^2$

Table 5.8: Subgame perfect equilibria strategies for Class 3†

	Condition	Best price	$\hat{U}_{DU}$	$\hat{U}_{DC}$
I	$a < \frac{C_{op}}{Max} + \frac{Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} = l_1$	0	0	0
II	$l_1 \le a$	$\frac{C_{op}}{Max} + \frac{Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$	$Max \cdot \left[a - \left[\frac{C_{op}}{Max} + \frac{Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]$	$\hat{U}^{1c}_{DC} = 0$

† In this class, data field A_j is requested at the partial granularity level, $n\cdot\alpha_{op} < Min$, and $Max^2 < n\cdot\gamma\cdot C_{op}$

Table 5.9: Subgame perfect equilibria strategies for Class 4†

	Condition	Best price	$\hat{U}_{DU}$	$\hat{U}_{DC}$
I	$a < \frac{C_{op}}{n\cdot\alpha_{op}} = l_1$	0	0	0
II/III	$l_1 \le a < \frac{3\cdot\alpha_{op}}{\gamma} = l_2$	$\frac{C_{op}}{n\cdot\alpha_{op}}$	$n\cdot\alpha_{op}\cdot\left[a - \left[\frac{C_{op}}{n\cdot\alpha_{op}}\right]\right]$	$\hat{U}^{2a}_{DC} = 0$
IV	$l_2 \le a \le \frac{4\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma} = l_3$	$\frac{C_{op}}{n\cdot\alpha_{op}}$ OR $\frac{\gamma\cdot a - \alpha_{op}}{2\cdot\gamma}$	$\max\{n\cdot\alpha_{op}\cdot\left[a - \left[\frac{C_{op}}{n\cdot\alpha_{op}}\right]\right]$ $\frac{n}{8\cdot\gamma}\cdot[\alpha_{op} + \gamma\cdot a]^2\}$	$\hat{U}^{2a}_{DC} = 0$ OR $\hat{U}^{2b}_{DC} = \frac{n}{16\cdot\gamma}\cdot[\alpha_{op} + \gamma\cdot a]^2 - C_{op}$
V	$l_3 \le a$	$\frac{C_{op}}{n\cdot\alpha_{op}}$ OR $\frac{2\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}$	$\max\{n\cdot\alpha_{op}\cdot\left[a - \left[\frac{C_{op}}{n\cdot\alpha_{op}}\right]\right]$ $Max\cdot\left[a - \left[\frac{2\cdot Max}{n\cdot\gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]\}$	$\hat{U}^{2a}_{DC} = 0$ OR $\hat{U}^{2b}_{DC} = \frac{Max^2}{n\cdot\gamma} - C_{op}$

† In this class, data field A_j is requested at the partial granularity level, $Min \le n\cdot\alpha_{op} \le Max$, and $n\cdot\gamma\cdot C_{op} < [n\cdot\alpha_{op}]^2 < Max^2$

Table 5.10: Subgame perfect equilibria strategies for Class 5†

	Condition	Best price	$\hat{U}_{DU}$	$\hat{U}_{DC}$
I	$a < 2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} = l_1$	0	0	0
II	$l_1 \leq a < 4 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma} = l_2$	$2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}$	$\sqrt{n \cdot \gamma \cdot C_{op}} \cdot \left[a - \left[2 \cdot \sqrt{\frac{C_{op}}{n \cdot \gamma}} - \frac{\alpha_{op}}{\gamma}\right]\right]$	$\hat{U}_{DC}^{2b} = 0$
III	$l_2 \leq a \leq \frac{4 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} = l_3$	$\frac{\gamma \cdot a - \alpha_{op}}{2 \cdot \gamma}$	$\frac{n}{8 \cdot \gamma} \cdot [\alpha_{op} + \gamma \cdot a]^2$	$\hat{U}_{DC}^{2b} = \frac{n}{16 \cdot \gamma} \cdot [\alpha_{op} + \gamma \cdot a]^2 - C_{op}$
IV	$l_3 \leq a$	$\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$	$Max \cdot \left[a - \left[\frac{2 \cdot Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]$	$\hat{U}_{DC}^{2b} = \frac{Max^2}{n \cdot \gamma} - C_{op}$

† In this class, data field A_j is requested at the partial granularity level, $Min \leq n \cdot \alpha_{op} \leq Max$, and $[n \cdot \alpha_{op}]^2 \leq n \cdot \gamma \cdot C_{op} \leq Max^2$

Table 5.11: Subgame perfect equilibria strategies for Class 6†

	Condition	Best price	$\hat{U}_{DU}$	$\hat{U}_{DC}$
I	$a < \frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} = l_1$	0	0	0
II	$l_1 \leq a$	$\frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$	$Max \cdot \left[a - \left[\frac{C_{op}}{Max} + \frac{Max}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}\right]\right]$	$\hat{U}_{DC}^{2c} = 0$

† In this class, data field A_j is requested at the partial granularity level, $Min \leq n \cdot \alpha_{op} \leq Max$, and $Max^2 < n \cdot \gamma \cdot C_{op}$

Table 5.12: Subgame perfect equilibria strategies for Class 7†

	Condition	Best price	$\hat{U}_{DU}$	$\hat{U}_{DC}$
I	$a < \frac{C_{op}}{Max} = l_1$	0	0	0
II	$l_1 \leq a$	$\frac{C_{op}}{Max}$	$Max \cdot \left[a - \left[\frac{C_{op}}{Max}\right]\right]$	$\hat{U}_{DC}^{3} = 0$

† In this class, data field A_j is requested at the partial granularity level, $Max < n \cdot \alpha_{op}$.

5.5 Case Studies and Application of the Results

We explain the usage and implications of the classes found in Section 5.4.2 via two synthetic case studies.

Case study 1: The first case represents an abstract instance of the problem. Consider a situation where the data user requires data field A_j for only one year.Therefore, it is enough to analyze offers, *op* and *oe* with the same format explained in Section 5.4.

For any offer of partial information request, *op*, let α_{op} denote the probability of a data provider providing data for A_j at the partial granularity level with zero incentive and C_{op} represent the cost of providing the data user with the dataset according to *op*. Therefore, $\alpha_{op} = \tau_0 + \tau_1 \cdot h_g(0) + \cdots + \tau_j \cdot h_g(1) + \tau_{j+1} \cdot h_g(0) + \cdots + \tau_m \cdot h_g(0) + \theta \cdot h_r(1)$ and $C_{op} = G + B$. Similar to α_{op} and C_{op}, we use the notations $\alpha_{oe} = \tau_0 + \tau_1 \cdot h_g(0) + \cdots + \tau_j \cdot h_g(2) + \tau_{j+1} \cdot h_g(0) + \cdots + \tau_m \cdot h_g(0) + \theta \cdot h_r(1)$ and $C_{oe} = B$ to represent the same concepts where the offer is made for the exact granularity level, *oe*.

Consider an abstract instance of the problem with the following setups:

Condition *op*(a): $n \cdot \alpha_{op} < Min$,

Condition *op*(b): $n \cdot \gamma \cdot C_{op} < Min^2$,

Condition *op*(c): $\frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \leq a \leq \frac{4Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma}$,

Condition *oe*(a): $n \cdot \alpha_{oe} < Min$,

Condition *oe*(b): $n \cdot \gamma \cdot C_{oe} < Min^2$ and

Condition *oe*(c): $\frac{C_{oe}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{oe}}{\gamma} \leq b \leq \frac{4Min}{n \cdot \gamma} - \frac{\alpha_{oe}}{\gamma}$

The first three conditions only match with the second row of Table(5.6). In other words, the environmental conditions and boundaries on a match classes 1-II and 1-III. In both of these classes the game's subgame perfect equilibria are the set of strategies specified in

case $p1$ from Section 5.4.1 or requesting data at the exact granularity level. As a result, by requesting data at the partial granularity level, the data user expects the following maximum payoff:

$$\hat{U}^{p}_{DU} = \hat{U}^{p1}_{DU} = Min \cdot \left[a - \frac{C_{op}}{Min} - \frac{Min}{n \cdot \gamma} + \frac{\alpha_{op}}{\gamma} \right] \tag{5.82}$$

Similarly, by requesting data at the exact granularity level, the data user expects the following maximum payoff:

$$\hat{U}^{e}_{DU} = \hat{U}^{e1}_{DU} = Min \cdot \left[b - \frac{C_{oe}}{Min} - \frac{Min}{n \cdot \gamma} + \frac{\alpha_{oe}}{\gamma} \right] \tag{5.83}$$

The data user has to make a final decision between *op* and *oe* based on $\hat{U}^{p1}_{DU}$ and $\hat{U}^{e1}_{DU}$. The data user chooses to ask for exact information if the following holds:

$$Min \cdot \left[a - \left[\frac{C_{op}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{op}}{\gamma} \right] \right] < Min \cdot \left[b - \left[\frac{C_{oe}}{Min} + \frac{Min}{n \cdot \gamma} - \frac{\alpha_{oe}}{\gamma} \right] \right] \tag{5.84}$$

Since the privacy parameters offered in *op* and *oe* only differ in the value of g_j, the difference between α_{op} and α_{oe} is the difference between $h_g(1)$ and $h_g(2)$ magnified by the coefficient τ_j. In other words, $\alpha_{op} - \alpha_{oe} = \tau_j \cdot [h_g(1) - h_g(2)]$. Since there are 7 different levels of granularity, we assume that function $h_g(x) = 6 - x$ explains the privacy gain of setting a granularity level to value x. Considering this function instantiation, $\alpha_{op} - \alpha_{oe} = \tau_j$, $C_{op} - C_{oe} = G$, and we can rewrite the inequality in Equation 5.84 as:

$$\frac{\tau_j}{\gamma} - \frac{G}{Min} < b - a \tag{5.85}$$

Therefore, if Equation 5.85 holds, the data user asks for exact information on A_j. In case of an equality the data user would be indifferent between asking for exact or partial information. Finally, if $b - a < \frac{\tau_j}{\gamma} - \frac{G}{Min}$ the data user is better off by asking for partial information. This analysis shows the followings:

1. With larger values of τ_j, asking for partial granularity level becomes the most profitable option for the data user. Recall that $\tau_j = \beta_1 \cdot \lambda_j = \left[\frac{p_2 \cdot w_1}{\overline{v}} + \frac{p_3}{\overline{v'}} \right] \cdot \lambda_j$. A higher value of τ_j can be caused by higher values of w_1 and/or λ_j.

The former parameter, w_1, represents the weight of privacy gain within data providers' privacy/incentive trade-off analysis (see Chapter 4 Section 4.2.1.1). Therefore, a larger value for w_1 models a population of data providers with higher privacy awareness. Moreover, according to Equation 5.3, parameter λ_j in the definition of privacy gain function $h(\delta)$ is the weight assigned to $h_g(g_j)$ based on sensitivity of the data field A_j. Consequently, in a population of data providers with high privacy awareness who considers data field A_j sensitive, the firms are forced to collect only partial information rather than exact information.

2. As the influence of incentive becomes less important on the public's privacy decisions (as γ decreases), the only way to collect personal information is to protect individual's privacy by using generalization (or other perturbation methods) on data.

3. When the cost of generalization G increases, asking for data at the exact granularity level becomes the most profitable choice of the data user.

4. By increasing the minimum number of required records, the third party has no choice other than providing privacy to the data providers and ask for data at the partial granularity level.

Case study 2: As a more concrete case study, consider a situation where a pharmacy goods manufacturer (a data user) is planning to launch a new production line. To make the best decision on what kind of goods to produce, the manufacturer needs a dataset containing the pharmacy-related shopping habits of habitants in the area. Consequently, she decides to ask for such information from the largest supermarket (a data collector) in the city. The supermarket records the `date, total amount payable`, and `items` purchased when each customer pays for her basket. The supermarket can offer some discounts and seek for the

customer's permission to provide this information to the pharmacy goods manufacturer (for each shopping trip).

According to Westin's indexes [KC05] and Grossklags and Acquisti 's finding [GA07] (who claim that 25 cents is enough for data providers to trade privacy), a data provider can be a privacy unconcerned (with $p_1 = 0.16$), a privacy pragmatist (with $p_2 = 0.59$), or a privacy fundamentalist (with $p_3 = 0.25$) and parameters $w_1 = 0.25$ and $w_2 = 1$ specify the weights in privacy pragmatists' indifference curves. For this example, we assume that functions $h_g(g_i) = 6 - g_i$ and $h_r(r) = \frac{1}{1+r}$ define privacy gains of providing data field A_i at granularity level g_i and providing data for retention period r, respectively. Moreover, we make the assumption that sensitivities of data fields `date`, `total amount payable`, `items`, and retention, are $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, $\lambda_3 = 0.5$, and $\lambda_4 = 0.2$, respectively. With these characteristics, a data provider opts in for a combination of privacy parameter $\delta = \langle g_1, g_2, g_3, r \rangle$ and incentive, I, with the following probability:

$$\begin{aligned} Prob(OptIn) = \; & 0.16 + 0.0126 \times [6 - g_1] + 0.0063 \times [6 - g_2] + 0.0315 \times [6 - g_3] \\ & + 0.0126 \times \tfrac{1}{r+1} + 0.0524 \times I \end{aligned} \tag{5.86}$$

Where g_1, g_2, and g_3 represent granularity levels of `date, total amount payable`, and `items` respectively. Equation 5.86 defines each data provider's best response based on what explained in Section 5.3.1 (see Equation 5.10, Equation 5.11, and Equation 5.12).

The supermarket might not know all the underlying characteristics of the data providers but she is assumed to know its clients' privacy behavior (the opt in probability function) based on some past experience.

The data user is only interested in accessing information about `item` data field for a year. This information can be provided at the exact granularity level or at the partial granularity level (after values are generalized to broader categories). Other parameters of this problem instance are summarized in Table 5.13.

The data user must decide on the granularity level, g_3, of values collected for `item` data field and the price for each piece of information. She calculates α_{op} , C_{op}, α_{oe}, and C_{oe} as

Table 5.13: Parameter settings

	Parameter	Value		Parameter	Value
1	Max	$25,000$	2	Min	$15,000$
3	B	\$1,000	4	G	\$100
5	a	\$5	6	b	\$10
7	n	$30,000$			

follows:

$$\begin{aligned}
\alpha_{op} &= \tau_0 + \tau_1 \cdot [6 - g_1] + \tau_2 \cdot [6 - g_2] + \tau_3 \cdot [6 - g_3] + \theta \cdot \tfrac{1}{r+1} \\
&= 0.16 + 0.0126 \times [6 - 0] + 0.0063 \times [6 - 0] + 0.0315 \times [6 - 1] + 0.0126 \times \tfrac{1}{1+1} \\
&= 0.4377 \\
C_{op} &= G + B = 1,100 \\
\alpha_{oe} &= \tau_0 + \tau_1 \cdot [6 - g_1] + \tau_2 \cdot [6 - g_2] + \tau_3 \cdot [6 - g_3] + \theta \cdot \tfrac{1}{r+1} \\
&= 0.16 + 0.0126 \times [6 - 0] + 0.0063 \times [6 - 0] + 0.0315 \times [6 - 2] + 0.0126 \times \tfrac{1}{1+1} \\
&= 0.4061 \\
C_{oe} &= B = 1,000
\end{aligned} \tag{5.87}$$

We can easily see that the setting of this problem conforms to the conditions ***op*** **(a), (b), (c)** and ***oe*** **(a), (b), (c)**. Therefore we have:

$$\begin{aligned}
\hat{U}^p_{DU} &= Min \cdot \left[a - \tfrac{C_{op}}{Min} - \tfrac{Min}{n \cdot \gamma} + \tfrac{\alpha_{op}}{\gamma}\right] = 56077.97 \\
\hat{U}^e_{DU} &= Min \cdot \left[b - \tfrac{C_{oe}}{Min} - \tfrac{Min}{n \cdot \gamma} + \tfrac{\alpha_{oe}}{\gamma}\right] = 122152.50
\end{aligned} \tag{5.88}$$

Since $\hat{U}_{DU}[e] > \hat{U}_{DU}[p]$, the payoff to the data user would be higher if the exact granularity level is asked for. Therefore, in the stable state of the game the subgame perfect Equilibrium suggests the following payoffs and prices:

$$\begin{aligned}
p &= \tfrac{C_{oe}}{Min} - \tfrac{Min}{n \cdot \gamma} + \tfrac{\alpha_{oe}}{\gamma} = 1.8565 \\
U_{DU} &= Min \cdot \left[b - \tfrac{C_{oe}}{Min} - \tfrac{Min}{n \cdot \gamma} + \tfrac{\alpha_{oe}}{\gamma}\right] = 122152.50 \\
U_{DC} &= 0
\end{aligned} \tag{5.89}$$

Notice that the equilibrium results in $U_{DC} = 0$. This means that the the data collector will make just enough money to be compensated for all of her efforts and nothing more. Therefore, the she will be indifferent between accepting and rejecting. We have only analyzed those

equilibria in which the data collector accepts when she is indifferent. Although accepting the offer does not add any monetary value to the utility of the data collector, the opportunity of building a relationship with a possibly future supplier might convince the supermarket to collect information. The data collector sets the incentive to $\frac{Min - n \cdot \alpha_{oe}}{n \cdot \gamma} = 1.7898$.

5.6 Summary

In this chapter, we demonstrated how the generic game model can be tailored to address the problem of setting a balanced privacy policy. We chose a sample privacy policy declaration method similar to the P3P method and suggest 7 different levels of granularity control for each data field. The definitions of the privacy parameter δ and the privacy gain functions are discussed according to the context.

The game's subgame perfect equilibria are explained as the data user's most profitable choice among 15 different options: 7 options for asking the data field A_j at the partial granularity level such that the data collector accepts the offer, 7 options for asking the data field A_j at the exact granularity level such that the data collector accepts the offer, and the option of offering with price zero that leads to a "reject" response from the data collector. For each instance of the game a subset of these options are applicable and the one that leads to the highest profit to the data user is her winning strategy.

To provide a more in-depth game analysis, we assumed the simplifying scenario where the data user only required to retain the collected information for a year. Within this scenario, we were able to partition the problem space into 22×22 classes such that the setting of each game instance only matches the criteria of one of these classes. We compared the applicable choices (among the explained 15 options) and found the most profitable ones in each class. Finally, we demonstrated the applications and implications of the game's solutions using two case studies.

Chapter 6

Game Model Instantiation for Data Sanitization Methods

In this chapter we demonstrate how the generic game model (explained in Section 4.2) can be instantiated for different data sanitization methods. We describe and analyze two games to find stable privacy parameter values for a seminal non-interactive data sanitization technique (k-anonymity) and for a standard interactive method (differential privacy). For each game, the solution is explained and simulations of the results are provided. The last section of this chapter, discusses an important difficulty in comparing different sanitization systems and illustrates how equilibria of "privacy parameter setting" games can suggest a meaningful approach for such comparison.

6.1 Stable k Values in k-Anonymity

k-Anonymity is a seminal data sanitization method that includes a privacy parameter k. In this section, after providing a brief overview on k-anonymity and the Mondrian algorithm, we show how our generic game model can be used to balance the privacy/utility tradeoff involved in k-anonymity.[1]

6.1.1 k-Anonymity and the Mondrian Algorithm

We begin by assuming that 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

[1]This work is published in an international conference and awarded the best student paper award [KABSN12].

a single individual in the dataset via linking attacks. A subset of tuples in a 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-identifying attribute values in the dataset, there are at least $k-1$ other records that contain such a combination.

As explained in Section 2.2.1.3, there are several methods to achieve k-anonymity. The basic techniques use 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 the Mondrian algorithm [LDR06a] for k-anonymity. This greedy algorithm implements partition-based multi-dimensional global recoding which allows finer-grained search and thus often leads to a better data quality. In the Mondrian algorithm there is no cell suppression and the generalizations are not restricted by predefined generalization hierarchies.

The Mondrian algorithm first partitions records into non-overlapping classes and then uses summary statistics of classes to recode the records. The partitioning phase is done by adopting a spatial representation of tuples. In a table with d quasi-identifying attributes, a tuple is represented as a point in d-dimensional space. The algorithm starts with the whole table as a single d-dimensional rectangular box, chooses a dimension, splits the box along a point on the chosen dimension, and recursively splits each of the two newly created partitions. The algorithm stops when no more cut on any of the regions can create new regions with at least k tuples. Within each recursion, the candidate dimension is chosen heuristically and the cutting point is chosen to be the *median* value along the chosen dimension. With the choice of using the median as cutting points, LeFevre *et al.*[LDR05b] proves that if m denotes the maximum number of identical tuples in a region, there exists a rectangular box with $2 \cdot d \cdot [k-1] + m$ tuples that cannot be split into two boxes of at least k tuples. They

further show that $2 \cdot d \cdot [k-1] + m$ is the upper bound on the size of non-splittable regions. An algorithmic description of the partitioning procedure in the the Mondrian algorithm is provided in Figure 6.1.

```
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)
```

Figure 6.1: Mondrian Algorithm.

6.1.2 k-Anonymity Game Model

When k-anonymity is chosen as the data sanitization mechanism, an increase in the value of k leads to a higher privacy protection and as the value of k decreases the resulting data table has higher utility to a data user (k is a privacy enhancer). Consequently, $\langle k \rangle$ is the privacy parameter in k-anonymity (*i.e.*, $\delta = \langle k \rangle$) and the proposed game model in Section 4.2 can be used to find stable values of k. When the generic game is instantiated for k-anonymity, the rules of the game explained in Section 4.2.2 remain intact. However, in the generic game model, some of the function definitions are left unspecified. These functions must be defined according to details of the specific privacy protection method used in the game. In this section, we briefly review components of the game and concentrate on explaining the required function definitions when k-anonymity is used as the privacy protection method within the game model.

As with the generic game model, this game has three groups of players: n data providers DP, a data collector DC, and a data user DU.

6.1.2.1 Data Providers

Data providers decide whether to provide their personal information based on the privacy protection level implied by the value of k and the value of promised incentives. Unlike those data providers who are privacy unconcerned, privacy pragmatists and privacy fundamentalists subconsciously compare the promised privacy protection level (and incentive in case of privacy pragmatists) to their threshold to make a decision (see Section 4.2.1.1). Therefore, to define an instance of the game for k-anonymity, definitions of the privacy gain function $h(\delta)$ and the rewarding effects of the incentive $g(I)$ must be clarified. For simplicity, we only consider the identity function for the incentive and we use the logarithmic function to capture the privacy gain caused by the value of k. In other words :

$$h(k) = log_2(k) \quad \text{and} \quad g(I) = I \tag{6.1}$$

To understand our choice of the *log* function for $h(.)$, 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 ($log_2(1)$) to 1 ($log_2(2)$). 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 the entropy ($logk$) of this uniform probability distribution ($p = \frac{1}{k}$) is the best indicator for privacy protection.

6.1.2.2 Data User

A data user is the entity who requests a k-anonymized data table from the data collector for the purpose of conducting data analysis. In successful negotiations, a data user receives the whole k-anonymized data table to run her queries on. Recall that in Section 4.2.1.2, we considered Min and Max bounds on the number of data records that a data user considers beneficial. Here, for simplicity we will assume $Min = 0$ and $Max = \infty$. As a result, the

effective cardinality of the dataset $EN(N)$ and the cardinality of the dataset N are assumed to be the same. If b represents the true economic value of each record to the data user and p denotes the price she pays for each record, based on Equation 4.11, the data user's payoff function can be defined as:

$$U_{DU} = [b \cdot Precision(k, N) - p] \cdot N \tag{6.2}$$

The definition of data user's payoff function U_{DU} requires 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 a 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. We will demonstrate the calculations for a specific `SELECT` query with potential quantity imprecision problem and use the 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 the sensitive attribute, T^* is the anonymized dataset, `q` is one of the quasi-identifying attributes, and $\mathtt{v}_i$ 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 `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 those tuples with `q =` $\mathtt{v}_i$. The latter introduce some quantity imprecision in the result. LeFevre *et al.*[LDR08] introduce the Region-Based Utility Metric for Select Queries (see Section 2.2.3.2) to find the best cuts while running the Mondrian algorithm [LDR06a] on experimental datasets. After normalizing this metric, we define *Precision* as:

$$Precision = \frac{|Q_i(T)|}{|Q_i(T^*)|} \tag{6.3}$$

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 $\mathtt{q}$. To measure precision, we only consider queries that seek for information within the refined domain of attribute $\mathtt{q}$. For these queries we can still have $|Q_i(T)| = 0$ but we are guaranteed to have $|Q_i(T^*)| > 0$.

To provide an estimate on *Precision* we need to estimate $|Q_i(T)|$ and $|Q_i(T^*)|$. Let N represent the cardinality of the dataset and Pr_i denote the ratio of the records that have value $\mathtt{v}_i$ for the quasi-identifying attribute $\mathtt{q}$ in the dataset. The expected value of $|Q_i(T)|$ is:

$$|Q_i(T)| = Pr_i \cdot N \tag{6.4}$$

In Lemma 6.1.1 below, we use some facts about the Mondrian algorithm [LDR06a] to estimate the depth of the recursive calls during anonymization. This estimate is then used in Theorem 6.1.2 to estimate $|Q_i(T^*)|$.

According to Lefevre *et al.*'s [LDR06a] second theorem, the maximum number of records in each equivalence class is $2 \cdot d \cdot [k-1] + m$, where d is the number of quasi-identifying attributes and m denotes the maximum number of records with identical values for all quasi-identifying attributes. Moreover, in a k-anonymous dataset the minimum number of records in each class is k. Since the distribution of equivalence class sizes is 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:

$$ec_{AVG} = \frac{2 \cdot d \cdot [k-1] + m + k}{2} \tag{6.5}$$

Lemma 6.1.1. *Let N denote cardinality of the dataset. If the average size of each equivalence class is determined by Equation 6.5, then the depth of the recursive calls, σ, in the Mondrian algorithm can be estimated as:*

$$\sigma = log_2(\frac{2 \cdot N}{2 \cdot d \cdot [k-1] + m + k}) \tag{6.6}$$

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 are 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 Equation 6.5.

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

$$Size_x = \frac{N}{2^x} \tag{6.7}$$

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

$$\begin{array}{lcll} Size_\sigma & = & ec_{AVG} & \Rightarrow \\ \frac{N}{2^\sigma} & = & \frac{2 \cdot d \cdot [k-1]+m+k}{2} & \Rightarrow \\ \sigma & = & log_2(\frac{2 \cdot N}{2 \cdot d \cdot [k-1]+m+k}) & \end{array} \tag{6.8}$$

□

Lemma 6.1.2. *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:*

$$|Q_i(T^*)| = [1 - \frac{1}{2 \cdot d}]^\sigma \cdot N \tag{6.9}$$

where d is the number of quasi-identifying attributes and σ is the depth of recursive calls estimated in Lemma 6.1.1.

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 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 $\mathtt{q}$ or other dimensions. For this estimate, we assume that all dimensions are chosen with equal probability. Therefore, the algorithm chooses dimension $\mathtt{q}$ 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:

$$|Q_i(T^*)|_x = \sum_{j=1}^{j=s} |e_j| \tag{6.10}$$

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:

$$|e_j|_{x+1} = \frac{1}{d} \cdot \left[\frac{|e_j|}{2}\right] + \left[1 - \frac{1}{d}\right] \cdot |e_j| = \left[1 - \frac{1}{2 \cdot d}\right] \cdot |e_j| \tag{6.11}$$

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

$$\begin{aligned} |Q_i(T^*)|_{x+1} &= \textstyle\sum_{j=1}^{j=s} |e_j|_{x+1} \\ &= \textstyle\sum_{j=1}^{j=s} [1 - \frac{1}{2 \cdot d}] \cdot |e_j| \\ &= [1 - \tfrac{1}{2 \cdot d}] \cdot |Q_i(T^*)|_x \end{aligned} \tag{6.12}$$

By solving the recursive formula we get $|Q_i(T^*)|_\sigma = [1 - \frac{1}{2 \cdot d}]^\sigma \cdot N$. □

Consequently, *Precision* is can be defined as:

$$Precision = \frac{Pr_i \cdot N}{[1 - \frac{1}{2 \cdot d}]^\sigma \cdot N} = \frac{Pr_i}{[1 - \frac{1}{2 \cdot d}]^\sigma} \tag{6.13}$$

We can also use Lemma 6.1.2 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 6.1.2 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:

$$Pr_i = \frac{|Q_i(T^*_{k=1})|}{N} = [1 - \frac{1}{2 \cdot d}]^{log_2 \frac{2 \cdot N}{m+1}} \tag{6.14}$$

Now we can refine Equation 6.13 as:

$$Precision = \frac{[1 - \frac{1}{2 \cdot d}]^{log_2 \frac{2 \cdot N}{m+1}}}{[1 - \frac{1}{2 \cdot d}]^{\sigma}} \tag{6.15}$$

6.1.2.3 Data Collector

The data collector is responsible for collecting personal information from the data providers in exchange for some incentive. Once a private data table is collected, we assume that an honest data collector runs the Mondrian algorithm with the accepted value of k and sends a copy of the sanitized data table to the data user. The accumulative cost of data collection, storage, running the Mondrian algorithm, and providing the data user with a copy of the data table is represented as the fixed cost C. Since $EN(N) = N$, the data collector's payoff from Equation 4.14 can be re-written as:

$$U_{DC} = [p - I] \cdot N - C \tag{6.16}$$

6.1.3 Subgame Perfect Equilibria in k-Anonymity

Using the process of backward induction to find the game's subgame perfect equilibria, the first step is to find the data providers' best responses in "data collection subgames". Our game is modeled as an incomplete game where data providers can be of any of the three types: privacy unconcerned, privacy pragmatist, privacy fundamentalist each with a certain

probability. In Section 4.3.1 it is shown how to find the best responses of each data provider for each combination of δ and I (see Equation 4.24). When k-anonymity is used as the privacy protection mechanism, we consider $\delta = \langle k \rangle$ and $h(\delta) = h(k) = log_2(k)$ (see Equation 6.1). After substituting these function instantiations into Equation 4.24, a data provider opts in for a combination of k and I with probability $Prob(OptIn)$:

$$Prob(OptIn) = p_1 \cdot 1 + p_2 \cdot \frac{w_1 \cdot log_2(k) + w_2 \cdot I}{\overline{v}} + p_3 \cdot \frac{log_2(k)}{\overline{v}'} \tag{6.17}$$

where $\overline{v} = \max_{k,I}\{w_1.log_2(k) + w_2.I\}$ and $\overline{v}' = \max_k\{log_2(k)\}$.

Since we are assuming that data providers make their decisions independently, the expected cardinality of a dataset with n data providers can be defined as $N = n \cdot Prob(OptIn)$. Let $\beta_0 = p_1$, $\beta_1 = [p_2 \cdot w_1]/\overline{v} + p_3/\overline{v}'$, and $\beta_2 = [p_2 \cdot w_2]/\overline{v}$. Based on equation Equation 6.17, the expected cardinality of the dataset can be found with the following regression model:

$$N = n \cdot [\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot I] \tag{6.18}$$

Knowing the best responses of data providers, we can fold back the game tree one step and find the data collector's best response in subgames that start with the data collector's moves. In these subgames, a data collector has received an offer $Of = \langle k, p \rangle$ and must determine the optimum incentive $\hat{I}$ such that it maximizes U_{DC} from Equation 6.16 considering the expected cardinality as $N = n(k, I)$ (see Equation 6.18). If the data collector accepts the offer $Of = \langle k, p \rangle$ with incentive I, her payoff will be:

$$U_{DC} \quad = \quad [p - I] \cdot n \cdot [\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot I] - C \tag{6.19}$$

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

$$\frac{dU_{DC}}{dI} = -n \cdot [\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot I] + n \cdot \beta_2 \cdot [p - I] = 0 \;\; \Rightarrow \hat{I} = \frac{\beta_2 \cdot p - \beta_1 \cdot log_2(k) - \beta_0}{2 \cdot \beta_2} \tag{6.20}$$

$\hat{I}$ is a 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 $n \cdot [\beta_0 + \beta_1 \cdot log_2(k)]$. The lower bound on I leads us to consider two separate cases:

Case 1: $\beta_2 \cdot p \geq \beta_1 \cdot log_2(k) + \beta_0$: In this case the value of I which maximizes U_{DC} is $\hat{I} = \frac{\beta_2 \cdot p - \beta_1 \cdot log_2(k) - \beta_0}{2 \cdot \beta_2}$ and the maximum payoff to the data collector (in case of an acceptance) will be:

$$\begin{aligned} \hat{U}^1_{DC} &= [p - \tfrac{\beta_2 \cdot p - \beta_1 \cdot log_2(k) - \beta_0}{2 \cdot \beta_2}] \cdot n \cdot [\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot \tfrac{\beta_2 \cdot p - \beta_1 \cdot log_2(k) - \beta_0}{2 \cdot \beta_2}] - C \\ &= \tfrac{\beta_2 \cdot n}{4} \cdot [p + \tfrac{\beta_1 \cdot log_2(k) + \beta_0}{\beta_2}]^2 - C \end{aligned} \tag{6.21}$$

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:

$$p + \frac{\beta_1 \cdot log_2(k)}{\beta_2} \geq \sqrt{\frac{4 \cdot C}{\beta_2 \cdot n}} - \frac{\beta_0}{\beta_2} \tag{6.22}$$

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

$$N = n \cdot [\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot \hat{I}] = \frac{n}{2} \cdot [\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot p] \tag{6.23}$$

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

$$\hat{U}^2_{DC} = p \cdot n \cdot [\beta_0 + \beta_1 \cdot log_2(k)] - C \tag{6.24}$$

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

$$p \cdot n \cdot [\beta_0 + \beta_1 \cdot log_2(k)] \geq C \tag{6.25}$$

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

$$N = n \cdot [\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot 0] = n \cdot [\beta_0 + \beta_1 \cdot log_2(k)] \tag{6.26}$$

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

$$\hat{I} = \hat{i}(k,p) = \begin{cases} \frac{\beta_2 \cdot p - \beta_1 \cdot log_2(k) - \beta_0}{2 \cdot \beta_2} & if\ \beta_2 \cdot p \geq \beta_1 \cdot log_2(k) + \beta_0 \\ \\ 0 & Otherwise \end{cases} \tag{6.27}$$

Plugging this definition into Equation 4.30 from Section 4.3.3, we can define the cardinality of the private dataset as a piecewise function of k and p:

$$N = \begin{cases} n \cdot [\frac{\beta_0 + \beta_1 \cdot log_2(k) + \beta_2 \cdot p}{2}] & if\ \beta_2 \cdot p \geq \beta_1 \cdot log_2(k) + \beta_0\ \ \wedge p + \frac{\beta_1 \cdot log_2(k)}{\beta_2} \geq \sqrt{\frac{4 \cdot C}{\beta_2 \cdot n}} - \frac{\beta_0}{\beta_2} \\ \\ n \cdot [\beta_0 + \beta_1 \cdot log_2(k)] & if\ \beta_2 \cdot p < \beta_1 \cdot log_2(k) + \beta_0\ \ \wedge p \cdot n \cdot [\beta_0 + \beta_1 \cdot log_2(k)] \geq C \\ \\ 0 & Otherwise \end{cases} \tag{6.28}$$

If the new definition of N is plugged into the $Precision$ function (see Equation 6.15) precision becomes a function of k and p (*i.e.*, $Precision = prec(k,p)$). As a result, U_{DU} from Equation 6.2 can be rewritten as a function of k and p. The best strategy for the data user is to compute $\hat{k}$ and $\hat{p}$ such that they maximize her payoff:

$$\langle \hat{k}, \hat{p} \rangle = \arg\max_{k,p} U_{DU} = \arg\max_{k,p} [b \cdot prec(k,p) - p] \cdot N(k,p) \tag{6.29}$$

The maximizing values for k and p represent the optimum offer and solving Equation 6.29 completes the process of finding the game's subgame perfect equilibria.

6.1.4 Simulation Results for k-Anonymity

If players of the game are rational and have the required information, the game's subgame perfect equilibria would always conform to what Section 6.1.3 suggests since we used an

analytical method to find the game's equilibria. In this section, we provide the results of simulating the game with different settings to discuss how various problem settings affect stable values of the privacy parameter $\delta = \langle k \rangle$.

6.1.4.1 Problem Settings

In our method, a dataset does not exist before the game is complete and the specifications of the collected dataset depend on the values chosen for the parameters while the game is played. Therefore, running experiments on real databases does not provide meaningful results for this work. Alternatively, we choose to simulate the game and visualize the results by testing multiple problem settings using MATLAB R2008a. We test the effects of the cost C, the population size n, the maximum number of data providers with the same quasi-identifier values m, the privacy awareness, and the privacy trust on the equilibria of the game. In each simulation, values of all parameters are fixed while varying the values of one or two of the parameters that are the subjects of the test. The fixed values chosen for parameters are: $n = 20,000$, $b = \$10$, $d = 4$, $m = 5$, $C = \$20,000$, $p_1 = 16\%$, $p_2 = 59\%$, $p_3 = 25\%$, $w_1 = 25$, and $w_2 = 1$.

Values for n, b, d, and C are randomly selected as an estimate of reasonable values commonly used in real instances of the problem. We chose to initialize $m = 5$ because we believe that in practice this number cannot be very large otherwise the term quasi-identifier would not make sense. However, we show how the stable values of k change as this number increases.

To model different types of data providers, probabilities p_1, p_2, p_3 of having a privacy unconcerned, privacy pragmatist, and privacy fundamentalist are set to exact values that Westin estimates (see Chapter 2). Recall that w_1 and w_2 are marginal effects of privacy protection and incentive on privacy pragmatists' indifference curves. The choice of $w_1 = 25$ and $w_2 = 1$ is made to conform to Grossklags and Acquisti 's finding [GA07] who claim 25 cents is enough for data providers to trade privacy. With this choice the marginal rate of

substitution w_1/w_2 would be 25. In other words, with this settings we assume that a privacy pragmatist would be indifferent between the choices of receiving one extra unit of privacy protection or receiving 25 cents instead.

An increase in privacy awareness is modeled by increasing the ratio w_1/w_2. This means that a more privacy aware data provider is willing to substitute one unit of privacy with larger amounts of incentives. We fix w_2 to the value 1 and just increase the values of w_1. This increase changes the values of β_1 and β_2 as shown in Table 6.1. Moreover, to assess the effects of trust, we use the rationale that as trust (in the data collector and privacy protection mechanism) increases, a subset of privacy fundamentalists will consider data practices safe and will not be paranoid about their privacy anymore. This effect can be captured by having a population of data providers with less privacy fundamentalists and more privacy unconcerned. This change affects β_0 and β_1. Different values chosen for $p1$, $p2$, and $p3$ (and their effects on β_0, β_1, and β_2) to test various trust levels are summarized in Table 6.2.

Table 6.1: Problem settings for privacy awareness test in k-anonymity

Test	p_1	p_2	p_3	w_1	w_2	$\overline{v}$ †	$\overline{v}'$	β_0 ‡	β_1	β_2
1	0.16	0.59	0.25	10	1	1066.439	6.6438	0.16	0.0432	0.00055
2	0.16	0.59	0.25	20	1	1132.877	6.6438	0.16	0.0480	0.00052
3	0.16	0.59	0.25	30	1	1199.316	6.6438	0.16	0.0524	0.00049
4	0.16	0.59	0.25	40	1	1265.754	6.6438	0.16	0.0563	0.00047
5	0.16	0.59	0.25	50	1	1332.193	6.6438	0.16	0.0598	0.00044
6	0.16	0.59	0.25	60	1	1398.631	6.6438	0.16	0.0629	0.00042
7	0.16	0.59	0.25	70	1	1465.07	6.6438	0.16	0.0658	0.00040
8	0.16	0.59	0.25	80	1	1531.508	6.6438	0.16	0.0684	0.00039
9	0.16	0.59	0.25	90	1	1595.947	6.6438	0.16	0.0709	0.00037
10	0.16	0.59	0.25	100	1	1664.386	6.6438	0.16	0.0731	0.00035
11	0.16	0.59	0.25	110	1	1730.824	6.6438	0.16	0.0751	0.00034
12	0.16	0.59	0.25	120	1	1797.263	6.6438	0.16	0.0770	0.00033
13	0.16	0.59	0.25	130	1	1868.701	6.6438	0.16	0.0788	0.00032
14	0.16	0.59	0.25	140	1	1930.14	6.6438	0.16	0.0804	0.00031
15	0.16	0.59	0.25	150	1	1996.578	6.6438	0.16	0.0820	0.00030

† Considering maximum values of 100 and b for parameters k and I, we have $\overline{v} = w_1 \cdot \log_2(100) + w_2 \cdot b$ and $\overline{v}' = \log_2(100)$.

‡ Values of β_0, β_1, and β_2 are calculated as $\beta_0 = p_1$, $\beta_1 = [p_2 \cdot w_1]/\overline{v} + p_3/\overline{v}'$, and $\beta_2 = [p_2 \cdot w_2]/\overline{v}$.

6.1.4.2 Results and Discussion

The results of examining effects of m, C, n, privacy awareness, and privacy trust are shown in Figures 6.2 - 6.4.

Table 6.2: Problem settings for privacy trust test in k-anonymity

Test	p_1	p_2	p_3	w_1	w_2	$\overline{v}$ †	$\overline{v}'$	β_0 ‡	β_1	β_2
1	0.10	0.59	0.31	25	1	1166.096	6.6438	0.10	0.0593	0.00051
2	0.11	0.59	0.30	25	1	1166.096	6.6438	0.11	0.0578	0.00051
3	0.12	0.59	0.29	25	1	1166.096	6.6438	0.12	0.0563	0.00051
4	0.13	0.59	0.28	25	1	1166.096	6.6438	0.13	0.0548	0.00051
5	0.14	0.59	0.27	25	1	1166.096	6.6438	0.14	0.0533	0.00051
6	0.15	0.59	0.26	25	1	1166.096	6.6438	0.15	0.0518	0.00051
7	0.16	0.59	0.25	25	1	1166.096	6.6438	0.16	0.0503	0.00051
8	0.17	0.59	0.24	25	1	1166.096	6.6438	0.17	0.0488	0.00051
9	0.18	0.59	0.23	25	1	1166.096	6.6438	0.18	0.0473	0.00051
10	0.19	0.59	0.22	25	1	1166.096	6.6438	0.19	0.0458	0.00051
11	0.20	0.59	0.21	25	1	1166.096	6.6438	0.20	0.0443	0.00051
12	0.21	0.59	0.20	25	1	1166.096	6.6438	0.21	0.0428	0.00051
13	0.22	0.59	0.19	25	1	1166.096	6.6438	0.22	0.0412	0.00051
14	0.23	0.59	0.18	25	1	1166.096	6.6438	0.23	0.0397	0.00051
15	0.24	0.59	0.17	25	1	1166.096	6.6438	0.24	0.0382	0.00051

† Considering maximum values of 100 and b for parameters k and I, we have $\overline{v} = w_1 \cdot \log_2(100) + w_2 \cdot b$ and $\overline{v}' = \log_2(100)$.

‡ Values of β_0, β_1, and β_2 are calculated as $\beta_0 = p_1$, $\beta_1 = [p_1 \cdot w_1]/\overline{v} + p_3/\overline{v}'$, and $\beta_2 = [p_2 \cdot w_2]/\overline{v}$.

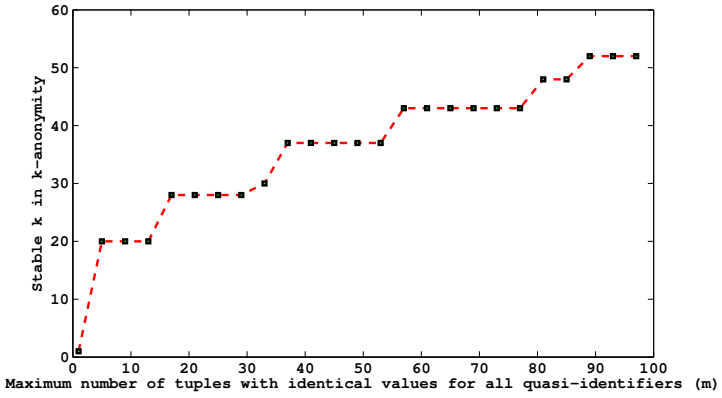
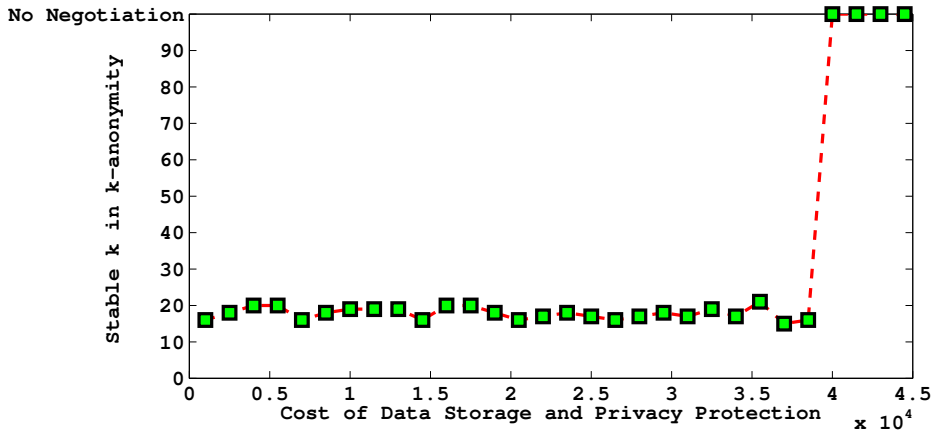


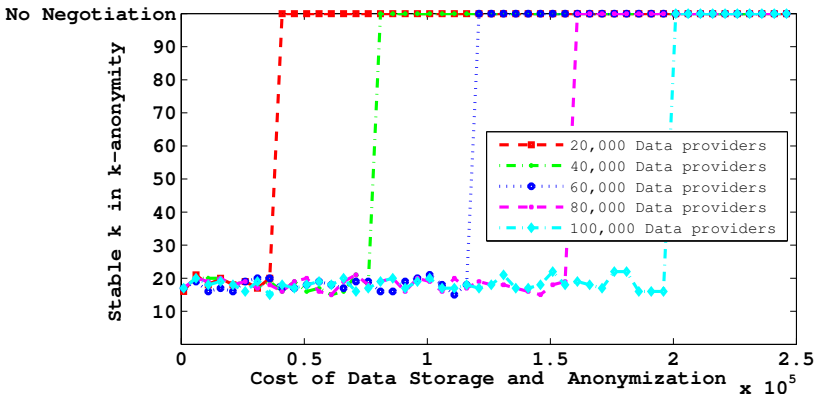
Figure 6.2: Changes to the stable values of k due to an increase in the maximum number of data providers with identical values for their quasi-identifiers m.

Figure 6.2 shows the effects of m (maximum number of data providers with identical quasi-identifier values) on the stable values of k. We choose values of m from $\{1, ..., 100\}$. As the value of m increases the stable values of k increase. To understand this counter-intuitive result, notice that as m increases less generalization will be needed to group the tuples in equivalence classes of size k. Therefore, compared to the cases with smaller m, the same precision can be achieved with higher values of k. Larger values of k attract more data providers without largely affecting the precision of query results and consequently, the data

user might be able to make more profit in this case. Since values of k are discrete the curve has a "staircase-like" shape; when m increases by one unit, the data user has the choice of increasing k which in turn decreases precision but increases the number of data providers who opt-in. However, she does not increase k unless the decrease in precision (by increasing k) can be justified by the amount of extra precision she is saving (by having a larger m) and extra number of data providers she attracts with a higher k.



(a)

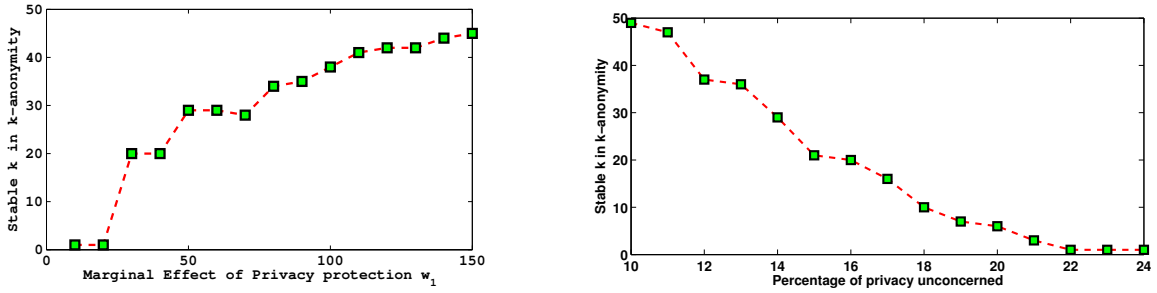


(b)

Figure 6.3: Changes to the stable k due to an increase in: (a) the cost of data anonymization and storage C; (b) the cost of data anonymization C and population of data providers n.

The effects of anonymization, and maintenance cost (C) on stable values of k are illustrated in Figure 6.3(a). Before the game goes to the phase where no negotiation is profitable (before $C \approx 40,000$) stable values of k are not drastically changed by increasing the cost

of privacy protection and maintenance. In fact, there is a zigzag trend in changes to values of stable k. The range of values of C can be decomposed into smaller intervals such that in every other interval having higher costs leads to higher privacy and in the rest of the intervals the data user is better off by compensating the data collector only through proposing a higher price. This fact shows that unlike the results found by Freuudiger *et al.*in the context of non-cooperative location privacy protection [FMHP09] increasing the cost of data sanitization is not a permanent solution to promote higher privacy protection levels. After a certain value for C, the game reaches a point where no combination of $\langle k, p\rangle$ can be found that is both acceptable by the data collector and has $U_{DU} \geq 0$. We have illustrated this case by the highest horizontal bar with label "No Negotiation". This situation represents an instance of *impractical* privacy protection method (see Chapter 4). Figure 6.3(b) illustrates tests of different sizes for the data providers' populations to show how a larger population of data providers justifies using a more expensive algorithm. As n increases, the unsuccessful negotiation threshold is pushed back to higher costs for privacy protection.



(a) (b)

Figure 6.4: Changes to the stable k due to an increase in: (a) data providers' privacy awareness; (b) data providers' trust in the data collector and the privacy protection method.

Figures in 6.4 represent the effects of the data providers' privacy awareness and trust (in the data collector and privacy protection mechanism) on stable values of k. According to Figure 6.4(a) as the marginal effect of privacy gain increases within the privacy pragmatists

indifference curves, a higher compensation is needed to convince a privacy pragmatist to give up one unit of privacy protection. As a result, a data user is better off by proposing higher privacy protection levels to attract more data providers. This shows that when privacy has a more significant impact on data providers' decisions, data will be sanitized with larger values of k. The effect of trust is captured by the ratio of privacy unconcerned population to privacy fundamentalist population. As trust increases, fewer people would be paranoid about their privacy and therefore the population of privacy unconcerned increases. As Figure 6.4(b) shows, an increase in the data providers' trust can lead to a lower privacy protection level (and higher profits to the data collector and the data user). Since the data collector and data user benefit from this trust they are more likely to act honestly and within the terms of their contract to avoid distrust in future interaction. In this model we have assumed honesty for all players and this result shows that some incentive exists to encourage such behavior. However, without considering future interactions, the effect of this incentive might fade. An accurate study on parties' honesty must be done using a repetitive version of the game model (a future extension to this work). These diagrams show how the public's privacy awareness and distrust can force firms to protect privacy of data providers.

6.2 Stable Values of ϵ in Differential Privacy

Differential privacy is an interactive output perturbation mechanism to protect data providers' privacy. The value of the parameter ϵ in this method determines the magnitude of noise added to queries' results. When this method is used, a balanced privacy/utility tradeoff can be achieved by choosing stable values for the privacy parameter ϵ. In this section, after providing a short overview on differential privacy and its implementation, we show how to find stable values of parameter ϵ by instantiating our generic game for differential privacy.

6.2.1 Differential Privacy and Laplace Mechanism

The goal in differential privacy is to limit privacy risks of providing personal information. For each data provider, the risk of contributing to a data table is defined as how much this decision *increases* the *extra* amount of information which can be learned about her [Dwo11]. As discussed in Section 2.2.2.1, differential privacy can achieve this goal by using a randomized function κ. The function κ adds random noise to the results of each query Q such that after receiving the noisy results of Q an adversary cannot distinguish (with a high probability) between any two possible underlying data tables T and T' that only differ in a single record. More formally, the randomized function κ guarantees ϵ-differential privacy if for every pair of data tables T and T' differing in only one row and for all $S \subseteq Range(\kappa)$ the following holds [Dwo11]:

$$Pr[\kappa(T) \in S] \leq e^{\epsilon} \cdot Pr[\kappa(T') \in S] \tag{6.30}$$

parameter ϵ is public and must be selected according to privacy and utility requirements. As ϵ increases, the allowable difference between the two probabilities in Equation 6.30 increases exponentially and an adversary can distinguish between the two possible reconstructions T and T' with a higher probability. Therefore, ϵ is a *utility enhancer* (see Section 4.1).

One realization of the function κ adds independent noise to query result sets with a Laplace distribution. This distribution has a parameter z that must be chosen based on what the noise is designed to hide. Differential privacy aims at concealing the signs of existence/nonexistence of a data provider's record in the data table. Let R_i be the record of personal information about data provider i. Let T be a data table without record R_i and T' be the same data table that also includes the record R_i (*i.e.*, $T' = T \cup \{R_i\}$). The difference between outputs of a query Q over T and T' is the sign that Laplace noise must hide to protect data provider i's privacy. To generalize this concept over all possible individuals, the *sensitivity* of query Q, denoted by ΔQ, is defined as the maximum difference between any of the two outputs it might generate over any two data tables T and T' differing in only one

record.

It has been shown [Dwo11, DMNS06, Dwo06] that a Laplace noise with parameter $z = \frac{\Delta Q}{\epsilon}$, denoted by $Lap(\frac{\Delta Q}{\epsilon})$, can provide ϵ-differential privacy for a query Q with sensitivity ΔQ. The proof is also available in Section 2.2.2.2. As a result, in this mechanism if a data user asks query Q with sensitivity ΔQ from the data collector (in the interactive settings the data user does not have direct access to the data table), the data collector runs the query on the original data table, adds noise with the Laplace distribution $Lap(\frac{\Delta Q}{\epsilon})$ to the results, and returns the noisy answer to the data user.

If more than one query is asked by the data user, then the noise added to the output of *each* query must be chosen based on the additive sensitivities of all queries [Dwo11]. Let the sequence of Q_1, Q_2, ..., Q_ρ be the queries a data user is willing to run on the data table. The data collector must know the sensitivity ΔQ_i of each query Q_i in the sequence *before answering any of the queries* to compute the parameter $z = \frac{\sum_{1 \leq i \leq \rho} \Delta Q_i}{\epsilon}$. After announcing all of the sensitivities, the data user asks each query Q_i and the data collector adds noise with distribution $Lap(z)$ to the query results. Therefore, even though the data collector and the data user interact in ongoing rounds of query-response, the decision on the privacy protection (on values of ϵ and additive sensitivity $\sum_{1 \leq i \leq \rho} \Delta Q_i$) is made *once and for all* at round zero of the interactions. This property of differential privacy enables us to model it within our game theoretic framework to find stable results at round zero. Future interactions between the data collector and the data user (in a query-response fashion) do not affect the privacy protection level of the data table as long as the data collector verifies that the requested queries have the initially acknowledged sensitivities. As a result, the game only models round zero of the interactions where the negotiation over the privacy protection level occurs.

6.2.2 Differential Privacy Game Model

Parameters of differential privacy affect the privacy protection and the precision of the data analysis differently. Parameter ϵ is the only parameter that affects privacy protection level

of a data table. In this sense we can define the privacy parameter as $\delta = \langle \epsilon \rangle$. However, the amount of noise added to an output and hence the query precision depends on both ϵ and the additive query sensitivities $\sum \Delta Q_i$. Therefore, we can alternatively define the privacy parameter as $\delta = \langle \sum \Delta Q_i,\ \epsilon \rangle$. Although the latter alternative is more comprehensive, as we see later in this section the former is expressive enough to address privacy parameter setting challenges using a game model. As a result, we consider $\delta = \langle \epsilon \rangle$ for differential privacy.

Game theory can be applied to find stable values of ϵ in differential privacy. This section explains how to tailor the generic game model from Section 4.2 to the specifications of differential privacy. To avoid repeating content, we mainly focus on function instantiations and those elements of the game that are exclusive to differential privacy. The interested readers are encouraged to refer to Section 4.2 if more details about the game's components are required.

6.2.2.1 Players and Payoffs

As with the generic game model, this game has three groups of players: n data providers DP, a data collector DC, and a data user DU.

Data Providers - A data provider is a player who is informed that differential privacy is the data sanitization mechanism. Based on the announced values of ϵ for differential privacy and the incentive I, she decides whether or not to provide her personal information. As mentioned in Chapter 4, we assume the identity function $g(I) = I$ to explain the rewarding effects of the incentive on a privacy pragmatist's tradeoff analysis. However, the data providers' privacy gain function $h(\delta) = h(\epsilon)$ must be defined based on the role of ϵ in differential privacy.

A data provider is aware that by providing her information, the answer to some of the queries will be e^{ϵ} more (or less) likely. Therefore, the function e^{ϵ} can be interpreted as a privacy risk indicator. Here we assume $ln100$ as the upper bound of ϵ since a difference of magnitude $e^{\ln 100} = 100$ between likelihood of receiving the same output over two data tables

T and T' (differing in one row) is a clear indication of existence (or non-existence) of an individual's record in the data set. Notice that this upper bound is still much higher than Dwork's guess of $\ln 3$ [Dwo11]. As a result, we adopt the function $h(\epsilon) = 100 - e^{\epsilon}$ as the privacy gain function to assess trade-offs of privacy pragmatists and fundamentalists.

Data User - A data user is the entity who plans to query a private data table. Cardinality of the data table, queries she asks from the data collector, the amount of noise added to the output of each query, economic value of each record, and the price she pays for each record contribute to the payoff of a data user. Section 4.2 defines a function $EN(N)$ to explain the effective cardinality of a data table with N records after considering Min and Max number of records required by the data user. Here, we just use the simplifying assumption of $Min = 0$ and $Max = \infty$. With this assumption, the function $EN(N) = N$ explains the effective cardinality of the data table. If the data user pays price p for each record, her expenditure is:

$$expenditure_{DU} = p \cdot N \tag{6.31}$$

We consider a situation where a data user is willing to run one or more queries on the data table each with the same sensitivity ΔQ. We also assume that queries are equally important[2]. Let b represent the net revenue of each data record if the data user runs a query with sensitivity 1 on the data table for free. For a dataset of cardinality N, this revenue is $b \cdot N$. If the data user chooses to ask ρ queries each with sensitivity ΔQ, her income is $\rho \cdot \Delta Q \cdot b \cdot N$. In differential privacy,based on the value of ϵ some noise is added to the output of each query and therefore the income of the data provider decreases by a coefficient we refer to as $Precision$. Consequently we define the income of a data user as:

$$income_{DU} = \rho \cdot \Delta Q \cdot b \cdot N \cdot Precision(\epsilon) \tag{6.32}$$

[2]To model a more general case where each query has a different sensitivity and importance, the same approach must be taken to model data user's income. However, U_{DU} becomes a piecewise function with a sub function for every possible subset of queries chosen by the data user. The same approach can be used to solve this generic case but the solution becomes more complicated and has several cases.

The definition of the function *Precision* must be chosen based on implementation of differential privacy. With the Laplace mechanism, to achieve differential privacy a noise with Laplace distribution $Lap(\frac{\rho \cdot \Delta Q}{\epsilon})$ is added to the output of each query. This noise has the variance $2 \cdot [\frac{\rho \cdot \Delta Q}{\epsilon}]^2$ and standard deviation $\sqrt{2} \cdot [\frac{\rho \cdot \Delta Q}{\epsilon}]$. The standard deviation of the noise is considered as the expected error of the query [Dwo11]. Therefore, the reciprocal of this expected error can be used as an indicator of precision. When we plug $\frac{\epsilon}{\sqrt{2} \cdot \rho \cdot \Delta Q}$ as precision in Equation 6.32, the terms $\rho \cdot \Delta Q$ in the formula cancel each other:

$$\rho \cdot \Delta Q \cdot b \cdot N \cdot Precision(\epsilon) = [\rho \cdot \Delta Q] \cdot b \cdot N \cdot \frac{\epsilon}{\sqrt{2} \cdot [\rho \cdot \Delta Q]} = b \cdot N \cdot \frac{\epsilon}{\sqrt{2}} \tag{6.33}$$

This equation means that a data user will eventually be indifferent between choosing any subset of the queries to ask. As she asks more queries, more noise will be added to the outputs and the gain of asking more queries is cancelled by the loss of getting noisier outputs. Notice that this result is very well aligned with the original goal of differential privacy which is to prevent an adversary from increasing her odds of guessing private information by asking more queries or repeating the same query multiple times. Since the number of queries ρ and their sensitivities ΔQ do not affect the data user's payoff function, we only model the game with privacy parameter $\delta = \langle \epsilon \rangle$. The data user must still inform the data collector of the value of $\rho \cdot \Delta Q$ before starting to ask the queries but this information does not affect the decision of any of the players in the game and that is why we can model the game without considering ρ and ΔQ.

Based on Equation 6.33 precision is represented by the term $\frac{\epsilon}{\sqrt{2}}$. With the assumption of $\epsilon \leq \ln 100$ we can normalize the values of precision into the range of $[0, 1]$ by dividing $\frac{\epsilon}{\sqrt{2}}$ by $\frac{\ln 100}{\sqrt{2}}$. After this adjustment, we can rewrite the income to a data user as follows:

$$income_{DU} = b \cdot N \cdot \left[\frac{\epsilon}{\ln 100}\right] \tag{6.34}$$

Consequently, the payoff to the data user can be defined as:

$$U_{DU} = \left[b \cdot \left[\frac{\epsilon}{\ln 100}\right] - p\right] \cdot N \tag{6.35}$$

Data Collector- A data collector, is the player who makes certain privacy and incentive promises to data providers to collect their personal information. Once data is collected, the data collector interacts with the data user in multiple rounds of query-response to provide the output of queries requested by the data user. Since the data user pays price p for each record, the income of a data collector is $p \cdot EN(N) = p \cdot N$. We assume a fixed cost C for collecting the data table, storing it, and answering the data user's queries.[3] Besides C, the data collector is paying each data provider the incentive I. Therefore, the payoff to a data collector is defined as follows:

$$U_{DC} = [p - I] \cdot N - C \tag{6.36}$$

6.2.2.2 Rules of the Game

Differential privacy is an interactive mechanism and unlike anonymization techniques communication between the data user and the data collector does not terminate after the data table is collected (and anonymized). Interactions between the data user and the data collector start with the data user making an offer. In this offer the data user specifies the price she is willing to pay for each record, ϵ, and the additive sensitivities of all queries she is going to run. If the data collector accepts such an offer the contract between the data user and the data collector is sealed. The data collector starts collecting personal information and waits for the data user to ask her queries. From this point on, any query that the data user asks must conform to the agreed upon contract. In other words, as soon as the additive sensitivities of the queries submitted by the data user exceed the agreed on number in the contract, the data collector refuses to answer the query (as well as any subsequent query). In this sense, after sealing the contract, if the data collector is honest, subsequent interactions between the data user and the data collector can only affect the data user herself. Therefore, these interactions do not need to be included in the game model and this game has the exact

[3] If this cost becomes an increasing function of the number of queries then the data user always choses to ask minimum number of queries.

same rules as explained in the generic model in Section 4.2.2. Since the additive sensitivity value $\rho \cdot \Delta Q$ is only used for verifications in future rounds of interactions and does not affect any player's payoff, we can denote the offer of a data user as $Of = \langle \epsilon, p \rangle$ without explicitly including the additive sensitivity of queries. However, in real instances of such negotiation, the data user's offer must have the additive sensitivity as a third element.

6.2.3 Subgame Perfect Equilibria in Differential Privacy

When differential privacy is the underlying data sanitization method, the game's subgame perfect equilibria can be found by directly following the generic approach explained in Section 4.3. The details of applying the backward induction method closely resembles the game analysis provided for k-anonymity (see Section 6.1.3). However, due to the inherent differences between k-anonymity and differential privacy, the privacy protection level and query precision are measured differently between the two methods. These differences lead to different best response functions for the players. In this section, we mainly provide the final function definitions of players' best responses and refer the reader to Section 6.1.3 for the justifications.

Within the game tree, the data collection subgames are the smallest subgames in which each data provider chooses between opting in and out for the announced combination of ϵ and incentive I. As explained in Section 6.2.2.1 we use the function $h(\epsilon) = 100 - e^{\epsilon}$ to measure the privacy gain when analyzing tradeoffs of privacy pragmatists and fundamentalists. Consequently, similar to Equation 6.17 in the context of k-anonymity, the mixed strategy of a data provider is to opt in with the following probability:

$$Prob(OptIn) = p_1 \cdot 1 + p_2 \cdot \frac{w_1 \cdot [100 - e^{\epsilon}] + w_2 \cdot I}{\overline{v}} + p_3 \cdot \frac{100 - e^{\epsilon}}{\overline{v}'} \tag{6.37}$$

where $\overline{v} = \max_{\epsilon,I}\{w_1 \cdot [100 - e^{\epsilon}] + w_2 \cdot I\}$ and $\overline{v}' = \max_{\epsilon}\{100 - e^{\epsilon}\} = 100 - e^{0} = 99$. After replacing the constant term and coefficients of $h(\epsilon)$ and I with the expressions $\beta_0 = p_1$, $\beta_1 = [p_2 \cdot w_1]/\overline{v} + p_3/\overline{v}'$, and $\beta_2 = [p_2 \cdot w_2]/\overline{v}$ in Equation 6.37, we can calculate the expected

cardinality of a dataset after announcing values ϵ and I as follows:

$$N = n \cdot Prob(OptIn) = n \cdot [\beta_0 + \beta_1 \cdot [100 - e^{\epsilon}] + \beta_2 \cdot I] \tag{6.38}$$

By deducing how data providers react to each combination of ϵ and I, the data collector chooses her most profitable action in response to any offer she receives from the data collector. Similar to the analysis explained for k-anonymity in Section 6.1.3, the derivative of function U_{DC} (see Equation 6.36) with respect to I becomes zero when U_{DC} reaches its maximum. Therefore, the optimizing value of incentive $\hat{I}$ can be formulated as:

$$\hat{I} = \frac{p \cdot \beta_2 - \beta_1 \cdot [100 - e^{\epsilon}] - \beta_0}{2 \cdot \beta_2} \tag{6.39}$$

By definition $I \geq 0$. Therefore, if for an offer $Of = \langle p, \epsilon \rangle$ the corresponding $\hat{I}$ is less than zero, the most profitable incentive for the data collector would be setting $I = 0$. Based on the value of $\hat{I}$, the data collector's best response to a combination of $\langle \epsilon, p \rangle$ can be defined as:

$$BR_{DC} = \begin{cases} \{I = \frac{p \cdot \beta_2 - \beta_1 \cdot [100 - e^{\epsilon}] - \beta_0}{2 \cdot \beta_2}\} & if\ \beta_2 \cdot p \geq \beta_1 \cdot [100 - e^{\epsilon}] + \beta_0 > 2.\sqrt{\frac{C \cdot \beta_2}{n}} - \beta_2 \cdot p \\ \{I = \frac{p \cdot \beta_2 - \beta_1 \cdot [100 - e^{\epsilon}] - \beta_0}{2 \cdot \beta_2}, Reject\} & if\ \beta_2 \cdot p \geq \beta_1 \cdot [100 - e^{\epsilon}] + \beta_0 = 2.\sqrt{\frac{C \cdot \beta_2}{n}} - \beta_2 \cdot p \\ \{I = 0\} & if\ \beta_1 \cdot [100 - e^{\epsilon}] + \beta_0 > \max\{\beta_2 \cdot p,\ \frac{C}{n \cdot p}\} \\ \{I = 0, Reject\} & if\ \frac{C}{n \cdot p} = \beta_1 \cdot [100 - e^{\epsilon}] + \beta_0 > \beta_2 \cdot p \\ \{Reject\} & Otherwise \end{cases} \tag{6.40}$$

Considering the best response function BR_{DC} of the data collector from Equation 6.40 and assuming that the data collector accepts the offer when she is indifferent, we can deduce the expected cardinality of the dataset for each offer $Of = \langle \epsilon, p \rangle$ as follows:

$$N = \begin{cases} \frac{n}{2} \cdot [\beta_0 + \beta_1 \cdot [100 - e^{\epsilon}] + \beta_2 \cdot p] & if\ \beta_2 \cdot p \geq \beta_1 \cdot [100 - e^{\epsilon}] + \beta_0 \geq 2 \cdot \sqrt{\frac{C \cdot \beta_2}{n}} - \beta_2 \cdot p \\ n \cdot [\beta_0 + \beta_1 \cdot [100 - e^{\epsilon}]] & if\ \beta_1 \cdot [100 - e^{\epsilon}] + \beta_0 \geq \max\{\beta_2 \cdot p,\ \frac{C}{n}\} \\ 0 & Otherwise \end{cases} \tag{6.41}$$

When Equation 6.41 is used to define N, the data user's payoff function (see Equation 6.35) becomes a function of only two variables: ϵ and p. The best strategy for the data user is to compute $\hat{\epsilon}$ and $\hat{p}$ such that they maximize her payoff:

$$\langle \hat{\epsilon}, \hat{p} \rangle = \arg\max_{\epsilon,p} U_{DU} = \arg\max_{\epsilon,p} \left[b \cdot \frac{\epsilon}{\ln 100} - p \right] \cdot N(\epsilon, p) \tag{6.42}$$

Maximizing combinations of ϵ and p represent the optimum offers and determining these combinations completes the process of finding the game's subgame perfect equilibria.

6.2.4 Simulation Results for Differential Privacy

The game's subgame perfect equilibria suggest stable prescriptions for the privacy parameter ϵ for each instance of the problem. They can also be used to study how different problem settings affect steady values of the privacy parameter. As explained in Section 6.1.4, since an analytical approach is used to solve the game, experiments to verify the outcome is unnecessary; as long as players are rational and the settings of a privacy negotiation matches our assumptions, the game's equilibria would be the same as the ones suggested in Section 6.2.3. However, to provide a more concrete illustration of the game's outcome, we have modeled the game for some synthetic problem settings and provided the diagrams to show how aspects such as output perturbation cost, population size, privacy awareness, and privacy trust influence stable values of ϵ in the game's subgame perfect equilibria.

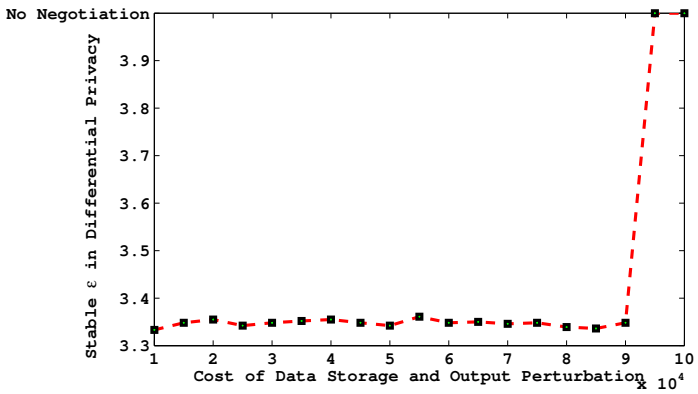
In the experiments, an instance of the problem with $n = 20,000$, $b = \$10.00$, and $C = \$20,000$ is used. A data provider is of any of the privacy unconcerned, privacy pragmatist, or privacy fundamentalist types, with probabilities $p_1 = 0.16$, $p_2 = 0.59$, and $p_3 = 0.25$ (following Westin's indexes). We also use $w_1 = 25$ and $w_2 = 1$ to define a privacy pragmatist's indifference curve.[4] In each experiment, all of these parameters are fixed except for the subjects of the test. The sample problem setting is deliberately chosen to be as similar as possible to the setting we used to run simulations on k-anonymity. This choice allows for

[4]This setting reflects findings of Grossklags and Acquisti [GA07].

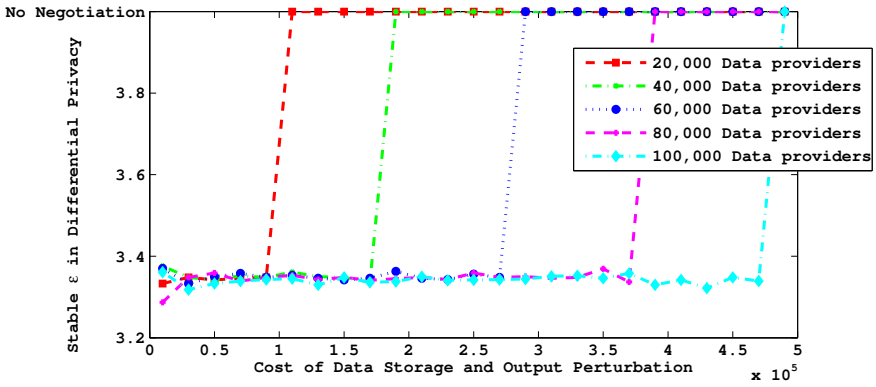
meaningful comparison between diagrams in this section and diagrams provided in Section 6.1.4.

Diagrams represented in Figure 6.5 show stable values of ϵ for different costs, C of data storage and output perturbation. In Figure 6.5(a), we see that increasing the cost C up to a certain point does not largely affect stable values of ϵ. In fact, the data user mostly compensates by offering a higher price per record rather than asking for more private query results. Such behavior is caused by the shape of the privacy gain function $h(\epsilon) = 100 - e^{\epsilon}$ at equilibrium values of ϵ ($3.3 \leq \epsilon \leq 3.4$). The derivative of function $h(\delta)$ exponentially decreases as ϵ increases. Therefore, outside the range of the equilibrium values, a decrease in ϵ does not largely increase the cardinality of the dataset and an increase in the value of ϵ significantly reduces the number of data providers who opt in. We conclude that similar to k-anonymity, increasing the cost of data storage and adding noise to query outputs does not lead to higher levels of privacy. Notice that in Figure 6.5(a) there is much less zigzag effects compared to Figure 6.3(a). This is because values of ϵ are continuous but values of k in k-anonymity are discrete. Analogous to k-anonymity, increasing the cost over a certain point ($90,000$ in Figure 6.5(a)) causes the game to enter the "No Negotiation" stage, where all profitable combinations of $\langle \epsilon, p \rangle$ for the data user will be rejected by the data collector. Compared to the corresponding diagram in k-anonymity, for the same problem settings, higher costs of data storage and sanitization are justifiable in differential privacy. (In our simulation for k-anonymity the game enters the "No Negotiation" stage for $C \approx 40,000$ while in the differential privacy instance $C \approx 90,000$ causes the same effect). This dissimilarity is caused by the difference between characteristics of $h(\epsilon)$ versus $h(k)$ and the definitions of the *Precision* functions in differential privacy versus k-anonymity. Because of the exponential rate of change in $h(\epsilon)$, stable values of ϵ generally lead to a higher cardinality for the data table compared to stable values of k in k-anonymity. Therefore, when the cost C increases, the effects of a small adjustment in price is magnified by the larger value for the expected

cardinality of the dataset and can still provide the data collector with a non-negative profit. As the total number of data providers n increases, the game enters the "No Negotiation" stage with a higher value of C. Therefore, with a larger population of data providers, higher costs of data storage and output perturbation are justifiable. This effect is captured in Figure 6.5(b) where each curve represents stable values of ϵ for different sizes of the data providers' population.



(a)



(b)

Figure 6.5: Changes to stable values of ϵ due to an increase in: (a) in the cost of data storage and data perturbation C; (b) the cost C of data perturbation and population of data providers n.

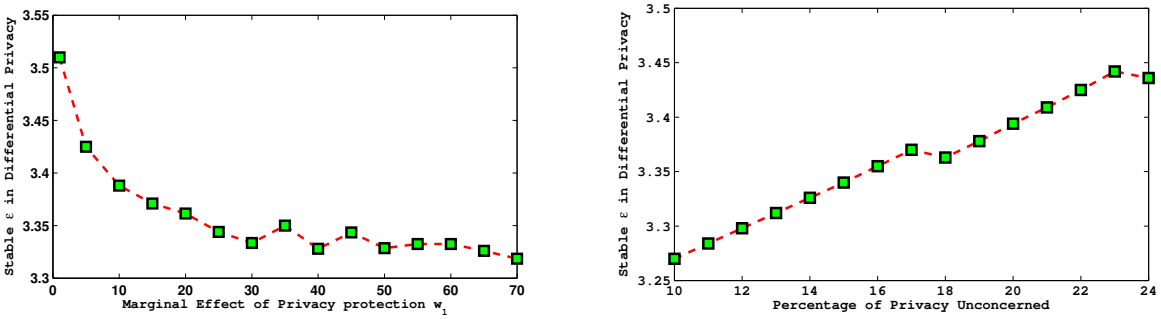
Figures 6.6(a) and 6.6(b) show the effects of data providers' privacy awareness and trust (in the data collector and privacy protection mechanism) on stable values of ϵ. As explained in Section 6.1.4, we model an increase in privacy awareness by increasing the ratio w_1/w_2

within the indifference curves of the privacy pragmatists. Since w_2 is always fixed to the value 1, a higher value of w_1 means that a privacy pragmatist requires a higher amount of incentive in exchange for giving up a unit of privacy protection. Figure 6.6(a) shows the influences of privacy awareness on stable values of ϵ. As w_1 increases (privacy awareness increases), promising higher privacy protection (lower ϵ) becomes more effective (compared to promising higher incentive) in collecting more records. Therefore, a population of data providers with higher privacy awareness encourages the data provider to make offers with higher privacy. On the contrary, Figure 6.6(b) shows how lower privacy protection levels would be offered to a more trusting population of data providers.[5] In a population of data providers who have higher trust in the data collector and the privacy protection mechanism, the percentage of privacy fundamentalists (who are paranoid about risks to their privacy) is expected to be lower. Therefore, a more trusting population of data providers is modeled with a higher percentage of privacy unconcerned (and lower percentage of privacy fundamentalists). With a larger population of privacy unconcerned, a data user can make the same profit without offering a high privacy protection level. These diagrams conform to our findings in k-anonymity on positive effects of the public's privacy awareness and distrust in receiving higher privacy protection.

6.3 Game Theory and Comparing Sanitization Mechanisms: An Illustration

Numerous sanitization methods have been proposed so far to protect privacy of individuals who have their personal information stored in data repositories. A subset of these methods are introduced in Section 2.2 but the provided list is by no means exhaustive. Every method introduces its own privacy protection parameters and is claimed to be superior to previously proposed methods until some of its shortcomings are discovered and a new sanitization

[5]In both Figures 6.6(a) and 6.6(b), jumping points in the curves happen where the data user finds it more beneficial to change the price rather than changing the value of ϵ.



(a) (b)

Figure 6.6: Changes to stable values of ϵ due to an increase in: (a) data providers' privacy awareness; (b) data providers' trust in the data collector and the privacy protection method.

method is designed to overcome the challenges. However, just because a second method addresses some issues that are overlooked in the first method we cannot conclude that the second method outperforms the first one.

Perhaps, an insuperable obstacle in providing a comprehensive framework for comparing different data sanitization methods is that each method embodies a unique and often incomparable privacy parameter. The details of the mechanisms and their outputs are often so dissimilar that one might argue that comparing these methods is analogous to "comparing apples to oranges". However, when a choice is to be made between buying an apple or an orange, making such comparison is inevitable. Unlike the common belief, it is not unreasonable to compare apples to oranges as long as a common criterion is chosen for the argument. For example, apples are red and oranges are orange (comparison based on color). Similarly, after defining the common grounds, we can examine different sanitization methods in contrast. The common grounds can be the data mining utility of the sanitized dataset, the dataset's vulnerabilities to the same class of attacks, or financial profitability of the methods (for different stakeholders of the dataset). To define these common evaluation criteria, some literature [SSNDA10, ASNB12] propose pragmatic and information theoretic approaches to evaluate privacy and utility levels of published datasets after applying various data sanitization methods. However, due to the nature of data sanitization methods, the

measurements provided in the literature inevitably depend on the privacy parameter values of each method. Therefore, even after choosing a common criterion for our comparison, we cannot directly use the measurements and declare that method x provides better privacy than method y. Instead, an appropriate usage of the existing measurements is to make the comparison statement that "when value δ_x is chosen for privacy parameter in method x, it provides a better privacy protection compared to situations where method y is used with privacy parameter δ_y". This brings out the question of which combinations of δ_x and δ_y should be used to compare methods x and y? Our game theoretic model can be used to suggest a possible answer to this question and put the final piece of the puzzle in place by nominating the stable values of δ_x and δ_y (found at the games' equilibria) as the reference points for the comparison.

Having specific problem settings for each sanitization method, *separate* instances of the game can be defined and solved to find steady values for their privacy parameters. It is crucial to note that the game instances are modeled and analyzed separately. Therefore, although the stable privacy parameter values (found at each game's equilibria) are influenced by the data providers' privacy decisions, in none of the scenarios the data providers are required to choose between two (or more) data sanitization methods. Once the stable privacy parameter values are plugged in each method, one can choose a common criterion such as the offered level of privacy protection or the method's profitability to compare sanitization methods (possibly based on the existing measurements in the literature). For example, a data user can compare solutions of the games for 2 (or more) different sanitization mechanism and choose the one that provides her with the highest payoff at the equilibrium point. The game model also provides the opportunity to conduct more complicated comparisons between sanitization methods. For instance, based on curves of stable privacy parameter values for different privacy awareness levels, one can investigate the effectiveness of investing on educating data providers about privacy risks involved in each privacy protection method.

In this section, we show how solutions of the sanitization games can be used to compare differential privacy and k-anonymity in a sample problem setting. Building a comprehensive comparison framework based on the game's results and the existing literature is left as a future extension to this work.

6.3.1 A Sample Comparison: k-Anonymity Vs. Differential Privacy

As a sample problem setting, we consider a situation where a data user intends to run `SELECT` queries of the type specified in Section 6.1.2.2 (for k-anonymity). These queries are equivalent to requesting a single histogram query that counts the number of records with each possible combination of values for `sensitiveAtt` and quasi-identifying attribute `q`. As Dwork shows [Dwo11], the entire histogram query has sensitivity $\Delta Q = 1$. Therefore, precision is estimated for the `SELECT` query in k-anonymity and the histogram for differential privacy. Moreover, as noted in the literature [Dwo11], statisticians generally prefer to have access to a dataset as a "table" rather than being allowed to only ask questions. Therefore, we choose a setting where the data user associates a higher economic value to each record if the sanitization method is k-anonymity (value of parameter b is fixed to \$26 in k-anonymity and to \$10 in differential privacy). Other common problem parameters between the two games are chosen to be the same and their values are equal to the ones explained in Sections 6.1.4 and 6.2.4.

Figure 6.7 illustrates how equilibrium values of the data user's payoff changes as the data providers' trust level increases. These changes are shown for both the k-anonymity and the differential privacy methods. The two curves cross each other when 16% of the population fall into the privacy unconcerned category (*i.e.*, there are 25% privacy fundamentalists within the data providers population). According to this diagram, when faced with a low trust population of data providers (when less than 16% of data providers are privacy unconcerned), a data user can make higher profit if differential privacy is used as the data sanitization method. However, as trust increases among data providers (a population with less than 25%

privacy fundamentalists), k-anonymity becomes a more profitable privacy protection option for the data user. As a result, if the data user has the option to choose between the two methods (when making her offer to the data collector), she would choose differential privacy for a low-trust, and k-anonymity for a high-trust community of data providers. This shows that depending on the current perception of the two methods in data providers population, differential privacy might not always be the winning method to provide privacy protection.

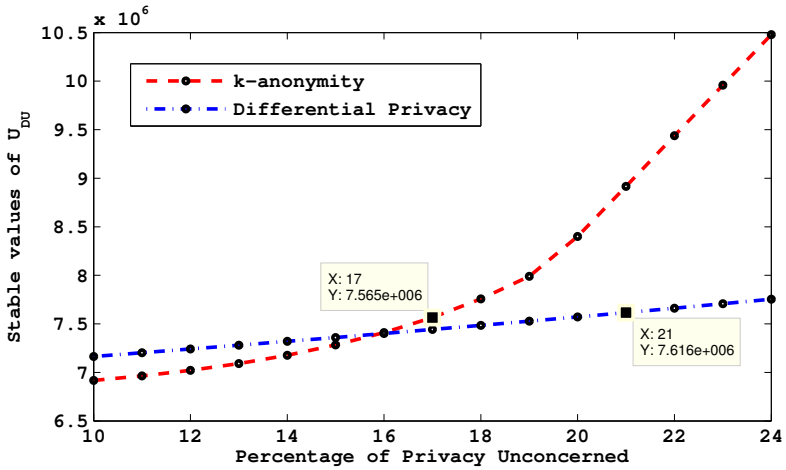
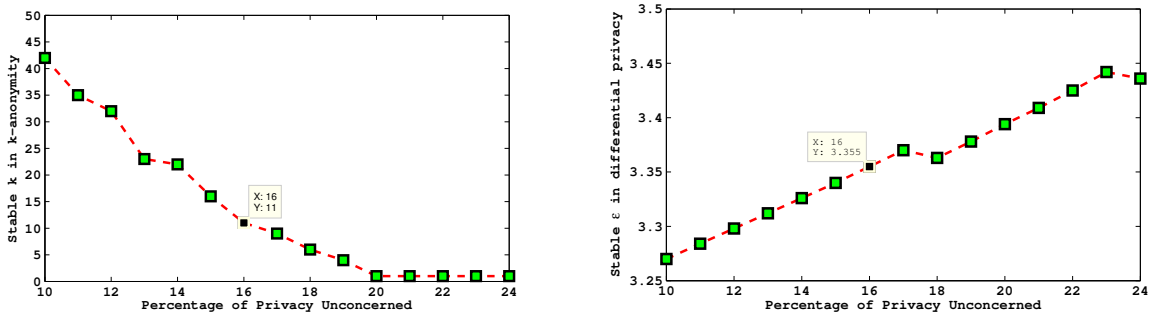


Figure 6.7: Data user's equilibrium payoff values for k-anonymity and differential privacy as data providers' trust level increases.

Another interesting result illustrated in Figure 6.7 is the effects of trust across two different sanitization method. Consider a situation where data providers have a higher trust in the differential privacy than in the k-anonymity method. This advantage of differential privacy can sometimes convince the data user to choose differential privacy over k-anonymity even in a high trust population of data providers. For example, consider a population of data providers such that when the k-anonymity method is used 24% of them feel paranoid about risks to their privacy while this percentage is 20% when differential privacy is used. This situation models a population of high trust for both k-anonymity and differential privacy, but data providers are assumed to have higher trust in differential privacy. With this setting, it can be seen in Figure 6.7 that the data user's payoff is still higher if she chooses

differential privacy over k-anonymity (even though she is facing a high-trust population of data providers). This example shows that using a privacy mechanism in which the public has higher trust can lead to a higher profit (up to a certain point).

Stable values of k and ϵ found through our game theoretic analysis can be used as concrete reference points for comparing privacy protection capabilities of k-anonymity and differential privacy. For instance, if the real world problem settings match the sample settings we used for this simulation then for a population following Westin's index of privacy (16% privacy unconcerned, 59% privacy pragmatists, and 25% privacy fundamentalists) a stable ϵ for differential privacy is $\epsilon = 3.35$ while a stable k for k-anonymity is $k = 11$ (see Figures 6.8(a) and 6.8(b).). To provide a reasonable comparison between differential privacy and k-anonymity (within the specified context), a framework that uniformly measures privacy and utility offered by the two methods (such as [SSNDA10, ASNB12]) can be used to clarify which one of the differential privacy with $\epsilon = 3.35$ or k-anonymity with $k = 11$ are more prone to privacy attacks.



(a) (b)

Figure 6.8: Effects of data providers' trust on stable values of : k in k-anonymity (a); and ϵ in differential privacy (b).

6.4 Summary

This Chapter described how the generic game model can be instantiated for two seminal data sanitization methods: k-anonymity (as a non-interactive method) and differential privacy (as an interactive method). We showed how *Precision* can be estimated based on the implementation of the methods and data applications. Following the guidelines provided for solving the generic game model in Chapter 4, the subgame perfect equilibria in both games are found. The games' solutions are simulated for sample problem settings to study the effects of sanitization costs, population of data providers, and data providers' privacy awareness and trust on the stable values of k and ϵ. The last Section of this Chapter, discusses how the solutions of privacy setting games can be used as reference points to compare different privacy protection methods. A sample comparison between k-anonymity and differential privacy is provided.

Chapter 7

Summary and Concluding Remarks

This study set out to guide the process of setting privacy parameters so that a balanced level of privacy/utility tradeoff is achieved. In this last Chapter we conclude by summarizing the progress made towards this goal in terms of a game that models the challenge and its instantiations to various privacy protection mechanisms. We also highlight the major contributions of this study and suggest some future research venues that can enhance the comprehensiveness of the model and provide directions on how to exploit the game's solutions in different applications.

7.1 Realization of the Research Objective

Massive personal data collection on the Web raises privacy concerns among data providers. To address these concerns, several privacy protection mechanisms have been proposed. Each privacy protection mechanism is associated with at least one privacy parameter that controls how much privacy is provided to the data providers. As the the guaranteed level of privacy (adjusted by the value of the privacy parameters) increases, more data manipulation is applied to the collected information and hence information utility of the dataset decreases. The tradeoff between conflicting needs for privacy and information utility was chosen as the research problem in this thesis.

To address the research problem, this thesis aimed at answering the question "How to establish a value for the privacy parameter to achieve a balanced tradeoff between privacy and utility?" The answer to this question must explain the meaning of a *balanced* privacy utility tradeoff and use this definition to develop a model that produces the balancing value for the privacy parameter based on the description of a privacy protection mechanism. Con-

sequently, the goal of this study was to propose an all-embracing framework to find balancing values of the privacy parameter in each privacy protection mechanism.

The underlying conception of *balanced* privacy/utility tradeoff is very critical to this study. Contrary to common practices, the notion of "optimum" privacy/utility combination is neither a clear nor a reasonable interpretation of a balanced tradeoff. Optimality "is in the eye of the beholder." An optimal combination of privacy/utility for users of the dataset is not necessarily optimal for a provider of the information. Therefore, even if a combination of privacy and utility is optimal for one party, the combination may never be achieved due to the inter-dependencies between decisions made by all stakeholders of the information. A more reasonable concept to pursue is the "stability" of the privacy/utility tradeoff. In particular, in this study, a balanced tradeoff is defined as the outcome of a state at which attempts of different parties to maximize their gain will reach an equilibrium.

The interpretation used for a balanced tradeoff and stable privacy parameter value immediately suggested the application of game theory. A generic game model is used here to find stable privacy parameter settings. The provided game is an extensive form game with incomplete and imperfect information. Players of the game were defined to be 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*. The data user attempts to propose the most profitable offer to the data collector, the data collector aims at maximizing her benefit for each offer by choosing the most profitable amount of incentive (or rejecting the offer), and the data providers choose to opt in if providing their information is their most rewarding choice.

The privacy parameter values found in the game's subgame perfect equilibria denote the stable privacy settings and represent shared agreements (or *consensus*) in which none of the players would attempt to behave differently. We used the generalized method of backward

induction to find the game's subgame perfect equilibria. The process of solving the game started with finding data providers' best responses to each combination of privacy and incentive. In this study, multiple types and subtypes were considered for the data providers and the mathematical expectation of the event that a randomly chosen data provider accepts the privacy/incentive combination was found. This expected value represents the best response of a data provider in terms of the probability of the event that she opts in for the announced combination of privacy and incentive. With this probability, the expected cardinality of the dataset is determined. When a data collector receives an offer, her best response was defined as choosing an incentive that maximizes her payoff considering the expected cardinality of the dataset. Multiple cases have to be considered to ensure that all semantic criteria hold while choosing the most profitable incentive. The last phase of solving the game was to find the data user's most profitable offers with regard to the best responses of the data collector and data providers. The winning offers specified the values of privacy parameter that are in the game's subgame equilibria.

We explained how the proposed all-embracing framework serves as a template that can capture the features of any privacy protection method (defined as either a privacy policy declaration or a data sanitization approach) with minor adjustments. In each instantiation of the game for an arbitrary privacy protection method: (1) the method's effective privacy parameter components were recognized and formalized as a vector, and (2) an information utility function was defined to estimate the precision of the target query's result set. The estimates are specific to each method and depend on the algorithm used for data masking.

In Chapters 5 and 6 after finding the game's solution in three different privacy protection methods (one from each class of data protection methods), we studied the game's equilibria considering some synthetic game settings. In the analysis, we showed how the stable values of a privacy parameter change as the problem settings change (such as the cost of privacy protection, population of data providers, and data providers' privacy awareness and trust

factor). These results are of independent use and further prove the interdependencies between party's decisions. Consequently, by proposing a generic game model and providing its solution in various applications, the research question is answered and the project objective is achieved.

7.2 Main Contributions

In this study we provided a novel game model, the solution of which suggests stable values for privacy parameters. We hope that this work could be the first step towards developing a comprehensive guidelines on choosing privacy parameter values that balance the privacy/utility tradeoff. The most important contributions of this thesis can be summarized as follows:

Adopting a Novel Perspective on the Dynamics of Private Data Collection:

Often, the existing literature on privacy/utility tradeoff approaches the challenge by contrasting the usefulness of data for an authorized data user (good utility) against its usefulness for a malicious attacker (bad utility) [LS08, LL09, SSNDA10, MKS11, ZO10]. These studies, provide in-depth theoretical and/or pragmatic evaluations of the levels of privacy and information utility provided by different sanitization methods with various settings. While these analyses are of independent interest and can provide some insight on how much privacy is sacrificed to achieve a certain level of information utility, their application in *choosing* a balanced privacy/utility tradeoff is still not clear. The studies in this area are either silent about this important decision or rely on some weights and thresholds to be assigned to the privacy and utility requirements. As a result, the final decision on privacy setting is still left to the data collector's personal opinion. Two drawbacks of delegating the the privacy setting decision to intuitions of a data collector *after* the information is collected are: (1) the data collector does not have enough justifications to support any of her decisions, and (2) in the event of a privacy breach, it will not be clear whether the data collector is to be blamed or not.

As an alternative approach, we tried to focus on explaining the data collector's true motivation behind providing a better privacy to data providers. Without discussing this overlooked issue, the advantage of providing better privacy protection cannot be clearly defined and compared against its disadvantage in decreasing information utility. In this thesis, attracting more data providers (and collecting a larger data table) is explained as the underlying motivation behind promising a better privacy protection. To the best of our knowledge, this is the first work that shifts the focus from an attacker to the opinion of the data providers in searching for a balancing privacy parameter value.

Proposing a Generic Game Model to Establish Stable Privacy Parameters:

Rather than fixing a privacy protection level and searching for the maximum information utility, this work proposes a process to find balanced combinations of privacy and utility. we adopted the solution concept of Nash equilibrium as the interpretation of a balanced tradeoff. This interpretation is new and we believe it is providing a different perspective because in any other combination, there is at least one party who has enough motivation to make a different decision and change the outcome. We provided an original game model that incorporates the decisions made by three different stakeholders of the private information: A data user, a data collector, and data providers. The solution to this game provides stable privacy parameter values that lead to a balanced privacy/utility tradeoff. The proposed game model is independent of any specific privacy protection mechanism.

Instantiating and Solving the Game for Three Privacy Protection Methods:

The broad range of applicability of our generic game model is shown by instantiating it for a sample privacy declaration method in Chapter 5 and two sample data sanitization methods (k-anonymity as an interactive and differential privacy as a non-interactive method) in Chapter 6. For each of these methods, we provided particular functions to estimate the amount of precision and privacy implications of the privacy parameters. The generic backward induction mechanism was used to find the games' subgame perfect equilibria.

Examining the Effects of Problem Settings on Stable Privacy Parameter Values:

In the game model, several parameters such as Min, Max, data user's valuation for each record, cost of protecting privacy of the data table, total number of data providers, and distribution of different types of data providers were described as properties of the problem settings. Parametric description of the game solutions provided the opportunity of assessing the impact of these parameters on the stable values of privacy parameters. For a sample privacy declaration method (in Chapter 5), a synthetic setting shows how values of Min, data user's valuation for records at partial and exact granularity levels, cost of privacy protection, and sensitivity of the requested data field can affect the stable privacy policies. With regard to data sanitization methods, the two seminal methods of k-anonymity and differential privacy are considered and the impacts of data sanitization cost and total population of data providers on the stable values of k and ϵ are discussed. More particularly, we found that increasing the cost over a threshold takes the game to the phase of no negotiation (impractical privacy protection) and this phase is delayed by having a larger population of data providers. For all of the three chosen privacy protection mechanisms, we also demonstrated how data providers' higher privacy awareness and lower trust (in the privacy protection mechanism) can lead to a stable privacy parameter that guarantees a higher privacy protection level. This fact provides further evidence of data providers' critical role in assessing privacy/utility tradeoff.

Utilizing the Games' Results in Comparing Privacy Protection Mechanisms:

Comparing different privacy protection mechanisms is a challenging but necessary problem to address. The mechanisms make different assumptions about the attackers' back ground knowledge, employ disparate processes to mask data, and embody their own distinct privacy parameters. A necessary step towards facilitating such a comparison is to develop metrics to uniformly measure the level of privacy and utility offered by each method at each privacy parameter value (see [SSNDA10, ASNB12] for such metrics). However, to use these

measurements practically, for each of the methods, we need to know which pair of measured privacy and utility levels must be examined in contrast to the pairs chosen for other methods. Using the notion of balanced privacy/utility tradeoff and the generic game model, we proposed the stable values of privacy parameters as suitable candidate reference points for such comparison. At these reference points two different privacy protection mechanisms can be compared based on how much utility they offer to the data user and how much privacy protection they promise (according to the measurements provided in the existing literature). As an illustration, we showed how the game model can be used to compare k-anonymity and differential privacy. With the help of an example, we also showed how the data providers' trust in different methods can determine the most profitable privacy protection mechanism.

Bringing Three Independent Bodies of Research Together:

Privacy is a multi-dimensional challenge studied in the realms of legal sciences, social sciences, economic, and computer science. In social science studies, the focus has been on providing indexes on data providers' privacy concerns and recognizing the factors that influence their privacy behaviour. Economic literature on privacy mainly revolves around profitability of providing privacy protection and game theory is one of the most commonly used tools in such an assessment. However, the economic literature mostly lack an in-depth analysis of the technological implications and the implementation of privacy protection mechanisms. We applied game theory and used the economic perspective of choosing the most profitable option for each player, given the available privacy protection mechanisms and their implementation details. We also used the social science findings to model different types of data providers for the game. Consequently, our approach is influenced by three different aspects of privacy (economic, social, and technological) and brings them together to provide a better understanding of the challenge.

7.3 Recommendations for Future Work

To the best of our knowledge, this study is the first attempt at finding balanced privacy/utility tradeoff using game theory and considering the effects of data providers' decisions. The proposed model still has a number of limitations. We identify several points of improvement to evolve this model towards a more practical and general solution guideline. The limitations, possible improvements, and unexplored potential applications of the model suggest variety of research directions that need to be pursued.

Although we applied some findings from the social science literature to model various types and subtypes of data providers, we made a number of assumptions that are yet to be proven via empirical studies on public's privacy behavior. We have assumed a uniform distribution for thresholds of privacy pragmatists and privacy fundamentalists. Moreover, to model the indifference curve for the privacy pragmatist group, we used a linear combination of privacy gain and rewarding effects of the incentive. The functions for privacy gain were chosen to be reasonable estimates based on the definition of each privacy protection mechanism. These assumptions limit practicality of the game and we have not established whether they reflect the reality. Another limitation of this study was the very large number of game's equilibria. This work only studied those equilibria in which a data collector accepts the offers whenever she is indifferent between accepting or rejecting. For future work it is recommended to examine validity of the assumptions (using experimental studies) and explore other equilibria of the game.

The proposed game model was designed with the simplest structure that is capable of capturing the essence of the privacy/utility tradeoff problem. This simplicity was necessary to establish the foundations of our novel approach in using game theory to balance the tradeoff of privacy versus utility. Several improvements are recommended to expand the model into a more realistic reflection of real-world situations. The one-shot negotiation between the data user and the data collector can be changed to infinite rounds of negotiations. This

improved version must consider the negative effects of long negotiations which eventually makes the most impatient party to stop the bargaining process. This negotiation can also capture the competition between multiple data users over a single data table. A game that models such competition, provides more reliable results but increases the complexity of the calculations. In the proposed game model, the data collector is always trustworthy. An improved version of this game can drop this assumption and extend the game to multiple rounds of repeated games (probably with different data users). With the iterative model, it is expected that the data collector acts honest to establish a good reputation for future interactions with the same data providers. The exact implementation and implications of this behaviour are interesting venues for future research. Unlike the approach used in this study, in real-life instances of the problem, a data user may not know the exact payoff function of a data collector. Therefore, another extension point to this work is to consider multiple types for the data collector each with a different payoff function. Moreover, throughout this work, we have always assumed that the data user and the data collector have the same belief about data providers' types and distributions. A more challenging analysis would be to consider different beliefs and solve the game. Finally, as mentioned in Chapter 4, this game only models the situations where data providers make their decisions individually and independent of each other. In some environments (especially social networking) this is not the case and a data provider's decision to opt-in encourages other data providers to opt in. Modelling these environments is another recommended future direction.

The main purpose of the proposed game model was to guide data collectors on choosing a balancing value for a privacy parameter. However, this model can potentially help choosing a more profitable privacy protection mechanism. In fact, a more comprehensive game can be built in which a data user starts the game by choosing the most profitable privacy protection method (among some candidate methods). Each emanating branch from this root node takes us to an instance of the proposed game that is instantiated for the chosen privacy protection

method. Once a parametric solution to this game is found, it can be used to partition the problem space into various classes. The conditions and equilibria in each class can shed light on the question: "In what situations method x is the data user's most profitable choice?" This question is still an open problem and we believe our model can possibly help to find an answer to it.

Bibliography

[ACdVS08] Claudio Agostino Ardagna, Marco Cremonini, Sabrina De Capitani di Vimercati, and Pierangela Samarati. A privacy-aware access control system. *Journal of Computer Security*, 16(4):369–397, 2008.

[ACR99] Mark S. Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM conference on Electronic commerce*, EC '99, pages 1–8. ACM, 1999.

[AG05] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 3(1):26–33, 2005.

[AHK+03] Paul Ashley, Satoshi Hada, Günter Karjoth, Calvin Powers, and Matthias Schunter. Enterprise Privacy Authorization Language (EPAL 1.2). Technical report, IBM, 2003.

[AKSX02] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. Hippocratic databases. In *Proceedings of the 28th international conference on Very Large Data Bases*, VLDB '02, pages 143–154. VLDB Endowment, 2002.

[And06] Horace E. Anderson. The privacy gambit: Toward a game theoretic approach to international data protection. *Vanderbilt Journal of Entertainment and Technology Law*, 9(1), 2006.

[ASNB12] Mina Askari, Reihaneh Safavi-Naini, and Ken Barker. An information theoretic privacy and utility measure for data sanitization mechanisms. In *Proceedings of the second ACM conference on Data and Application Security and Privacy*, CODASPY '12, pages 283–294. ACM, 2012.

[BA05] Roberto J. Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 217–228, 2005.

[BAB+09] Ken Barker, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam Majedi, Sampson Pun, and Adepele Williams. A data privacy taxonomy. In *Proceedings of the 26th British National Conference on Databases*, BNCOD 26, pages 42–54. Springer-Verlag, 2009.

[BK07] Rainer Böhme and Sven Koble. On the viability of privacy-enhancing technologies in a self-regulated business-to-consumer market: Will privacy remain a luxury good? In *6th Annual Workshop on the Economics of Information Security*, WEIS 2007, 2007.

[BP09] Joseph Bonneau and Sören Preibusch. The Privacy Jungle: On the Market for Data Protection in Social Networks. In *Proceedings of The Eighth Workshop on the Economics of Information Security*, WEIS 2009, pages 121–167. Springer US, 2009.

[CA99] Mary J. Culnan and Pamela K. Armstrong. Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science*, 10(1):104–115, January 1999.

[CDE+06] Lorrie Faith Cranor, Brooks Dobbs, Serge Egelman, Giles Hogben, Jack Humphrey, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, Joseph M. Reagle, Matthias Schunter, David A. Stampley, and Rigo Wenning. The platform for privacy preferences 1.1 (p3p1.1) specification. World Wide Web Consortium, Note NOTE-P3P11-20061113,`http://www.w3.org/TR/P3P11/`, November 2006. W3C Recommendation.

[CLM02a] Lorrie Faith Cranor, Marc Langheinrich, and Massimo Marchiori. A p3p preference exchange language 1.0 (appel 1.0). World Wide Web Consortium, Working Draft WD-P3P-preferences-20020415 , http://www.w3.org/TR/2002/WD-P3P-preferences-20020415, 2002.

[CLM+02b] Lorrie Faith Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph M. Reagle. The platform for privacy preferences 1.0 (p3p1.0) specification. World Wide Web Consortium, Recommendation REC-P3P-20020416, http://www.w3.org/TR/2002/REC-P3P-20020416, April 2002.

[CLS11] Jan Camenisch, Ronald Leenes, and Dieter Sommer, editors. *Digital Privacy - PRIME - Privacy and Identity Management for Europe*, volume 6545 of *Lecture Notes in Computer Science*. Springer, 2011.

[CP01] Giacomo Calzolari and Alessandro Pavan. Optimal design of privacy policies. Technical report, Gremaq, University of Toulouse, 2001.

[CRA00] Lorrie Faith Cranor, Joseph Reagle, and Mark S. Ackerman. Beyond concern: Understanding net users' attitudes about online privacy. In *The Internet Upheaval: Raising Questions, Seeking Answers in Communications Policy*, pages 47–70. MIT Press, 2000.

[DF09] Josep Domingo-Ferrer. The functionality-security-privacy game. In *Proceedings of the 6th International Conference on Modeling Decisions for Artificial Intelligence*, MDAI '09, pages 92–101. Springer-Verlag, 2009.

[DFS09] Josep Domingo-Ferrer and Yücel Saygin. Editorial: Recent progress in database privacy. *Data Knowl. Eng.*, 68(11):1157–1159, November 2009.

[DHR00] Yevgeniy Dodis, Shai Halevi, and Tal Rabin. A cryptographic solution to a game theoretic problem. In *Proceedings of the 20th Annual International Cryptology*

Conference on Advances in Cryptology, CRYPTO '00, pages 112–130. Springer-Verlag, 2000.

[Dix90] Avinash Kamalakar Dixit. *Optimization in Economic Theory*. Oxford University Press, 2 edition, 1990.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third conference on Theory of Cryptography*, TCC'06, pages 265–284. Springer-Verlag, 2006.

[Dut99] Parjit K. Dutta. *Strategies and Games: Theory and Practice*. MIT Press, Cambridge, Mass. [u.a.], 1999.

[Dwo06] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II*, ICALP '06, pages 1–12, 2006.

[Dwo11] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, January 2011.

[EG95] Edwin J. Elton and Martin J. Gruber. *Modern portfolio theory and investment analysis*. Portfolio Management Series. Wiley, 5. ed edition, 1995.

[fECOD02] Organisation for Economic Co-Operation and Development. *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. OECD Publishing, 2002.

[FMHP09] Julien Freudiger, Mohammad Hossein Manshaei, Jean-Pierre Hubaux, and David C. Parkes. On non-cooperative location privacy: a game-theoretic analysis. In *Proceedings of the 16th ACM conference on Computer and communications security*, CCS '09, pages 324–337. ACM, 2009.

[FWY05] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Top-down specialization for information and privacy preservation. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 205–216. IEEE Computer Society, 2005.

[GA07] Jens Grossklags and Alessandro Acquisti. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *6th Annual Workshop on the Economics of Information Security*, WEIS 2007, 2007.

[GD08] Gabriele Gianini and Ernesto Damiani. A game-theoretical approach to data-privacy protection from context-based inference attacks: A location-privacy protection case study. In *Secure Data Management*, volume 5159 of *Lecture Notes in Computer Science*, pages 133–150. Springer Berlin Heidelberg, 2008.

[HA96] Louis Harris and Associates. *1996 Equifax-Harris Consumer Privacy Survey*. Equifax Inc, 1996.

[HAF05] Bernardo A. Huberman, Eytan Adar, and Leslie R. Fine. Valuating privacy. *IEEE Security & Privacy*, 3(5):22–25, 2005.

[Har67] John C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii part i. the basic model. *Management Science*, 14(3):159–182, 1967.

[Har68a] John C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii. part iii. the basic probability distribution of the game. *Management Science*, 14(7):486–502, 1968.

[Har68b] John C. Harsanyi. Games with incomplete information played by "bayesian" players part ii. bayesian equilibrium points. *Management Science*, 14(5):320–334, 1968.

[Har73] John C. Harsanyi. Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory*, 2(1):1–23, December 1973.

[JAL09] Leslie K. John, Alessandro Acquisti, and George F. Loewenstein. The Best of Strangers: Context Dependent Willingness to Divulge Personal Information. *Social Science Research Network Working Paper Series*, July 2009.

[KABSN12] Rosa Karimi Adl, Mina Askari, Ken Barker, and Reihaneh Safavi-Naini. Privacy consensus in anonymization systems via game theory. In *Data and Applications Security and Privacy XXVI - 26th Annual IFIP WG 11.3 Conference (DBSec 2012)*, volume 7371 of *Lecture Notes in Computer Science*, pages 74–89. Springer, 2012.

[Kat08] Jonathan Katz. Bridging game theory and cryptography: recent results and future directions. In *Proceedings of the 5th conference on Theory of cryptography*, TCC'08, pages 251–272. Springer-Verlag, 2008.

[KC05] Ponnurangam Kamaraguru and Lorrie Faith Cranor. Privacy indexes: A survey of westin's studies. Technical Report CMU-ISRI-5-138, Institute for Software Research International, Carnegie Mellon University, 2005.

[KDL07] Hillol Kargupta, Kamalika Das, and Kun Liu. Multi-party, privacy-preserving distributed data mining using a game theoretic framework. In *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD 2007, pages 523–531. Springer-Verlag, 2007.

[KM11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, SIGMOD '11, pages 193–204. ACM, 2011.

[KMN05] Krishnaram Kenthapadi, Nina Mishra, and Kobbi Nissim. Simulatable auditing. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '05, pages 118–127. ACM, 2005.

[Kob07] Alfred Kobsa. Privacy-enhanced web personalization. In *The Adaptive Web: Methods and Strategies of Web Personalization*, chapter 21, pages 628–670. Springer-Verlag, 2007.

[KPM11] Charles A. Kamhoua, Niki Pissinou, and Kia Makki. Game theoretic modeling and evolution of trust in autonomous multi-hop networks: Application to network security and privacy. In *Proceedings of IEEE International Conference on Communications*, ICC 2011, pages 1 –6. IEEE, 2011.

[KPR00] Jon Kleinberg, Christos Papadimitriou, and Prabhakar Raghavan. Auditing boolean attributes. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '00, pages 86–91. ACM, 2000.

[KPR01] Jon Kleinberg, Christos H. Papadimitriou, and Prabhakar Raghavan. On the value of private information. In *Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge*, TARK '01, pages 249–257. Morgan Kaufmann Publishers Inc., 2001.

[LDR05a] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, SIGMOD '05, pages 49–60. ACM, 2005.

[LDR05b] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Multidimensional *k*-anonymity. Technical Report 1521, University of Wisconsin, 2005.

[LDR06a] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, ICDE '06, pages 25–36. IEEE Computer Society, 2006.

[LDR06b] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 277–286, 2006.

[LDR08] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems*, 33:17:1–17:47, September 2008.

[LL09] Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 517–526. ACM, 2009.

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *International Conference on Data Engineering*, ICDE '07, pages 106–115, 2007.

[LMPS04] Matt Lepinski, Silvio Micali, Chris Peikert, and Abhi Shelat. Completely fair sfe and coalition-safe cheap talk. In *Proceedings of the twenty-third annual ACM symposium on Principles of distributed computing*, PODC '04, pages 1–10. ACM, 2004.

[LS08] Grigorios Loukides and Jianhua Shao. Data utility and privacy protection trade-off in k-anonymisation. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, PAIS '08, pages 36–45. ACM, 2008.

[LWFP06] Jiuyong Li, Raymond Chi-Wing Wong, Ada Wai-Chee Fu, and Jian Pei. Achieving k-anonymity by clustering in attribute hierarchical structures. In *Proceedings of the 8th international conference on Data Warehousing and Knowledge Discovery*, DaWaK'06, pages 405–416, 2006.

[MG93] George R. Milne and Mary Ellen Gordon. Direct mail privacy-efficiency tradeoffs within an implied social contract framework. *Journal of Public Policy & Marketing*, 12(2):206–215, 1993.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3:1–3:52, 2007.

[MKS11] Ashwin Machanavajjhala, Aleksandra Korolova, and Atish Das Sarma. Personalized social recommendations: accurate or private. *Proceedings of the VLDB Endowment*, 4(7):440–450, 2011.

[MN12] Sara Miner More and Pavel Naumov. Calculus of cooperation and game-based reasoning about protocol privacy. *ACM Trans. Comput. Logic*, 13(3):22:1–22:21, August 2012.

[Mos05a] Tim Moses. eXtensible access control markup language TC v2.0 (XACML). OASIS Standard, `http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf`, 2005.

[Mos05b] Tim Moses. Privacy policy profile of xacml v2.0. OASIS Standard, `http://docs.oasisopen.org/xacml/2.0/access_control-xacml-2.0-privacy_profile-spec-os.pdf`, 2005.

[MP05] Marco Casassa Mont and Siani Pearson. An adaptive privacy management system for data repositories. In *Proceedings of the Second international conference*

on Trust, Privacy, and Security in Digital Business, TrustBus'05, pages 236–245. Springer-Verlag, 2005.

[MR11] Atsuko Miyaji and Mohammad Shahriar Rahman. Privacy-preserving data mining: a game-theoretic approach. In *Proceedings of the 25th annual IFIP WG 11.3 conference on Data and applications security and privacy*, DBSec'11, pages 186–200. Springer-Verlag, 2011.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103. IEEE Computer Society, 2007.

[MY04] Edwin Mansfield and Gary Wynn Yohe. *Microeconomics: Theory/Applications*. W.W. Norton, 11, illustrated edition, 2004.

[NMK+06] Shubha U. Nabar, Bhaskara Marthi, Krishnaram Kenthapadi, Nina Mishra, and Rajeev Motwani. Towards robustness in query auditing. Technical Report 2006-16, Stanford InfoLab, June 2006.

[Osb03] Martin J. Osborne. *An Introduction to Game Theory*. Oxford University Press, USA, August 2003.

[Pre06] Sören Preibusch. Implementing privacy negotiations in e-commerce. In *Proceedings of the 8th Asia-Pacific Web conference on Frontiers of WWW Research and Development*, APWeb'06, pages 604–615. Springer, 2006.

[Rot00] Marc Rotenberg. *Privacy Law Sourcebook 2000: United States Law, International Law, and Recent Developments*. Electronic Privacy Information Center, 2000.

[RS10] Lisa Rajbhandari and Einar Arthur Snekkenes. Using game theory to analyze risk to privacy: An initial insight. In *Privacy and Identity Management for Life*,

volume 352 of *IFIP Advances in Information and Communication Technology*, pages 41–51. Springer Boston, 2010.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[SGB01] Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, EC '01, pages 38–47. ACM, 2001.

[SH95] Knut Sydsaeter and Peter J. Hammond. *Mathematics for economic analysis.* Prentice-Hall International editions. Prentice-Hall International, 1995.

[SMC93] Eleanor Singer, Nancy A. Mathiowetz, and Mick P. Couper. The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 u.s. census. *The Public Opinion Quarterly*, 57(4):465–482, 1993.

[Sol08] Daniel J. Solove. *Understanding privacy.* Number v. 10 in Understanding privacy. Harvard University Press, 2008.

[SS98] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, PODS '98, page 188, 1998.

[SSNDA10] Michal Sramka, Reihaneh Safavi-Naini, Jörg Denzinger, and Mina Askari. A practice-oriented framework for measuring privacy and utility in data sanitization systems. In *Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT '10, pages 27:1–27:10. ACM, 2010.

[SSP09] Anna Cinzia Squicciarini, Mohamed Shehab, and Federica Paci. Collective privacy management in social networks. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 521–530. ACM, 2009.

[Swe00] Latanya Sweeney. Uniqueness of simple demographics in the US population. *LIDAP-WP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA*, 2000.

[Swe02a] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10:2002, 2002.

[Swe02b] Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[Tur03] Joseph Turow. Americans & online privacy: The system is broken. Technical report, Annenberg Public Policy Center, University of Pennsylvania, 2003.

[Var09] Hal R. Varian. Economic aspects of personal privacy. In *Internet Policy and Economics*, pages 101–109. Springer US, 2009.

[Var10] Hal R. Varian. *Intermediate Microeconomics: A Modern Approach*. W.W. Norton & Company, 2010.

[XT06] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 229–240. ACM, 2006.

[ZO10] Marek P. Zielinski and Martin S. Olivier. On the use of economic price theory to find the optimum levels of privacy and information utility in non-perturbative

microdata anonymisation. *Data and Knowledge Engineering*, 69(5):399–423, 2010.

Appendix A

Complete Proof of Theorem 5.3.1

Theorem 5.3.1. Let $o1 = \langle\langle g_1, \cdots, g_j, \cdots, g_m, r\rangle, p\rangle$ be an offer such that at least one of the g_i's with $i \neq j$ is set to a level higher than zero. Recall that A_j is the data field over which the predicate of the `COUNT`-query is defined. The data user can do at least as good as $o1$ by making an offer $o2 = \langle\langle 0, 0, \cdots, g_j, \cdots, 0, r\rangle, p\rangle$ or $o2' = \langle\langle 0, 0, \cdots, g_j, \cdots, 0, r\rangle 0\rangle$.

Proof. Consider the description of α_o given in Equation 5.13. Since all parameters τ_0,..., τ_m are greater than zero, function $h_g(\cdot)$ is a strictly decreasing function of granularity level, and there is at least one g_i that is zero in $o2$ but more than zero in $o1$, we have $\alpha_{o1} < \alpha_{o2}$. Moreover, as one of the g_i's in $o2$ changes from zero to another granularity level in $o1$, $CG(g_i)$ from Equation 5.7 either increases or stays the same. Thus, the inequality $C_{o2} \leq C_{o1}$ holds. With these two facts we show that for all meaningful combinations of cases (from Section 5.3.2):

- **part 1 -** If the data collector accepts $o1$ her will also accept $o2$.
- **part 2 -** Cardinality of the data table, N, after accepting offer $o2$, is expected to be at least as large as accepting offer $o1$.
- **part 3 -** By offering $o2$ or $o2'$, the data user's profits are at least as large as her profit when she offers $o1$.

Since $\alpha_{o1} < \alpha_{o2}$ not all combination of cases apply to offers $o1$ and $o2$. All possible combinations are enumerated in Table A.1. For each combination, we prove parts 1 and 2 and then justify part 3 of the theorem.

Proof of parts 1 and 2 :

To prove parts 1 and 2 for each applicable combination of cases, we first explain how accepting

Table A.1: Possible combinations of cases for $o1$ and $o2$

Case for $o2$ \ Case for $o1$	1a	1b	1c	2a	2b	2c	3
1a	✓	✗	✗	✗	✗	✗	✗
1b	✓	✓	✗	✗	✗	✗	✗
1c	✓	✓	✓	✗	✗	✗	✗
2a	✓	✓	✗	✓	✓	✗	✗
2b	✓	✓	✗	✗	✓	✗	✗
2c	✓	✓	✓	✗	✓	✓	✗
3	✓	✓	✓	✓	✓	✓	✓

both offers $o1$ and $o2$ with the optimum incentive, affect the data collector's maximum payoff and cardinality of the data table. Then we use facts about conditions that apply to the offers and the shape of payoff functions to show that the data collector's payoff after accepting $o2$, is at least as large as her payoff in case of accepting $o1$.

Case 1a for *o1* vs. case 1a for *o2* - If the data collector accepts offer $o1$ in case 1a, then her payoff $\hat{U}_{DC}^{1a}$ for offer $o1$ is greater than or equal to zero[1]. In other words:

$$0 \leq Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} = U_{DC}[o1] \tag{A.1}$$

Since $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, we have:

$$Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} \leq Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o2}}{n \cdot \gamma}\right] - C_{o2} = U_{DC}[o2] \tag{A.2}$$

Consequently, the data collector would also accept offer $o2$.

According to Table 5.1 if the data collector accepts any offer in case 1a the expected cardinality of the dataset will be Min. Therefore, the expected cardinality of the data table after accepting offer $o2$ is at least as large as offer $o1$.

Case 1a for *o1* vs. case 1b for *o2* - If the data collector accepts offer $o1$ in case 1a, then her payoff $\hat{U}_{DC}^{1a}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq Min \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} = U_{DC}[o1] \tag{A.3}$$

[1]In the rest of the proof, we denote the data collector's payoff for offer o as $U_{DC}[o]$ and don't include the superscripts. The superscripts can be deduced based on the case that applies to the offer.

Based on Equation 5.16, the maximum of the U_{DC} in case 1 happens at $\hat{I} = \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}$ and this maximum yields $\hat{U}_{DC}^{1b}$. Therefore for every offer, $\hat{U}_{DC}^{1a} \leq \hat{U}_{DC}^{1b}$. We have:

$$Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \right] - C_{o1} \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2} \right]^2 - C_{o1} \tag{A.4}$$

Since $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, we have:

$$0 \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2} \right]^2 - C_{o1} \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o2} + \gamma \cdot p}{2} \right]^2 - C_{o2} = \ U_{DC}[o2] \tag{A.5}$$

Therefore, $o2$ will be accepted.

According to Table 5.1 if the data collector accepts an offer in case 1a the cardinality of the dataset will be Min and if her accepts an offer in case 1b the expected cardinality would be $n.\left[\frac{\alpha_{o2}+\gamma \cdot p}{2}\right]$. Since Case 1b applies to offer $o2$, the condition for this case is satisfied:

$$\frac{Min - n \cdot \alpha_{o2}}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma} \Rightarrow Min \leq n \cdot \left[\frac{\gamma \cdot p + \alpha_{o2}}{2} \right] \tag{A.6}$$

Consequently, the expected cardinality of the data table after accepting offer $o2$ is at least as large as accepting offer $o1$.

Case 1a for *o1* vs. case 1c for *o2* - If the data collector accepts offer $o1$ in case 1a, then the data collector's payoff $\hat{U}_{DC}^{1a}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \right] - C_{o1} = U_{DC}[o1] \tag{A.7}$$

Since $Min \leq Max$ and $C_{o2} \leq C_{o1}$, we have:

$$Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \right] - C_{o1} \leq Max \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \right] - C_{o2} \tag{A.8}$$

We now show that $p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \leq p - \frac{Max - n \cdot \alpha_{o2}}{n \cdot \gamma}$ and therefore $0 \leq U_{DC}[o1] \leq U_{DC}[o2]$.

Since case 1c applies to $o2$, based on the condition of this case in Table 5.1 we have:

$$\frac{Max - n \cdot \alpha_{o2}}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma} \tag{A.9}$$

Since $\alpha_{o1} < \alpha_{o2}$, we also have:

$$\frac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma} \tag{A.10}$$

Finally, condition 1a that applies to $o1$ requires:

$$\frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma} \leq \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \tag{A.11}$$

Based on the past three inequalities, we have:

$$\begin{aligned}
&\tfrac{Max - n\cdot\alpha_{o2}}{n\cdot\gamma} \leq \tfrac{Min - n\cdot\alpha_{o1}}{n\cdot\gamma} \Rightarrow \\
&-\tfrac{Min - n\cdot\alpha_{o1}}{n\cdot\gamma} \leq -\tfrac{Max - n\cdot\alpha_{o2}}{n\cdot\gamma} \Rightarrow \\
&p - \tfrac{Min - n\cdot\alpha_{o1}}{n\cdot\gamma} \leq p - \tfrac{Max - n\cdot\alpha_{o2}}{n\cdot\gamma} \Rightarrow \\
&Min \cdot \left[p - \tfrac{Min - n\cdot\alpha_{o1}}{n\cdot\gamma}\right] - C_{o1} \leq Max \cdot \left[p - \tfrac{Max - n\cdot\alpha_{o2}}{n\cdot\gamma}\right] - C_{o2} \Rightarrow \\
&0 \leq U_{DC}[o1] \leq U_{DC}[o2]
\end{aligned} \tag{A.12}$$

Therefore, $o2$ will be accepted.

According to Table 5.1 if the data collector accepts an offer in case 1a the dataset's cardinality will be Min and if she accepts an offer in case 1c the expected cardinality would be Max. Since $Min \leq Max$, the expected cardinality of the data table after accepting offer $o2$ is at least as large as offer $o1$.

Case 1a for *o1* vs. case 2a for *o2* - If the data collector accepts offer $o1$ in case 1a, then her payoff $\hat{U}_{DC}^{1a}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n\gamma}\right] - C_{o1} = U_{DC}[o1] \tag{A.13}$$

Since $C_{o2} \leq C_{o1}$, we have:

$$0 \leq Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} \leq Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o2} \tag{A.14}$$

According to Table 5.1 the condition for case 1 is $n \cdot \alpha_{o1} < Min$. Therefore, by substituting the term $n \cdot \alpha_{o1}$ with Min we get:

$$\begin{aligned}
Min \cdot \left[p - \tfrac{Min - n\cdot\alpha_{o1}}{n\cdot\gamma}\right] - C_{o2} &\leq Min \cdot \left[p - \tfrac{Min - Min}{n\cdot\gamma}\right] - C_{o2} \\
&= Min \cdot p - C_{o2}
\end{aligned} \tag{A.15}$$

Finally, the condition for case 2 (in Table 5.2) is $Min \leq n \cdot \alpha_{o2}$. With this property we substitute the term Min with $n \cdot \alpha_{o2}$ to get the following:

$$Min \cdot p - C_{o2} \leq n \cdot \alpha_{o2} \cdot p - C_{o2} = U_{DC}[o2] \tag{A.16}$$

Based on the past four inequalities we conclude that the data collector's payoff $\hat{U}_{DC}^{2a}$ for offer $o2$ is greater than or equal to zero.

According to Table 5.1 if the data collector accepts an offer in case 1a the dataset's cardinality will be Min and if her accepts an offer in case 2a the expected cardinality would be $n \cdot \alpha_{o2}$. The condition for case 2 (see Table 5.2) is $Min \leq n \cdot \alpha_{o2}$. Therefore, the expected cardinality of the dataset after accepting offer $o2$ is at least as large as offer $o1$.

Case 1a for *o1* vs. case 2b for *o2* - If the data collector accepts offer $o1$ in case 1a, then she would also accept offer $o2$. The proof of this claim is identical to "case 1a for $o1$ vs. case 1b for $o2$".

According to Table 5.1 if the data collector accepts an offer in case 1a the dataset cardinality will be Min and if her accepts in case 2b the expected cardinality would be $n \cdot \left[\frac{\alpha_{o2}+\gamma \cdot p}{2}\right]$. Part of the condition for case 2b (see Table 5.2) is:

$$\begin{aligned} & 0 \leq \tfrac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma} && \Rightarrow \\ & \alpha_{o2} \leq \gamma \cdot p && \Rightarrow \\ & n \cdot \left[\tfrac{\alpha_{o2}+\alpha o2}{2}\right] \leq n \cdot \left[\tfrac{\alpha_{o2}+\gamma \cdot p}{2}\right] && \Rightarrow \\ & n \cdot \alpha_{o2} \leq n \cdot \left[\tfrac{\alpha_{o2}+\gamma \cdot p}{2}\right] \end{aligned} \tag{A.17}$$

Moreover, the condition for case 2 is $Min \leq n \cdot \alpha_{o2}$. Therefore, the expected database size after accepting offer $o2$ is at least as large as offer $o1$.

Case 1a for *o1* vs. case 2c for *o2* - If the data collector accepts offer $o1$ in case 1a, she would also accept the offer $o2$ in case 2c because $\hat{U}_{DC}^{2c} = \hat{U}_{DC}^{1c}$ and the proof is identical to "case 1a for $o1$ vs. case 1c for $o2$".

The expected cardinality of the dataset after accepting offer $o2$ (in case $2c$) is at least the same as the expected cardinality of the dataset if $o1$ is accepted (in case $1a$) since $Min \leq Max$.

Case 1a for *o1* vs. case 3 for *o2* - If the data collector accepts offer $o1$ in case 1a,

then her payoff $\hat{U}_{DC}^{1a}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \right] - C_{o1} = U_{DC}[o1] \tag{A.18}$$

The condition for case 1 is $n \cdot \alpha_{o1} < Min$ (see Table 5.1). As a result:

$$\begin{array}{l} 0 \leq \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \quad \Rightarrow \\ -\frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \leq 0 \end{array} \tag{A.19}$$

Consequently, the following holds:

$$Min \cdot \left[p - \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \right] - C_{o1} \leq Min \cdot p - C_{o1} \tag{A.20}$$

Since $Min \leq Max$ and $C_{o2} \leq C_{o1}$, we have:

$$Min \cdot p - C_{o1} \leq Max \cdot p - C_{o2} = U_{DC}[o2] \tag{A.21}$$

The equations prove that in this case offer $o2$ would also be accepted.

The expected cardinality of the dataset after accepting offer $o1$ (in case $1a$) is Min (see Table 5.2). The expected cardinality of the dataset after accepting offer $o2$ (in case 3) is $n \cdot \alpha_{o2}$ (see Table 5.3). Since $Min \leq Max \leq n \cdot \alpha_{o2}$ (the condition for case 3), if offer $o2$ is accepted the expected cardinality is at least the same as offer $o1$.

Case 1b for *o1* vs. case 1b for *o2* - If the data collector accepts offer $o1$ in case 1b, then her payoff $\hat{U}_{DC}^{1b}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2} \right]^2 - C_{o1} = U_{DC}[o1] \tag{A.22}$$

Since $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, we have:

$$\frac{n}{\gamma} \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2} \right]^2 - C_{o1} \leq \frac{n}{\gamma} \cdot \left[\frac{\alpha_{o2} + \gamma \cdot p}{2} \right]^2 - C_{o2} = U_{DC}[o2] \tag{A.23}$$

Therefore, the data collector would also accept offer $o2$.

The expected cardinality of the dataset if offer $o1$ is accepted would be $n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2} \right]$. The dataset cardinality would be $n \cdot \left[\frac{\alpha_{o2} + \gamma \cdot p}{2} \right]$ if offer $o2$ is accepted . Since $\alpha_{o1} < \alpha_{o2}$, if offer $o2$ is accepted the expected cardinality of the dataset is at least as large as accepting offer $o1$.

Case 1b for *o1* vs. case 1c for *o2* - The payoff in case 1b can be rewritten as:

$$\hat{U}_{DC}^{1b} = \left[p - \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}\right] \cdot n \cdot \left[\alpha_o + \gamma \cdot \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}\right] - C_o \tag{A.24}$$

This equation is the result of plugging the optimum incentive $\hat{I} = \frac{\gamma \cdot p - \alpha_o}{2 \cdot \gamma}$ in the second piece of the U_{DC} function in Equation 5.9.

If the data collector accepts offer $o1$ in case 1b, her payoff $\hat{U}_{DC}^{1b}$ is greater than or equal to zero. In other words:

$$\begin{aligned} &0 \leq \left[p - \tfrac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\alpha_{o1} + \gamma \cdot \tfrac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] - C_{o1} = U_{DC}[o1] \quad \Rightarrow \\ &0 \leq \left[p - \tfrac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\tfrac{\gamma \cdot p}{2} + \tfrac{\alpha_{o1}}{2}\right] - C_{o1} = U_{DC}[o1] \end{aligned} \tag{A.25}$$

According to Table 5.1, part of the condition for case 1b is:

$$\begin{aligned} &\tfrac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma} \leq \tfrac{Max - n \cdot \alpha_{o1}}{n \cdot \gamma} \quad \Rightarrow \\ &\tfrac{\gamma \cdot p}{2} - \tfrac{\alpha_{o1}}{2} + \alpha_{o1} \leq \tfrac{Max}{n} \quad \Rightarrow \\ &n \cdot \left[\tfrac{\gamma \cdot p}{2} + \tfrac{\alpha_{o1}}{2}\right] \leq Max \end{aligned} \tag{A.26}$$

Based on this inequality we have:

$$\left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq \left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot Max - C_{o1} \tag{A.27}$$

Since $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, the following holds:

$$\left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot Max - C_{o1} \leq \left[p - \frac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma}\right] \cdot Max - C_{o2} \tag{A.28}$$

Based on the condition for case 1c in Table 5.1, we know:

$$\begin{aligned} &\tfrac{Max - n \cdot \alpha_{o2}}{n \cdot \gamma} \leq \tfrac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma} \quad \Rightarrow \\ &-\tfrac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma} \leq -\tfrac{Max - n \cdot \alpha_{o2}}{n \cdot \gamma} \end{aligned} \tag{A.29}$$

Consequently we have:

$$\left[p - \frac{\gamma \cdot p - \alpha_{o2}}{2 \cdot \gamma}\right] \cdot Max - C_{o2} \leq \left[p - \frac{Max - n \cdot \alpha_{o2}}{n \cdot \gamma}\right] \cdot Max - C_{o2} = U_{DC}[o2] \tag{A.30}$$

Therefore, the data collector would also accept offer $o2$ since the payoff would be greater than or equal to zero.

The expected cardinality of the dataset is $n \cdot \left[\frac{\alpha_{o1}+\gamma \cdot p}{2}\right]$ in case of accepting offer $o1$, and Max in case of accepting $o2$. Based on Equation A.26, $n \cdot \left[\frac{\alpha_{o1}+\gamma \cdot p}{2}\right] \leq Max$ and accepting offer $o2$ would result in a dataset at least as large as the case where offer $o1$ is accepted.

Case 1b for *o1* vs. case 2a for *o2* - If the data collector accepts offer $o1$ in case 1b, her payoff $\hat{U}_{DC}^{1b}$ (see the version in Equation A.24) is greater than or equal to zero. In other words:

$$0 \leq \left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} = U_{DC}[o1] \tag{A.31}$$

The condition for case 1 is $n \cdot \alpha_{o1} < Min$. Combining this fact with the condition in case 1b (see Table 5.1), proves the following inequality:

$$0 \leq \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma} \tag{A.32}$$

As a result we have:

$$\left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2\gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \tag{A.33}$$

Case 2a for $o2$ (see Table 5.2) implies that $\gamma \cdot p < \alpha_{o2}$. We also know that $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$. Therefore we have:

$$p \cdot n \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot n \cdot \left[\frac{\alpha_{o2}}{2} + \frac{\alpha_{o2}}{2}\right] - C_{o2} = U_{DC}[o2] \tag{A.34}$$

The expected cardinality of the dataset after accepting offer $o1$ would be $n \cdot \left[\frac{\alpha_{o1}+\gamma \cdot p}{2}\right]$, and after accepting offer $o2$ would be $n \cdot \alpha_{o2}$. Since $\alpha_{o1} < \alpha_{o2}$ and $\gamma \cdot p < \alpha_{o2}$ (see condition 2a in Table 5.2), we have:

$$n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right] \leq n \cdot \left[\frac{\alpha_{o2} + \gamma \cdot p}{2}\right] \leq n \cdot \left[\frac{\alpha_{o2} + \alpha_{o2}}{2}\right] = n \cdot \alpha_{o2} \tag{A.35}$$

Therefore, accepting offer $o2$ would result in a dataset at least as large as the case where offer $o1$ is accepted.

Case 1b for *o1* vs. case 2b for *o2* - Since $\hat{U}_{DC}^{1b} = \hat{U}_{DC}^{2b}$ and the proof of "case 1b for $o1$ vs. case 1b for $o2$" only relies on the facts that $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, the proof of this case is identical to the proof of "case 1b for $o1$ vs. case 1b for $o2$".

Case 1b for *o1* vs. case 2c for *o2* - The proof of this case is identical to the proof of "case 1b for $o1$ vs. case 1c for $o2$". This is due to the facts that $\hat{U}_{DC}^{1c} = \hat{U}_{DC}^{2c}$, condition 1c (from Table 5.1) is the same as condition 2c (from Table 5.2), and the other inequalities used to prove "case 1b for $o1$ vs. case 1c for $o2$" are either $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, or related to $o1$.

Case 1b for *o1* vs. case 3 for *o2* - If the data collector accepts offer $o1$ in case 1b, her payoff $\hat{U}_{DC}^{1b}$ (see the version in Equation A.24) is greater than or equal to zero. In other words:

$$0 \leq \left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} = U_{DC}[o1] \tag{A.36}$$

The condition for case 1 is $n \cdot \alpha_{o1} < Min$. Combining this fact with the condition in case 1b (see Table 5.1), proves the following inequality:

$$0 \leq \frac{Min - n \cdot \alpha_{o1}}{n \cdot \gamma} \leq \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma} \tag{A.37}$$

As a result we have:

$$\left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \tag{A.38}$$

Since $C_{o2} \leq C_{o1}$ and $n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] \leq Max$ (see Equation A.26), the following holds:

$$p \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot Max - C_{o2} = U_{DC}[o2] \tag{A.39}$$

This proves that if offer $o1$ is accepted by the data collector, offer $o2$ is also accepted.

The expected cardinality of the dataset after accepting offer $o1$ would be $n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right]$ and after accepting offer $o2$ would be $n\alpha_{o2}$. Based on Equation A.26 we know $n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] \leq Max$. The condition for case 3 (see Table 5.3) is $Max \leq n \cdot \alpha_{o2}$. These two facts prove the following:

$$n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right] \leq Max \leq n \cdot \alpha_{o2} \tag{A.40}$$

Therefore, accepting offer $o2$ would result in a dataset at least as large as the case where offer $o1$ is accepted.

Case 1c for *o1* vs. case 1c for *o2* - If the data collector accepts offer $o1$ in case 1c, then her payoff $\hat{U}_{DC}^{1c}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq Max \cdot \left[p - \frac{Max - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} = U_{DC}[o1] \tag{A.41}$$

Since $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, we have:

$$Max \cdot \left[p - \frac{Max - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} \leq Max \cdot \left[p - \frac{Max - n \cdot \alpha_{o2}}{n \cdot \gamma}\right] - C_{o2} = U_{DC}[o2] \tag{A.42}$$

Therefore, the data collector would also accept offer $o2$.

If any of the offers $o1$ or $o2$ are accepted the expected cardinality of the dataset would be Max.

Case 1c for *o1* vs. case 2c for *o2* - Since the condition 1c (from Table 5.1) is the same as the condition 2c (from Table 5.2) and $\hat{U}_{DC}^{1c} = \hat{U}_{DC}^{2c}$, the proof of this case is identical to "case 1c for $o1$ vs. case 1c for $o2$".

Case 1c for *o1* vs. case 3 for *o2* - If the data collector accepts offer $o1$ in case 1c, then her payoff $\hat{U}_{DC}^{1c}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq Max \cdot \left[p - \frac{Max - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} = U_{DC}[o1] \tag{A.43}$$

Based on the condition for case 1 (see Table 5.1), we know:

$$n \cdot \alpha_{o1} < Min < Max \Rightarrow 0 < Max - n \cdot \alpha_{o1} \Rightarrow -\frac{Max - n \cdot \alpha_{o1}}{n\gamma} < 0 \tag{A.44}$$

Considering this inequality and the fact that $C_{o2} \leq C_{o1}$, we have:

$$Max \cdot \left[p - \frac{Max - n \cdot \alpha_{o1}}{n \cdot \gamma}\right] - C_{o1} \leq Max \cdot p - C_{o1} \leq Max \cdot p - C_{o2} = U_{DC}[o2] \tag{A.45}$$

This proves that if offer $o1$ is accepted by the data collector, offer $o2$ would also be accepted.

The expected cardinality of the dataset after accepting offer $o1$ would be Max and after accepting offer $o2$ it would be $n \cdot \alpha_{o2}$. Based on the condition for case 3 (see Table 5.3), we know $Max \leq n \cdot \alpha_{o2}$. Therefore, accepting offer $o2$ would result in a dataset at least as large as the case where offer $o1$ is accepted.

Case 2a for *o1* vs. case 2a for *o2* - If the data collector accepts offer $o1$ in case 2a then her payoff $\hat{U}_{DC}^{2a}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq n \cdot \alpha_{o1} \cdot p - C_{o1} = U_{DC}[o1] \tag{A.46}$$

Since $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, we have:

$$n \cdot \alpha_{o1} \cdot p - C_{o1} \leq n \cdot \alpha_{o2} \cdot p - C_{o2} = U_{DC}[o2] \tag{A.47}$$

Therefore, the data collector would also accept offer $o2$.

The expected cardinality of the dataset would be $n \cdot \alpha_{o1}$ if offer $o1$ is accepted and $n \cdot \alpha_{o2}$ if offer $o2$ is accepted. Since $\alpha_{o1} < \alpha_{o2}$, the expected cardinality of the dataset after accepting offer $o2$ is at least as large as accepting offer $o1$.

Case 2a for *o1* vs. case 3 for *o2* - If the data collector accepts offer $o1$ in case 2a then her payoff $\hat{U}_{DC}^{2a}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq n \cdot \alpha_{o1} \cdot p - C_{o1} = U_{DC}[o1] \tag{A.48}$$

The condition for case 2 (see Table 5.2) is $n \cdot \alpha_{o1} \leq Max$. This inequality and the facts that $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$, prove the following:

$$n \cdot \alpha_{o1} \cdot p - C_{o1} \leq Max \cdot p - C_{o1} \leq Max \cdot p - C_{o2} = U_{DC}[o2] \tag{A.49}$$

Therefore, the data collector would also accept offer $o2$.

The expected cardinality of the dataset would be $n \cdot \alpha_{o1}$ if offer $o1$ is accepted and $n \cdot \alpha_{o2}$ if offer $o2$ is accepted. Since $\alpha_{o1} < \alpha_{o2}$, the expected cardinality of the dataset after accepting offer $o2$ is at least as large as accepting offer $o1$.

Case 2b for *o1* vs. case 2a for *o2* - If the data collector accepts offer $o1$ in case 2b, her payoff $\hat{U}_{DC}^{2b}$ (see the version in Equation A.24) is greater than or equal to zero. In other words:

$$0 \leq \left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} = U_{DC}[o1] \tag{A.50}$$

The condition for case 2b is $0 \leq \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}$. Therefore, the following inequality holds:

$$\left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \tag{A.51}$$

Case 2a for $o2$ (see Table 5.2) implies that $\gamma \cdot p < \alpha_{o2}$. We also know that $\alpha_{o1} < \alpha_{o2}$ and $C_{o2} \leq C_{o1}$. Therefore we have:

$$p \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot n \cdot \left[\frac{\alpha_{o2}}{2} + \frac{\alpha_{o2}}{2}\right] - C_{o2} = U_{DC}[o2] \tag{A.52}$$

The expected cardinality of the dataset after accepting offer $o1$ would be $n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right]$, and after accepting offer $o2$ would be $n \cdot \alpha_{o2}$. Since $\alpha_{o1} < \alpha_{o2}$ and $\gamma \cdot p < \alpha_{o2}$ (see condition 2a in Table 5.2), we have:

$$n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right] \leq n \cdot \left[\frac{\alpha_{o2} + \gamma \cdot p}{2}\right] \leq n \cdot \left[\frac{\alpha_{o2} + \alpha_{o2}}{2}\right] = n \cdot \alpha_{o2} \tag{A.53}$$

Therefore, accepting offer $o2$ would result in a dataset at least as large as the case where offer $o1$ is accepted.

Case 2b for *o1* vs. case 2b for *o2* - Since $\hat{U}_{DC}^{1b} = \hat{U}_{DC}^{2b}$, the proof of this case is identical to the proof of "case 1b for $o1$ vs. case 1b for $o2$".

Case 2b for *o1* vs. case 2c for *o2* - The proof of this case is identical to the proof of "case 1b for $o1$ and case 1c for $o2$". This is due to the facts that $\hat{U}_{DC}^{1b} = \hat{U}_{DC}^{2b}$, $\hat{U}_{DC}^{1c} = \hat{U}_{DC}^{2c}$, and conditions used to prove "case 1b for $o1$ vs. case 1c for $o2$" also hold in "case 2b for $o1$ vs. case 2c for $o2$".

Case 2b for *o1* vs. case 3 for *o2* - If the data collector accepts offer $o1$ in case 2b, her payoff $\hat{U}_{DC}^{2b}$ is greater than or equal to zero. In other words:

$$0 \leq \left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} = U_{DC}[o1] \tag{A.54}$$

According to Table 5.2, the condition for case 2b is:

$$0 \leq \frac{\gamma \cdot p - \alpha_{o1}}{2 \cdot \gamma} \tag{A.55}$$

As a result we have:

$$\left[p - \frac{\gamma \cdot p - \alpha_{o1}}{2\gamma}\right] \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \tag{A.56}$$

Since $C_{o2} \leq C_{o1}$ and $n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] \leq Max$ (see Equation A.26), the following holds:

$$p \cdot n \cdot \left[\frac{\gamma \cdot p}{2} + \frac{\alpha_{o1}}{2}\right] - C_{o1} \leq p \cdot Max - C_{o2} = U_{DC}[o2] \tag{A.57}$$

This proves that if offer $o1$ is accepted by the data collector, offer $o2$ is also accepted.

The expected cardinality of the dataset after accepting offer $o1$ would be $n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right]$ and after accepting offer $o2$ would be $n \cdot \alpha_{o2}$. Based on Equation A.26 we know $n \cdot \left[\frac{\gamma p}{2} + \frac{\alpha_{o1}}{2}\right] \leq Max$. The condition for case 3 (see Table 5.3) is $Max \leq n \cdot \alpha_{o2}$. These two facts prove the following:

$$n \cdot \left[\frac{\alpha_{o1} + \gamma \cdot p}{2}\right] \leq Max \leq n \cdot \alpha_{o2} \tag{A.58}$$

Therefore, accepting offer $o2$ would result in a dataset at least as large as the case where offer $o1$ is accepted.

Case 2c for *o1* vs. case 2c for *o2* - Since the condition 1c (from Table 5.1) is the same as the condition 2c (from Table 5.2) and $\hat{U}^{1c}_{DC} = \hat{U}^{2c}_{DC}$, the proof of this case is identical to "case 1c for $o1$ vs. case 1c for $o2$".

Case 2c for *o1* vs. case 3 for *o2* - The proof of this case is identical to the proof of "case 1c for $o1$ and case 3 for $o2$". This is due to the facts that $\hat{U}^{1c}_{DC} = \hat{U}^{2c}_{DC}$ and all of the conditions used to prove "case 1c for $o1$ vs. case 3 for $o2$" also hold in "case 2c for $o1$ vs. case 3 for $o2$".

Case 3 for *o1* vs. case 3 for *o2* - If the data collector accepts offer $o1$ in case 3 then her payoff $\hat{U}^{3}_{DC}$ for offer $o1$ is greater than or equal to zero. In other words:

$$0 \leq Max \cdot p - C_{o1} = U_{DC}[o1] \tag{A.59}$$

Since $C_{o2} \leq C_{o1}$, we have:

$$Max \cdot p - C_{o1} \leq Max \cdot p - C_{o2} = U_{DC}[o2] \tag{A.60}$$

Therefore, the data collector would also accept offer $o2$.

The expected cardinality of the dataset would be $n \cdot \alpha_{o1}$ if offer $o1$ is accepted and $n \cdot \alpha_{o2}$ if offer $o2$ is accepted. Since $\alpha_{o1} < \alpha_{o2}$, the expected cardinality of the dataset after accepting offer $o2$ is larger than the case of accepting offer $o1$.

Once parts 1 and 2 are proven for every possible combination of cases that apply to offers $o1$ and $o2$, it is possible to show that offering $o2$ or $o2'$ (an offer with price 0) provides the data user with a payoff at least as large as her payoff for offering $o1$. We prove part 3 by considering three different scenarios as follows:

Scenario 1 - *Both offers get accepted by the data collector:* In part 2 of the proof we showed that if both offers are accepted then cardinality of the data table after accepting offer $o2$ is at least as large as its cardinality after accepting offer $o1$. Since both offers $o1$ and $o2$ have the same values for g_j (granularity level of the attribute A_j) and price, p, the payoff to the data user for offering $o2$ is at least as large as her payoff for offering $o1$.

Scenario 2 - *Offer $o1$ does not get accepted but offer $o2$ gets accepted by the data collector:* In this scenario the data user's payoff for offer $o1$ is zero. The data user can achieve a higher (or the same amount of) profit by choosing to offer $o2$ if such offer provides her with a non-negative payoff or choosing to offer $o2'$ (with a guaranteed payoff of zero), otherwise.

Scenario 3 - *None of the offers get accepted by the data collector:* In this scenario the data user's payoff for offering $o1$ and $o2$ are both zero. □