QUANTIFYING THE COSTS AND BENEFITS OF PRIVACY-PRESERVING HEALTH DATA PUBLISHING

RASHID HUSSAIN KHOKHAR

A THESIS

IN

THE CONCORDIA INSTITUTE FOR INFORMATION SYSTEMS ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF APPLIED SCIENCE IN INFORMATION SYSTEMS
SECURITY
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

APRIL 2013

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Rashid Hussain Khokhar**

Entitled: **Quantifying the Costs and Benefits of Privacy-Preserving Health Data Publishing**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science in Information Systems Security

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

	Dr. Simon Li	Chair
	Dr. Anjali Awasthi	CIISE Examiner
	Dr. Wen-Fang Xie	External Examiner
	Dr. Benjamin Fung	Supervisor
Approved by	**Dr. Amr Youssef**	Graduate Program Director

_______________ 20 ______ ____

Dr. Robin A. L. Drew, Dean

Faculty of Engineering and Computer Science

Abstract

Quantifying the Costs and Benefits of Privacy-Preserving Health Data Publishing

Rashid Hussain Khokhar

Cost-benefit analysis is required for making good business decision. This analysis is crucial in the field of privacy-preserving data publishing. In the economic trade of data privacy and utility, organization has the obligation to respect privacy of individuals. They intend to maximize the utility in order to earn revenue and also aim to achieve the acceptable level of privacy. In this thesis, we study the privacy and utility trade-offs and propose an analytical cost model which can help organization in better decision making subject to sharing customer data with another party. We examine the relevant cost factors associated with earning the revenue and the potential damage cost. Our proposed model is suitable for health information custodians (HICs) who share raw patient electronic health records (EHRs) with another health center or health insurer for research and commercial purposes. Health data in its raw form contain significant volume of sensitive data and sharing this data raises issues of privacy breach. Our analytical cost model could be utilized for non-perturbative and perturbative anonymization techniques for relational data. We show that our approach can achieve optimal value as per selection of each privacy model, namely,

K-*anonymity*, LKC-*privacy*, and ϵ-*differential privacy* and their anonymization algorithm and level, through extensive experiments on a real-life dataset.

Acknowledgments

It is my great pleasure to express my profound gratitude to my supervisor, Dr. Benjamin C. M. Fung. I am deeply indebted to him for all the guidance, constructive criticism, and persistent support through out the period of my study and thesis writing. His kind attitude, patience and motivation is admirable.

I am very much obliged to my friends Mr. Bukhari and Mr. Saleem for their encouragement in completing my thesis.

Also, I am highly grateful to all my family members especially my beloved parents, my wife and my daughter for their invaluable moral support and prays.

“Positive thinking won’t let you do anything but it will let you do everything better than negative thinking will.” - Zig Ziglar

To my wife *Irsa* and my daughter *Muneeba*.

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

EHR Electronic Health Record

HIC Health Information Custodian

HIPAA Health Insurance Portability and Accountability Act

HITECH Health Information Technology for Economic and Clinical Health

HHS United States Department of Health and Human Services

AHP Analytic Hierarchy Process

HIV Human Immunodeficiency Virus

PRAM Post Randomization Method

QID Quasi-Identifier

TDS Top-Down Specialization

DiffGen Differentially-private Generalization

DR Discernibility Ratio

DM Discernibility Metric

BA Baseline Accuracy

CA Classification Accuracy

Chapter 1

Introduction

The development of Electronic Health Record (EHR) systems proliferates in recent years [NCGC09]. Typically, an EHR system provides a stable and secure storage for large volume of health-related data including patient-specific medical history, laboratory test results, demographics, and billing records. The centralized storage not only facilitates the daily operations of different health service providers but also provides an ideal environment for supporting effective health data mining. The goal of health data mining is to efficiently and effectively extract hidden knowledge from large volume of health data with the goal of improving the operations of health service providers or supporting medical research. Data mining on EHRs has been proven to be effective and beneficial to health service providers, researchers, patients, and health insurers [KT05].

To achieve effective health data mining, the precondition is having access to high quality health data. Yet, health data by default is sensitive, and health information custodians (HICs) have the obligation to preserve the privacy of patients [BH03] [AFWM10] [BEP00]. Nowadays, the health information sharing activities are primarily based on obtaining consensus from the patients; however, HICs have faced notable privacy breaches of different nature [KCG11] [Swe02], which are due to either negligence of administrative staff or employment of weak de-identification methods.

In the past 15 years, many new privacy-enhancing techniques have been proposed to thwart different types of privacy attacks [FWFY10]. New privacy models and data anonymization methods have been iteratively proposed, broken, and patched with new models and methods [MCFY11] [KM11]. Thus, it is very difficult to claim that the published data is bulletproof for all privacy attacks. Suppose a health information custodians (HIC) would like to share its patient-specific data to another party for research purpose. The HIC would like to know the answers to the following questions:

- Which privacy model and anonymization algorithm should be employed?
- Given an anonymization algorithm, how to choose the parameters to provide adequate privacy protection to the patients?
- How useful the data is after anonymization?
- What is the probability of a privacy breach on the released data?
- What are the costs in case of a patient privacy breach?

A practical approach is to identify, minimize, and accept the risks by studying the trade-off between privacy protection and information utility. A study on patient privacy and data security [Pon12] shows that the number of health service providers' reporting cases of data privacy breach is increasing every year. The data loss includes patient sensitive information, medical files, billing, and insurance records. Average economic impact of data breaches over the last two years is $\$2.4$ million. These data loss incidents not only create negative impacts of the HICs' images in the general public but also result in possible civil lawsuits from patients for claiming compensation [Wit07] [BL11]. Thus, the measure of economic impact of privacy breach is complex. The objective of this thesis is to model the associated costs and benefits of sharing person-specific information with different data anonymization methods at different privacy protection levels with respect to the information utility for health data mining in terms of monetary value.

1.1 Contributions

The contributions of this thesis are summarized as follows. We study the challenges for sharing patient-specific EHRs faced by health information custodians (HICs), and develop an analytical cost model to search for optimal value. To thwart the potential privacy attacks on the released data, different privacy models, such as LKC-*privacy* [MFHL09], and ϵ-*differential privacy* [Dwo06], have been proposed. Applying these privacy models would result in degrade of data quality and loss of information. The goal of our proposed model is to evaluate the cost of data distortion, the likelihood of a privacy breach, the cost of lawsuit, and the compensation cost, so that the HICs can compare with benefits of releasing the data for health data mining, such as general data analysis and classification analysis. The cost model can help HICs in better decision making on secondary and commercial usages of health data. Finally, we evaluate the proposed model on a real-life person-specific data set.

1.2 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, we provide the literature review. In Chapter 3, we first present some measures to quantify the degree of privacy protection and information utility followed by an overview of a data anonymization algorithm and a problem definition. In Chapter 4, we provide details of our proposed solution to quantify optimal cost. In Chapter 5, we evaluate our proposed method to search for optimal value by performing experiments on real-life person-specific data set. Finally, in Chapter 6, we conclude the thesis and discuss possible future work.

Chapter 2

Literature Review

The research topic of privacy-preserving data publishing has received a lot of attention in different research communities, from economic implications to anonymization algorithms. The state-of-the-arts are summarized as follows.

2.1 Privacy Costs and Benefits in Commercial Setting

Cost and benefit analysis is the key concept in economic decision making system. Different techniques are in practice for problem solving and decision making such as Grid Analysis, Analytic Hierarchy Process, Conjoint Analysis, Decision Trees, and Pareto Analysis. Grid Analysis is also known as Decision Matrix Analysis, in which weights are assigned to different factors in making a decision [Ghu10]. Analytic Hierarchy Process is based on mathematics and psychology, in which weights are assigned to each qualitative and quantitative factors for organizing and analyzing complex decisions [Saa08]. Conjoint Analysis is a statistical technique used in market research to measure buyer preferences [PMJ02]. Decision Tree is basically a decision support tool that uses a tree-like model to analyze possible consequences of a decision, quantify the values of outcomes, and provide guidance in making good decisions [Min13]. Pareto Analysis employs the 80/20 Rule, in which the idea is to focus on the most important problems by doing 20% of work that can generate 80% advantage of doing entire job [Koc11].

Ku and Fan [KF09] use mathematical based Analytic Hierarchy Process (AHP), which is a multi-objective decision making tool to analyze the relative weights of the nine fundamental factors assigned by consumers who intend to purchase travel products from online travel agents. They build pairwise comparison matrix for criteria and subcriteria to obtain each hierarchical factor weight. They found that privacy is one of the most important factor considered by consumers when purchasing travel products on the Internet.

Phillips et al. [PMJ02] use conjoint analysis method to examine preferences for HIV testing. They conduct a survey and define attributes of HIV tests, i.e., location, price, sample collection, timeliness/accuracy, privacy/anonymity, and counseling with their levels. Price is defined as an attribute so that respondents can make decision between price and other attributes. Respondents are asked to choose from "Test A or B" where each test is described using a series of attribute levels. They calculate the mean on the absolute value of coefficients across levels of attributes in order to compare which attribute in general most important to respondents. Privacy/anonymity is found to be the most important attribute for respondents in trade-off among other attributes. To measure the validity they use three approaches (1) consistency of preferences, (2) willingness to trade, and (3) consistency with theoretical predictions. They set baseline scenario for attribute privacy/anonymity as results given in person but not linked to name. By using conjoint analysis they found that respondents did not prefer testing results by phone or in-person with linking of names to results.

Yassine and Shirmohammadi [YS08] discuss the negotiation process between online consumers and sellers by which consumers can capitalize based on the value of their personal information. They examine only the case of a monopoly in their study and employ risk-based premium method [DD05] to determine the consumer's payoff. The quantified privacy risk is context-dependent for each consumer. Similar to other business risks, privacy risk could significantly affect the revenue of a company.

Jentzsch et al. [JPH12] analyze the monetization of privacy and find that many consumers prefer service providers with lower price even though they are more privacy invasive. If the products and prices are similar, then the service provider that collects less personal information get significant share of the market due to offering privacy-friendly online services. They use duopoly model that allows consumers to select service provider depending on their privacy concerns and the offers made by service providers. They recommend that if portability of personal profiles for consumers among service providers is mandated, consumers will face reduced potential switching costs in personalization, but the transfer of personal profiles should be dependent on the consent of the consumer.

Our notion is close to Zielinski and Olivier's work [ZO10] which aims at maximizing both privacy and information utility. They present an approach based on price theory to achieve optimum levels of privacy and utility by the use of constraint optimization. They solve the optimization problem by using the Lagrange multipliers method. In contrast, our work employs a mathematical model by analyzing different cost factors associated with earning the revenue and the potential damage cost. Furthermore, their work is limited to non-perturbative patient-specific data anonymization and is applicable when global recoding is used as the anonymization technique. Our proposed method is applicable to both perturbative and non-perturbative data.

2.2 Privacy Trade-offs in Secondary Use

A family of previous work [LS08] [LL09] [GRS09] [AAC^{+}11] discusses the trade-off between privacy and utility but not in terms of monetary value. Some areas for secondary use of electronic health data include clinical research and development, public health surveillance, health system planning and management, quality improvement, post-market surveillance of drugs, mandatory or discretionary reporting and health insurance.

Loukides and Shao [LS08] present a distance-based quality measure approach that handles both quasi-identifiers and sensitive attributes on equal terms by optimizing the weighted sum of the amount of generalization of quasi-identifiers and the amount of protection of sensitive attributes for K-anonymous data. They design an efficient threshold based clustering algorithm that use heuristics technique and perform greedy search in forming data groups. This approach first partitions the data into sub-spaces and then clusters each sub-space separately. Multi-dimensional local recoding strategy is used to achieve both quality and efficiency in the anonymization process.

Li and Li [LL09] suggest that it is inappropriate to directly compare privacy with utility. They identify three critical characteristics about privacy and utility. The first characteristic states that acquiring specific knowledge about a small group of individuals has a significant impact on privacy, while acquiring aggregate information about a large group of individuals has a significant impact on utility. The second characteristic states that privacy is an individual concept, and utility is an aggregate concept. The third characteristic states that any information learned by the adversary if it deviates from prior belief either correct or incorrect may result in privacy loss, but only correct information contributes to utility. They observe that the trade-off between privacy and utility in publishing data is similar to the risk-return trade-off in financial investment [EG95], where the aim is to determine the appropriate level of risk. They use JS-divergence distance for measuring privacy loss. For utility loss they compare the anonymized data with the original data.

Dwork et al. [DMNS06] discuss differential privacy model which ensures that the addition or removal of a single database record does not significantly affect the overall privacy of a database and it also guarantees protection independent of an adversary's background knowledge. Ghosh et al. [GRS09] follow mechanisms that guarantee near-optimal utility to every potential user, independent of its side information and preferences. They model the side information as a prior probability distribution over the query results, and preferences using a loss function. They show that user can derive as much utility from geometric mechanism as it can be derive from differentially private mechanism.

Alvim et al. [AAC$^+$11] model the database query system as an information-theoretic channel and measure the information that an attacker can learn by posting queries on database and analyzing the response. They prove that ϵ-differential privacy provides protection by implying the bound on the information leakage and utility. This bound is strong enough to prevent attacks using prior distributions. They use binary gain function to measure the utility of a query result.

2.3 Disclosure Control Methods for Privacy Protection

Many non-perturbative and perturbative disclosure control methods, such as global and local recoding [WW98] [Tak99], suppression and local suppression [WW98] [Lit93], sampling [SMOW94], micro-aggregation [DFMS02], noise addition [Kim86], data swapping [DR82], post randomization [KWG97], adopted in the past so far in the vision to provide confidentiality and privacy in publishing person-specific data. These methods according to Gehrke [Geh10] do not formally state how much an attacker can learn and they preserve confidentiality by hiding the parameters used. Below we summarize the works in disclosure control methods.

Global recoding: In global recoding, specific attribute values are mapped into same generalized value in all records. Global recoding is the preferable method when there are many unsafe combinations to eliminate in the person-specific data and to obtain uniform categorization of attributes [WW98]. It applies to categorical variables and continuous variables. For example, recoding the age attribute values into a set of 5-year age groups, or recoding the occupation attribute values into two groups ‘White-collar’ and ‘Blue-collar’. This method helps in de-identification of the records.

Local recoding: In local recoding, attribute values may be generalized to different generalization group [WF10]. For example, attribute value Age=10 appear in two records may map to different anonymization age groups [5-10] and [10-15]. Local recoding provides less information loss in comparison to global recoding, but the analysis result could be difficult to interpret.

Local suppression: Local suppression is the process in which specific value of an attribute in a record changes to ‘missing’ value but the attribute values in other records remain unchanged [WW99]. It is a special case of generalization, which aims to reduce the information either in cell or record to prevent from identification. For example, suppose the combination "Occupation=Manager; Age=24; Salary>50K" is at risk of identification due to its uniqueness. By suppressing the age value 24 reduces the risk of information disclosure and increases the frequency count of similar records.

Sampling: Sampling is suitable for categorical microdata, in which sample S of the original set of records is published instead of publishing the original microdata file. This method is not effective for continuous microdata because it does not mask continuous variable for all records in S [HDFF$^+$12].

Micro-aggregation: In micro-aggregation, records are aggregated into groups and instead of releasing original value for a given record, the average of the group to which the record belongs is released [HDFF$^+$12]. Groups are constructed using a criterion of maximal similarity in order to prevent disclosure of individual data. Univariate micro-aggregation and Multivariate micro-aggregation methods are used to deal with one or more variables at a time. Univariate micro-aggregation is also called individual ranking in which micro-aggregation apply to each variable independently [DFOTMS02]. Multivariate micro-aggregation is NP-hard and in this approach groups are formed by considering all or subset of the variables at the same time [NHT08]. In general multivariate micro-aggregation method offer better disclosure control than univariate micro-aggregation.

Noise addition: Additive noise method is used to perturb continuous data in order to preserve privacy of the data but on other side it may affects the quality of data for legitimate use. It consists of adding a random value which chosen from uniform, normal, exponential, and other distributions with zero mean [Sra10]. Differential privacy mechanism adds noise using exponential distribution to achieve acceptable level of privacy [Dwo08].

Data swapping: Data-swapping is a method used for masking microdata file without altering marginal frequency counts. It consists of altering a fraction of the original records in a file by switching values of a subset of variables between selected pairs or swap pairs of records. It is simple to implement and it protects the univariate distribution of the variable by removing the relationship between the record and the respondent. It is commonly used to protect sample uniqueness to avoid risk of re-identification. Furthermore swapping data values between two or more variables can disturb multivariate relationships which would affect the utility of the data for research analysis [Moo96].

Post randomization: Post randomization method (PRAM) is a randomized version of data swapping to avoid disclosure of data. PRAM uses probabilistic mechanism (also known as transition matrix) to change the score on some categorical variables for certain records with respect to the score in the original microdata file [BM05]. The resulting perturbed microdata file may contain inconsistencies, e.g., 18-year old professor or 12-year old widow. These inconsistencies may appear between different records as well as between different variables of same record. To remove these inconsistencies edits check run use to eliminate invalid combinations. Its use is limited in practice due to little practical knowledge available for information utility.

2.4 Privacy Compliance

The health information custodians (HICs) who are liable to share electronic health records (EHRs) for health data mining and clinical research should ensure the compliance of health regulatory bodies. The Health Insurance Portability and Accountability Act (HIPAA) requires patient consent before the disclosure of health information between health service providers [Dep13]. Health Information Technology for Economic and Clinical Health (HITECH) Act builds on the HIPAA Act of 1996 to strengthen the privacy and security rules. Under the HITECH Act, HIPAA covered entities are required to report data breaches that affect 500 or more individuals to the U.S. Department of Health and Human Services (HHS) and the media, in addition to notifying the affected individuals [Lic12]. The HIC would face substantial breach notification costs and enforcement risks if there is any lapse occurred in non-compliance with the law at other side. HIPAA privacy rule provides two methods by which health information can be designated as de-identified. The first is the *Expert Determination* method which requires that an expert certifies the re-identification risk inherent in the data is sufficiently low. The second is the *Safe Harbor* method which requires the removal of a list of 18 identifiers [Off12].

The final rule under the HITECH Act augments an individual's privacy protections, expands individuals new rights to their health information, and includes revisions to the penalties applied to each HIPAA violation category for healthcare data breaches. Section 160.404 refers to the new HITECH penalty scheme [Dep13], as follows: (1) for violations in which the covered entity *did not know* and, by applying persistent efforts, would not have found within the scope of knowledge that the covered entity violated a provision, an amount not less than $100 or more than $50,000 for each violation (2) for a violation in which it is known that the violation was due to *reasonable cause* and not to *willful neglect*, an amount not less than $1,000 or more than $50,000 for each violation (3) for a violation in which it is known that the violation was due to *willful neglect* and was *timely corrected*, an amount not less than $10,000 or more than $50,000 for each violation and (4) for a violation in which it is known that the violation was due to *willful neglect* and was *not timely corrected*, an amount not less than $50,000 for each violation. A penalty for repeat violations in a calendar year can be hold upto $1.5 million across all HIPAA violation categories of an identical provision.

2.5 Privacy Breach and Impact

In this section we provide the details for personal data privacy breaches and their impacts. The top three causes for a personal data breach are [BRL12]: (1) Accidental disclosure, (2) Loss, and (3) Unauthorized access, use or disclosure.

Accidental disclosure: Incidents where a company mistakenly exposed personal information to unintended recipients. For example, bank confidential letter was sent to the wrong address through human, mechanical, or system error, or at work place inadvertently email personal information of an employee to the wrong recipient, personal data file made publicly accessible on a company's website by means of some technical error [BRL12].

Loss: Incidents where personal information is lost by a company. For example, stolen of laptop, CD/DVD, tape drive, hard drive, usb drive, flash drive, or any other removable media, or paper documents [BRL12].

Unauthorized access, use or disclosure: Incidents where personal information is illegally accessed from company's database by an intruder in order to acquire sensitive information of individual. Also in the cases where company's employees access or disclose personal information outside the requirements or authorisation of their assigned job [BRL12].

A study on patient privacy and data security [Pon12] shows that the number of health service providers' reporting cases of data breach has increased significantly during the last two years. The average economic impact of data breaches is measured at $2.4 million. To ensure the compliance, organizations are required to implement internal safe guard mechanisms to minimize the risks, costs, and the impact of a data breach. These safe guards may have a significant cost to mitigate the effects of a data breach including costs involved in sending mandatory breach notifications, dealing with regulatory investigations, hiring external auditors, facing class action litigation and loss of business goodwill due to decreased consumer loyalty [BRL12]. In addition to applying the safeguards, organizations are required to prepare incident response plan in mitigating the effects of a data breach. These response measures may serve as a reference guide for certain best practices in dealing with a data breach. These include gathering necessary information related to a data breach, finding number of affected individuals, and investigating possible consequences of the data breach. Subsequently, companies are also required to prepare a report and send notification of the data breach to concern data protection authorities on the set format as per applicable laws.

Electronic Health Record (EHR) provides benefits by connecting different health service providers through health data networks. Though integration on common platform enable health service providers to do effective health data mining and clinical research, on the other hand it is very challenging for them to manage, maintain, and control patient information. Health records by its nature are very sensitive and sharing even de-identified records may raise issues of patient privacy breach. Data privacy breach incidents not only create negative impacts of these health service providers in the general public but also result in possible civil lawsuits from patients for claiming compensation. Each person has its own intrinsic value and it is hard to settle the compensation cost for every individual. The economic analysis of litigation suggests that individuals are more likely to submit a file suit when their expected rewards exceed their expected costs [CU08]. In civil cases an individual can sue against another party who could be an individual, a company or corporation. Additionally, the parties may be two companies, organizations, or corporations [EH13].

Below we mention some cases of monetary fines due to data breach incidents.

Monetary fines due to data breach:

Zappos, an online shoe retailer owned by Amazon, suffered a massive data breach affecting 24 million consumers. In this data breach incident, reported in January 2012, hackers gained access to the company's internal network and stole details of their consumers including their name, e-mail addresses, billing and shipping addresses, the last four digits of the user's credit card and a cryptographically scrambled version of their website password [Sch12]. Such incidents can significantly impact company's reputation and loss of consumer trust. As stated by Zappos CEO Tony Hsieh following this incident, "We have spent over 12 years building our reputation and trust, it is painful to see us take so many steps back due to a single incident."

Global Payments initially from investigation estimated that 1.5 million accounts were exposed but later news reports suggested that nearly 7 million accounts were exposed due to data breach [Inf13]. According to Global Payments, the data breach it revealed in April 2012 cost the company around $94 million. The breakdown of breach costs include $60 million for expenses in investigation, remediation, and protection insurance, and $35.9 million for the estimated fraud loses, fines, and other potential penalties that would be imposed by the card networks.

South Shore Hospital pays $750,000 to settle charges against violation of HIPAA privacy and security rule that exposed the confidential health information of more than 800,000 individuals in 2010 [RB12]. Attorney General filed the lawsuit against the hospital under the state Consumer Protection Act and the federal HIPAA Act for this data breach incident. Hospital pays a civil penalty of $250,000, payment of $225,000 for an education fund to be used by the Attorney General's Office, and $275,000 for the hospital to implement the security measures.

Western Health, which manages hospitals and clinics at Newfoundland region, fired an employee who was involved in accessing the confidential medical records of 1,043 patients [CBC12]. Barbara Hynes filed a class-action lawsuit against the concerned authority for breach of her personal and confidential health information.

Hospice of North Idaho, a non-profit patient care facility, agreed to pay $50,000 in settlement with the U.S. Department of HHS for violating the HIPAA Act [DeG13]. In February 2010, they reported to HHS that an unencrypted laptop containing sensitive personal information of 441 patients has been stolen. This is the first breach settlement affecting less than 500 patients. Leon Rodriguez is the Director of the Office for Civil Right said "This action sends a strong message to the health care industry that, regardless of size, covered entities must take action and will be held accountable for safeguarding their patients' health information."

In many data breach lawsuits, plaintiffs seek remedy for claimed harms, such as actual financial loss incurred from the identity theft, emotional distress, compensation for the credit monitoring, or possible future losses [RHA12]. A study by Ponemon Institute found that identity theft 61% is the most significant privacy concern followed by 56% increase in government surveillance. [Pon13]

Chapter 3

Preliminaries

In this chapter, we first present some measures to quantify the degree of privacy protection and information utility followed by an overview of a data anonymization algorithm and a problem definition.

3.1 Quantifying Privacy

A HIC wants to share a person-specific data table with a health data miner, such as a medical practitioner or a health insurance company for research purposes. A person-specific data set for classification analysis typically contains four types of attributes, namely the explicit identifiers, the quasi-identifier (QID), the sensitive attribute, and the class attribute. Explicit identifiers (such as name, social security number, and telephone number, etc.) are those which belongs to personal unique identification. QID (such as birth date, sex, race, and postal code, etc.) is a set of attributes having values may not be unique but their combination may reveal the identity of an individual. Sensitive attributes (such as disease, salary, marital-status, etc.) are those attributes that contain sensitive information of an individual. Class attributes are the attributes that the health data miner wants to perform classification analysis. Let $D(A_1, \ldots, A_n, Sens, Class)$ be a data table with explicit identifiers removed, where $\{A_1, \ldots, A_n\}$ are quasi-identifiers that can be either categorical or numerical attributes, $Sens$ is a sensitive attribute, and $Class$ is a class attribute. A record in D has the form $\langle v_1, v_2, \ldots, v_n, s, cls\rangle$, where v_i is a value of A_i, s is a sensitive value of $Sens$, and cls is a class value of $Class$.

3.1.1 Privacy Threats

The following example illustrates two types of privacy attacks, namely *record linkage* and *attribute linkage* [FWCY10].

***Example* 1.** Consider the de-identified raw patient data in Table 1, where each record corresponds to the personal and health information of a patient, where QID = $\{Age$, $Gender$, $Occupation\}$, $Sens$ = $\{Disease\}$, and $Class$ = $\{$*Blood transfusion*$\}$. The HIC wants to release Table 1 to a researcher for the purpose of classification analysis on the class attribute *Blood transfusion* which has two values Yes and No, indicating whether or not the patient need transfusion of blood. Without loss of generality, we assume that the only sensitive value in $Disease$ is HIV in this example.

Table 1: De-Identified Raw Patient Data

	Quasi-identifier (QID)			*Sensitive*	*Class*
Rec#	**Age**	**Gender**	**Occupation**	**Disease**	**Blood transfusion**
1	29	M	Doctor	Migraine	N
2	38	F	Cleaner	HIV	Y
3	64	M	Welder	Asthma	Y
4	38	F	Painter	HIV	Y
5	56	M	Painter	Migraine	N
6	24	F	Lawyer	Migraine	Y
7	36	F	Cleaner	HIV	Y
8	61	M	Lawyer	Asthma	Y
9	39	F	Painter	HIV	Y
10	24	M	Technician	Asthma	N
11	52	M	Painter	HIV	Y
12	41	F	Lawyer	Asthma	N
13	28	M	Lawyer	Migraine	Y
14	37	M	Cleaner	HIV	Y
15	66	M	Welder	Asthma	N
16	36	F	Painter	HIV	Y
17	44	M	Painter	HIV	Y

The first type of attack is called *record linkage* [FWCY10]. In this attack, an adversary attempts to link a real-life patient to a data record in the released data table. In other words, the adversary wants to the identify the record of a target victim from the table. Suppose an adversary has gathered some prior knowledge about the target victim who is a female painter, denoted by $qid = \langle F, Painter\rangle$. By matching qid with the records in the table, the adversary attempts to identify the records in the data table that are consistent with the prior knowledge qid. The group of consistent records of a qid is denoted by $D[qid]$. If the group size $|D[qid]|$ is small, then the adversary may identify the victim's record and the victim's sensitive value. The probability of a successful record linkage is $1/|D[qid]|$. In this particular example, $D[qid] = \{Rec\#4, 9, 16\}$.

The second type of attack is called *attribute linkage* [FWCY10]. In this attack, an adversary may not able to identify the exact record of a target victim, but could infer his/her sensitive values with high confidence from the released data table. Suppose an adversary has prior knowledge qid of a target victim. The adversary can first identify $D[qid]$ and infer that the victim has sensitive value s with confidence $P(s|qid) = \frac{|D[qid \wedge s]|}{|D[qid]|}$, where $D[qid \wedge s]$ denotes the set the records containing both qid and s. $P(s|qid)$ is the percentage of the records in $D[qid]$ containing s. The privacy of the target victim is at risk if $P(s|qid)$ is high. For example, given $qid = \langle M, Painter \rangle$ in Table 1, $D[qid \wedge HIV] = \{Rec\#11, 17\}$ and $D[qid] = \{Rec\#5, 11, 17\}$, therefore $P(HIV|qid) = 2/3 = 66.67\%$. ■

3.1.2 Privacy Models

Various privacy models have been proposed to protect against the aforementioned linkages to an individual patient in the released data. In this subsection, we discuss the most widely adopted models in the literature, namely *K-anonymity*, *LKC-privacy*, and *ϵ-differential privacy*.

Definition 1. *(K-*anonymity*)* [SS98]. Let $D(A_1, \ldots, A_n)$ be a table and QID be the quasi-identifier associated with it. D satisfies K-anonymity if and only if each record on QID in D appears with at least $K - 1$ other records in D. ■

K-anonymity does not provide privacy if sensitive values in an equivalence class lack diversity so it is subject to attribute linkage attack. Furthermore, due to the curse of high dimensionality as discussed in [Agg05], enforcing K-anonymity on high-dimensional data would result in significant information loss. To overcome this bottleneck, Mohammed et al. [MFHL09] pointed out that in real-life privacy attack it is very difficult for an adversary to acquire all QID attributes of a target victim, and proposed the *LKC-privacy* model in which the adversary's prior knowledge qid is assumed to be bounded by at most L values of the QID attributes.

Definition 2. *(LKC-*privacy*)* [MFHL09]. Let L be the maximum number of values in the prior knowledge of an adversary on a target victim. Let $S \subseteq Sens$ be a set of sensitive values. A data table D satisfies *LKC-privacy* if and only if for any qid with $|qid| \leq L$,

1. $|D[qid]| \geq K$, where $K > 0$ is an integer anonymity threshold, and

2. for any $s \in S$, $P(s|qid) \leq C$, where $0 < C \leq 1$ is a real number confidence threshold. ■

Intuitively, LKC*-privacy* model prevents both record and attribute linkage attacks and also ensures that every combination of values in $QID_i \subseteq QID$ with maximum length L in the data table D is shared by at least K records, and the confidence of inferring any sensitive values in S is not greater than C, where L, K, C are thresholds and S is a set of sensitive values specified by the HIC. LKC-privacy bounds the probability of a successful record linkage to be $\leq 1/K$ and the probability of a successful attribute linkage to be $\leq C$, provided that the adversary's prior knowledge qid does not exceed L values. LKC is more flexible than K-anonymity in adjusting the privacy and utility trade-off. Dwork [Dwo06] proposed a privacy model called *differential privacy* which provides strong privacy guarantees independent of an adversary's background knowledge and computational power.

Definition 3. *(*ϵ-differential privacy*)* [Dwo06]. Given a sanitization mechanism M_r provides ϵ*-differential privacy* in non-interactive setting, if a real value $\epsilon > 0$, and for any two data sets D_1 and D_2 their symmetric difference contains at most one record (i.e., $|D_1 \triangle D_2| \leq 1$), and for any possible anonymized data sets D^*.

$$\text{Pr}[M_r(D_1) = D^*] \leq e^{\epsilon} \times \text{Pr}[M_r(D_2) = D^*], \tag{1}$$

where the probabilities are taken over the randomness of M_r. ■

Differential privacy is a privacy model which originates from the field of statistical disclosure control. It ensures that the addition or removal of a single database record does not affect the outcome of any query significantly. It follows that no risk is incurred by joining a statistical database.

3.2 Quantifying Utility

The data utility of a released data table depends on the data analysis task to be performed by the data recipient. We examine two utility measures in this thesis. The first measure, discernibility ratio (DR), aims at quantifying the impact of anonymization on the overall data distortion for supporting general analysis.

$$DR = \frac{\sum_{qid} |D[qid]|^2}{|D|^2} \tag{2}$$

DR is the normalized discernibility cost with the range of $0 \leq DR \leq 1$. Lower value of DR represents higher data quality.

The second measure aims at quantifying the impact of anonymization on classification quality. To determine the impact on classification data analysis, we build a classifier on 2/3 of the anonymized records as the training set, and measure the *classification error* (CE) on 1/3 of the anonymized records as the testing set. We measure classification error (CE) using $C4.5$ classifier [Qui93] for classification analysis. *Baseline Error* (BE) is measured on the raw data without anonymization. $CE - BE$ represents the impact of anonymization on classification quality.

3.3 Data Anonymization Algorithm

3.3.1 Top-Down Specialization Algorithm

Algorithm 1 provides an overview of the *Top-Down Specialization* (*TDS*) algorithm [FWY07]. Initially, all values in QID are generalized to the top most value in their taxonomy tree as depicted in Figure 1.

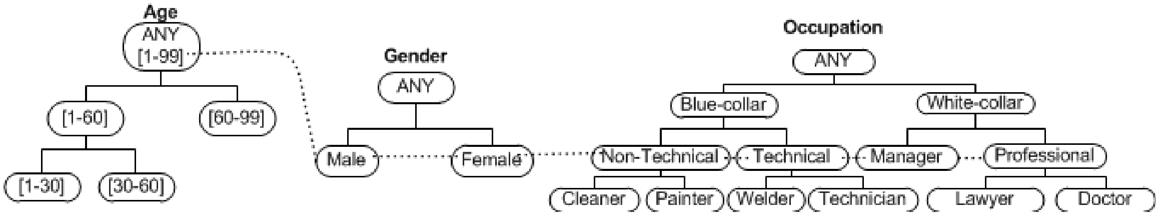

Figure 1: Taxonomy Trees

It is assumed that a taxonomy tree is specified for each categorical attribute in QID. For each continuous attribute in QID, a taxonomy tree can dynamically grown at runtime, forming a binary tree structure in which each non-leaf node has exactly two child nodes that represent a split of the parent interval. $Mark_i$ contains the topmost value for each attribute A_i in a taxonomy tree. At each iteration, the TDS algorithm performs the $Best$ specialization, which has the highest $Score$ among the *candidates* that are valid specializations in $\cup Mark_i$ (Line 4). Then, apply $Best$ to D and update $\cup Mark_i$ (Line 5). Finally, update the $Score$ and validity of the candidates in $\cup Mark_i$ (Line 6). The algorithm is efficient in updating the $Score$ and maintaining the statistics for candidates in $\cup Mark_i$ by directly accessing the data records and it terminates if any further specialization would lead to a violation of the privacy requirement. The specialization process can be viewed as pushing the "mark" of each taxonomy tree downwards, which effect in increase the utility and decrease anonymity as value of the records are become more distinguishable. Figure 1 exhibits a solution mark indicated by the dotted lines representing the anonymous Table 2.

Algorithm 1 Top-Down Specialization (TDS) Algorithm [FWY07]

```
Initialize every value in $D$ to the topmost value.
Initialize $Mark_i$ to include the topmost value.
while some $x \in \cup Mark_i$ is valid do
    Find the $Best$ specialization from $\cup Mark_i$.
    Perform $Best$ on $D$ and update $\cup Mark_i$.
    Update $Score(x)$ and validity for $x \in \cup Mark_i$.
end while;
Output $D$ and $\cup Mark_i$.
```

Table 2: Anonymous data ($L = 2$, $K = 2$, $C = 0.5$)

	Quasi-identifier (QID)			*Sensitive*	*Class*
Rec#	**Age**	**Gender**	**Occupation**	**Disease**	**Blood transfusion**
1	$[1-99]$	M	Professional	Migraine	N
2	$[1-99]$	F	Non-Technical	HIV	Y
3	$[1-99]$	M	Technical	Asthma	Y
4	$[1-99]$	F	Non-Technical	HIV	Y
5	$[1-99]$	M	Non-Technical	Migraine	N
6	$[1-99]$	F	Professional	Migraine	Y
7	$[1-99]$	F	Non-Technical	HIV	Y
8	$[1-99]$	M	Professional	Asthma	Y
9	$[1-99]$	F	Non-Technical	HIV	Y
10	$[1-99]$	M	Technical	Asthma	N
11	$[1-99]$	M	Non-Technical	HIV	Y
12	$[1-99]$	F	Professional	Asthma	N
13	$[1-99]$	M	Professional	Migraine	Y
14	$[1-99]$	M	Non-Technical	HIV	Y
15	$[1-99]$	M	Technical	Asthma	N
16	$[1-99]$	F	Non-Technical	HIV	Y
17	$[1-99]$	M	Non-Technical	HIV	Y

We provide the details of the $Score$ function for general and classification data analysis as follows.

Score for General Analysis: In some cases, data is shared for general data analysis without a specific data analysis task, or the data analysis task is unknown at the time of data release. For these cases, we use discernibility metric [Slo92] to measure the data distortion in the anonymized data table. The discernibility metric charges a cost to each record for being identical from other records. For each record in an equivalence group qid, the penalty cost is $|D[qid]|$. Thus, the penalty cost on a group is $|D[qid]|^2$. To minimize the discernibility penalty cost, we choose the specialization $v \rightarrow child(v)$ that maximizes the value of over all qid containing v, denoted by qid_v.

$$Score(v) = DM(v) = \sum_{qid_v} |D[qid_v]|^2 \tag{3}$$

Score for Classification Analysis: In the case of classification analysis, we use information gain, denoted by $InfoGain(v)$, to measure the *goodness* of a specialization. Our selection criterion, $Score(v)$, is to keep the specialization $v \rightarrow child(v)$ that has the maximum $InfoGain(v)$:

$$Score(v) = InfoGain(v). \tag{4}$$

InfoGain(v): Let D_x denote the set of records in D generalized to the value x. Let $freq(D_x, cls)$ denote the number of records in D_x having the class value cls. Note that $|D_v| = \sum_c |D_c|$, where $c \in child(v)$. So, we have

$$InfoGain(v) = H(D_v) - \sum_c \frac{|D_c|}{|D_v|} H(D_c), \tag{5}$$

$$H(D_x) = -\sum_{cls} \frac{freq(D_x, cls)}{|D_x|} \times log_2 \frac{freq(D_x, cls)}{|D_x|}, \tag{6}$$

where $H(D_x)$ measures the *entropy* of classes for the records in D_x [Qui93], and $InfoGain(v)$ measures the reduction of the entropy by specializing v into $c \in child(v)$. The smaller the entropy $H(D_x)$ implies the higher purity of the partition with respect to the class values. Example 2 shows the computation of $InfoGain(v)$.

***Example* 2.** Consider Table 1 with $L = 2$, $K = 2$, $C = 50\%$, and $QID = \{Age, Gender, Occupation\}$. Initially, all data records are generalized to $\langle$*[1-99]*$, ANY\_Gender, ANY\_Occupation\rangle$, and $\cup Mark_i = \{$*[1-99]*$, ANY\_Gender, ANY\_Occupation\}$. To find the $Best$ specialization among the candidates in $\cup Mark_i$, we compute $Score($*[1-99]*$)$, $Score(ANY\_Gender)$, and $Score(ANY\_Occupation)$. Below we show the computation of $Score(ANY\_Occupation)$ and DR.

For the specialization:

$$ANY\_Occupation \rightarrow \{\textit{Blue-collar}, \textit{White-collar}\}.$$

For general analysis:

$Score(ANY\_Occupation) = 12^2 + 5^2 = 169.$

$DR = \frac{3^2+2^2+5^2+3^2+4^2}{17^2} = 0.217993.$

For classification analysis:

$H(D_{\textit{ANY\_Occupation}}) = -\frac{12}{17} \times log_2\frac{12}{17} - \frac{5}{17} \times log_2\frac{5}{17} = 0.8739$

$H(D_{\textit{Blue-collar}}) = -\frac{9}{12} \times log_2\frac{9}{12} - \frac{3}{12} \times log_2\frac{3}{12} = 0.8112$

$H(D_{\textit{White-collar}}) = -\frac{3}{5} \times log_2\frac{3}{5} - \frac{2}{5} \times log_2\frac{2}{5} = 0.9709$

$InfoGain(ANY\_Occupation) = H(D_{\textit{ANY\_Occupation}}) - (\frac{12}{17} \times H(D_{\textit{Blue-collar}}) + \frac{5}{17} \times H(D_{\textit{White-collar}})) = 0.0156$

$Score(ANY\_Occupation) = InfoGain(ANY\_Occupation) = 0.0156.$ ■

3.3.2 Differential Generalization Algorithm

We employ differentially-private generalization anonymization algorithm *(DiffGen)* [MCFY11] for classification analysis. *DiffGen* achieves ϵ-differential privacy by making two major extensions on TDS. First, *DiffGen* selects the $Best$ specialization based on exponential mechanism. Second, *DiffGen* adds the Laplacian noise to the generalized contingency table, i.e., the published qid counts. The noise is calibrated according to *sensitivity* of the function which defines as the maximum difference of its outputs from two data sets that differ in at most one record. ϵ is a user-specified privacy threshold. A lower value of ϵ implies higher level of privacy protection.

3.4 Problem Definition

This thesis aims at answering the questions pointed out in Chapter 1 by proposing an analytical cost model. The research problem is formally described as follows. Let D be a raw patient-specific data table. A HIC would like to anonymize D and share the anonymized version D' to a third party. The HIC wants to quantify the cost and benefit of D' with respect to the level of privacy protection and information utility with respect to a data analysis or data mining task. Thus, the research problem is to propose an analytical cost model that covers both aspects of data privacy and utility in terms of monetary value. The model provides guidance in finding the optimal solution based on the choice of privacy models, anonymization algorithm, and privacy protection levels. The value of optimal cost continuously changes with respect to the variations and uncertainties in different qualitative and quantitative variables that influence the outcome of the decision on the basis of their values. The model considers the sensitivity of the dataset, size of the dataset, cost of distortion, cost in terms of classification quality, likelihood of privacy breach, cost of lawsuit, trend in compensation cost, probability of attack and potential damage cost.

Chapter 4

Proposed Solution

In this chapter, we propose a solution to quantify the trade-off between privacy and utility in terms of monetary value of data anonymization for health data mining. Our analytical cost model is applicable to both perturbative and non-perturbative anonymization techniques. In the subsequent analysis, we focus the analysis on person-specific relational data but the model is also applicable to other types of data such as set-valued data and sequential data. We assume no randomization is allowed in order to maintain the data truthfulness of records. Our proposed model will be evaluated with respect to some common privacy models, namely K*-anonymity*, LKC*-privacy*, and ϵ*-differential privacy*. Section 4.1 presents the analytical cost model, and Section 4.2 discusses the relevant cost factors of revenue earning and the factors that affect potential damage cost, followed by the attack model and performance measures.

4.1 Analytical Cost Model

Figure 2 depicts an overview of the proposed cost and benefit model. Nodes represent different types of variables such as general variables, chance variables, and objective variable. An influence diagram shows the dependency of one variable on another with an arrow. For example, the arrow connecting the *Size of Dataset* to *Monetary Value of Raw Dataset* indicates that the dependency of *Monetary Value of Raw Dataset* on the *Size of Dataset*. Our model allows the user to choose privacy models and their anonymization algorithms and privacy parameters. Then our model analyzes the impact of privacy protection with respect to the information utility for health data mining in terms of monetary value. It helps in identifying the economic consequences of sharing patient health data. Revenue depends upon the *Monetary Value of Raw Dataset* and the *Cost of Anonymization*. Data anonymization may impact on revenue by hiding potentially relevant information, but on the other hand it may provide benefits in reducing the risk of privacy breach and costs of compensation. The variable *Cost of Anonymization* in the model represents CoD for general data analysis and CQC for classification analysis. *Optimal Cost* is the model's objective and evaluates the overall value or desirability of possible outcomes. This model can help HICs in making better decisions to quantify the value of their earn, impact of a privacy breach, possible costs of compensation when person-specific health data is shared for secondary and commercial purposes.

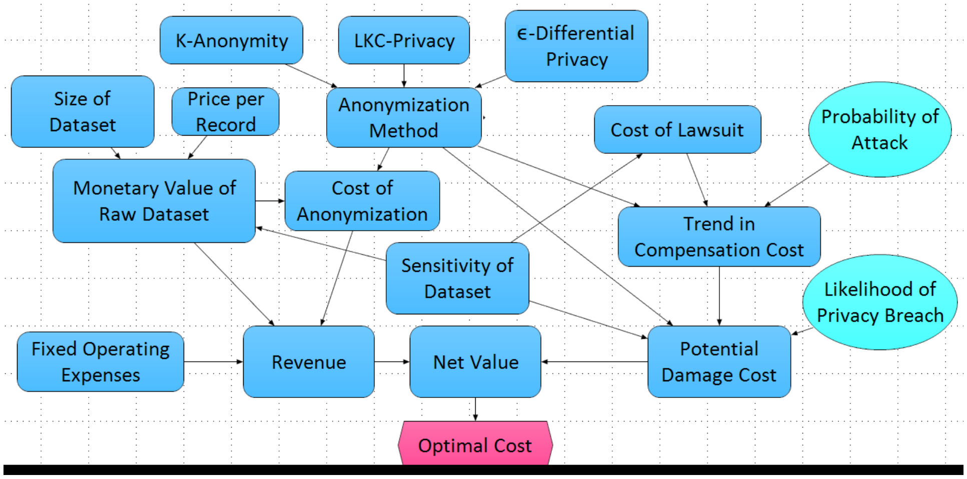

Figure 2: Analytical Cost Model

4.2 Cost Factors

In order to build analytical cost model as depicted in Figure 2, we need to identify and study the relevant quantitative and qualitative cost factors. We perform quantitative analysis to find the variable cost in terms of monetary value. Qualitative analysis is used to measure the relative magnitudes of cost factors.

4.2.1 Sensitivity of Dataset

The sensitivity of a dataset SD is a given qualitative factor and its level $l(l = 1, \ldots, z)$ represents the importance of data privacy. The higher the sensitivity level of l implies the higher monetary value of a raw dataset, and the higher would be the impact on the potential damage cost. Data privacy risk increases as the level of data sensitivity increases.

4.2.2 Size of Dataset

The size of a dataset $Size_{ds}$ is a quantitative factor and it represents the total number of records X in the dataset. $Size_{ds}$ increases as the number of records in the dataset increases. Each record has a monetary value. As the number of records increases, cost of raw dataset also increases.

4.2.3 Price per Record

The price per record Pr_{rec} is a quantitative factor and it represents the unit price Y of record. The cost of a raw dataset increases as the unit price per record increases.

4.2.4 Monetary Value of Raw Dataset

The monetary value of a raw dataset $Cost_{rd}$ is the product factor of sensitivity of the dataset SD, size of the dataset $Size_{ds}$, and price per record Pr_{rec}.

$$Cost_{rd} = SD \times Size_{ds} \times Pr_{rec} \tag{7}$$

4.2.5 Cost of Distortion

To determine the cost of distortion CoD, we first need to determine the discernibility ratio (DR) on the anonymized data as described in Section 3.2. Recall that the discernibility cost charges a penalty to each record for being indistinguishable from other records. Using Equation 2, the cost of distortion is calculated as follows:

$$CoD = Cost_{rd} \times DR \tag{8}$$

4.2.6 Cost in terms of Classification Quality

To determine the cost in terms of classification quality CQC, we first use all records for anonymization, build a classifier on 2/3 of the anonymized records as the training set, and measure the *classification error* (CE) on 1/3 of the anonymized records as the testing set. *Classification Accuracy* (CA) is measured using the form $(1 - CE)$. We use the well-known $C4.5$ classifier [Qui93] for classification model. *Baseline Accuracy* (BA) is the accuracy measured on the raw data without anonymization. $BA - CA$ represents the cost of anonymization in terms of classification accuracy. So, the cost in terms of classification quality is defined as:

$$CQC = Cost_{rd} \times (BA - CA) \tag{9}$$

4.2.7 Revenue

Revenue is the monetary value received by an institution after selling the anonymized version of the dataset for research or commercial purpose. $Cost_{rd}$ is impacted by CoD and

CQC for general analysis and classification data analysis, respectively. Operating expenses have an impact on a company's revenue but here we consider only fixed operating expenses F_{oe}, which is considered to be the same for both cases.

Revenue for the general analysis case, denoted by Rev_{ga}, is defined as:

$$Rev_{ga} = Cost_{rd} - CoD - F_{oe} \tag{10}$$

Revenue for the classification analysis case, denoted by Rev_{ca}, is defined as:

$$Rev_{ca} = Cost_{rd} - CQC - F_{oe} \tag{11}$$

4.2.8 Likelihood of Privacy Breach

Likelihood of privacy breach L_{PB} is calculated by applying adversary background knowledge to infer the sensitive attribute value of a victim in percentage using the attack model. The details are given in Section 4.3. Let us assume that victim record is in the released dataset and adversary knows the victim's QID. Formally, L_{PB} for general and classification analysis case is defined as:

$$L_{PB} = \frac{\textit{\#. of Records for} Sen_{val}}{\textit{\#. of Records on class label} Sen_{attr}} \tag{12}$$

where Sen_{val} denotes the value of the sensitive attribute and Sen_{attr} denotes the sensitive attribute in the dataset.

4.2.9 Cost of Lawsuit

The cost of lawsuit $Cost_{lwst}$ is based on the monetary fines or penalties applicable by law in case of privacy breach. It is a qualitative factor because its monetary value may vary depending on the disclosure of sensitive information. Approximate value of $Cost_{lwst}$

could be considered subject to the historical trends of privacy breach. The cost of lawsuit increases as the level of data sensitivity l increases.

4.2.10 Probability of Attack

The probability of attack $Prob_{atk}$ is taken by calculating the F-measure on sensitive attribute value Sen_{val}. F-measure is a weighted harmonic mean of recall and precision. Formally, $Prob_{atk}$ for general and classification analysis is defined as:

$$Prob_{atk} = \frac{2 \times (\textit{Precision on } Sen_{val} \times \textit{Recall on } Sen_{val})}{\textit{Precision on } Sen_{val} + \textit{Recall on } Sen_{val}} \tag{13}$$

4.2.11 Trend in Compensation Cost

The trend in compensation cost TC_{cost} means how the compensation cost would vary in presence of an attack and its severity level. TC_{cost} is impacted by the choice of the privacy model (e.g., K-anonymity and LKC-privacy) and its level of privacy protection. So, the higher the value of privacy parameter implies less chance of privacy attack. We hypothesize that the privacy attacks would have an exponential impact on compensation cost due to costly litigation processes [Off09]. There is no specific monetary value for compensation cost highlighted in [Off09], but a person who suffers from the financial loss due to disclosure of his sensitive information may claim for compensation. Every personal record has its own intrinsic value and it is hard to settle the compensation cost for each individual. As the probability of attack $Prob_{atk}$ increases, TC_{cost} also increases. Formally, TC_{cost} for general and classification analysis case is defined as:

$$TC_{cost} = \exp(Prob_{atk}) \times Cost_{lwst} \tag{14}$$

4.2.12 Potential Damage Cost

Potential damage cost PDC means the costs in mitigating the effects of a data breach. It may include significant costs in sending mandatory breach notifications, dealing with regulatory investigations, hiring external auditors, facing class action litigation and loss of goodwill in the general public due to decreased patient loyalty [BRL12]. PDC to the HICs is impacted by the likelihood of privacy breach L_{PB} and the trend in compensation cost TC_{cost}. We hypothesize that L_{PB} would have an exponential impact on potential damage cost [BL11] [AFT06] because a plaintiff seek remedy for alleged harms, such as actual financial loss incurred from the identity theft, emotional distress, or possible future losses [RHA12]. Formally, PDC for general and classification analysis is defined as:

$$PDC = \exp(L_{PB}) \times TC_{cost} \tag{15}$$

4.2.13 Net Value

Net value NV shows due diligence in evaluating the cost factors. NV is used in cost-benefit analysis to quantify the difference between the monetary value of revenue and potential damage cost on different privacy protection levels. Formally, NV_{ga} for general analysis and NV_{ca} for classification analysis are calculated as follows respectively:

$$NV_{ga} = Rev_{ga} - PDC \tag{16}$$

$$NV_{ca} = Rev_{ca} - PDC \tag{17}$$

4.2.14 Optimal Cost

Optimal Cost Opt_{cost} is calculated by taking the maximum of net value NV for general analysis or classification analysis on different privacy protection level. Opt_{cost} is formally defined as:

$$Opt_{cost} = \max(NV) \tag{18}$$

4.3 Attack Model

Let D be the raw patient data as shown in Table 1, and D' be the anonymized version of patient data as shown in Table 2. Suppose $Disease$ is the sensitive attribute and $Blood\,transfusion$ is the class attribute. Assume the anonymized data table D' is released together with the classifier. The adversary may have some additional background knowledge about a victim. Suppose he knows that victim is in the table and the victim's qid. Our attack model is similar to [Kif09] in the nous that we are thinking from adversary's perspective and predicting sensitive attribute value of a target victim who is a participant in the anonymized training data. An adversary cannot link a record to an individual, although he can infer some sensitive values with high confidence in percentage. We set the sensitive attribute $Disease$ as the class label and then use classification algorithm $C4.5$ to infer the sensitive attribute of individuals. In our attack model we use precision and recall measures to evaluate the quality of results on the class label $Disease$, which has three values $Migraine$, HIV, and $Asthma$. Below we provide the details of these measures followed by an example of confusion matrix.

4.3.1 Precision

Precision is a measure of exactness or quality which is formally defined as the number of correctly classified positive elements divided by the total number of all classified elements as positive.

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

4.3.2 Recall

Recall is a measure of completeness or quantity which is formally defined as the number of correctly classified positive elements divided by the total number of actual positive elements.

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

4.3.3 F-measure

F-measure is the harmonic mean of precision and recall and it is formally defined as:

$$\textit{F-measure} = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{21}$$

4.3.4 Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a well-known classification model. Performance of classification model is evaluated using the data in the matrix.

***Example* 3.** Consider the anonymous table D' in Table 2. An adversary sets the sensitive attribute $Disease$ as a class on the anonymized version of received data table. It results in new data table which is denoted by D^*. We consider HIV as a sensitive value in the sensitive attribute $Disease$, and then use classification model $C4.5$ to infer sensitive attributes of individuals. The confusion matrix for the three class classifier is shown in Table 3.

		Predicted class		
		A	**B**	**C**
Actual class	HIV (**A**)	3	0	0
	Asthma (**B**)	0	1	0
	Migraine (**C**)	0	1	0

Table 3: Confusion Matrix

The rows correspond to the Actual class of the data, i.e., the class labels in the data. The columns correspond to the Predictions made by the model. The values which are shown along the diagonal represent the number of correctly classified instances; other values show the errors. Below we show the calculation of the performance measures for the above confusion matrix on sensitive value.

True positives refer to the positive records that were correctly classified by the classifier, e.g., $TP = 3$ shows correctly classified. False negatives are the positive records that were incorrectly classified, e.g., $FN = 0$ shows no incorrectly classified. False positives are the negative records that were incorrectly classified, e.g., $FP = 0$ shows no incorrectly classified. So, the values of performance measures, i.e., $Precision = 1$, $Recall = 1$, and $F\text{-}measure = 1$ are calculated by using Equation 19, Equation 20, and Equation 21, respectively. An adversary may use these performance measures to determine the success rate of a privacy attack. F-measure represents the probability of attack $Prob_{atk}$, so when its value equal to 1 it implies that there are 100 percent chances of a successful attack.

4.3.5 Background Knowledge Attack

In continuation of the attack model as discussed in Example 3, an adversary may apply $C4.5$ classifier on data table D^* to predict the sensitive attribute value of an individual who is a part of an anonymized training data. In addition, assume the adversary knows that the victim is in the table and also know the victim's qid, i.e., $female$ and her occupation is

$Painter$. By applying this knowledge on the anonymized training data, he finds a total of 4 records on class attribute $Disease$ which all belong to sensitive value HIV. So, the likelihood of privacy breach L_{PB} for this case becomes 4/4 which is calculated based on Equation 12, implying that an adversary has 100% confidence on inferring sensitive disease of the victim.

Chapter 5

Empirical Study

In this chapter, our objectives are to study the impact of enforcing different data anonymization methods at different privacy protection levels on person-specific dataset with respect to the information utility for data mining in terms of monetary value. We perform the following tests on person-specific dataset (1) to measure the classification accuracy on the class label and on the sensitive attribute, (2) to measure the cost of distortion, (3) to measure the cost in terms of classification quality, (4) to estimate the probability of attack by using precision and recall performance measures, (5) to quantify the likelihood of privacy breach which impacted by adversary prior knowledge about victim, and (6) to perform net cost-benefit analysis to measure the revenue, potential damage cost and the optimal cost on the released data.

The real-life dataset *Adult* obtained from the UCI Machine Learning Repository is employed in our experiments. This dataset has been widely used for different research purposes and is the *de-facto* benchmark for comparing performance of anonymization algorithms [HJM07] [YC11] [FWY07]. It comprises of $45,222$ records on 8 categorical attributes, 6 numerical attributes, and a binary $Income$ class from the US Census database after removing the records with unknown instances. In our study, we set $Married\text{-}civ\text{-}spouse$ and $Divorced$ in the attribute $MaritalStatus$ as sensitive, and the remaining 13 attributes as QID. All experiments were performed on a machine with an Intel dual core 1.8GHz processor with 2GB memory.

5.1 Classification Accuracy on the Class Label and on the Sensitive Attribute

Let $Income$ be the class attribute, denoted by $Class\_Income$. Let $MaritalStatus$ be the sensitive attribute, denoted by $Sens\_MaritalStatus$.

Figure 3 depicts the classification accuracy CA for general data analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that CA on the class attribute $Income$ *generally* decreases as K or L increases, but not monotonically. For example, CA on $Income$ increases slightly by 0.1% when K increases from 40 to 50 for $L = 2$. In comparison to CA on the sensitive attribute $MaritalStatus$ *generally* decreases as K or L increases, but with some irregularities. For example, CA on $MaritalStatus$ increases by 0.4% when K increases from 10 to 20 for $L = 2$, and increases by 0.2% when K increases from 30 to 40 for $L = 6$. CA increases in this case because generalization removes the noise. However, as L increases to 6, the CA of LKC-privacy equals to the CA of traditional K-anonymity for both class attribute $Income$ and sensitive attribute $MaritalStatus$, but the CA remains unchanged with respect to the change of confidence threshold $10\% \leq C \leq 50\%$.

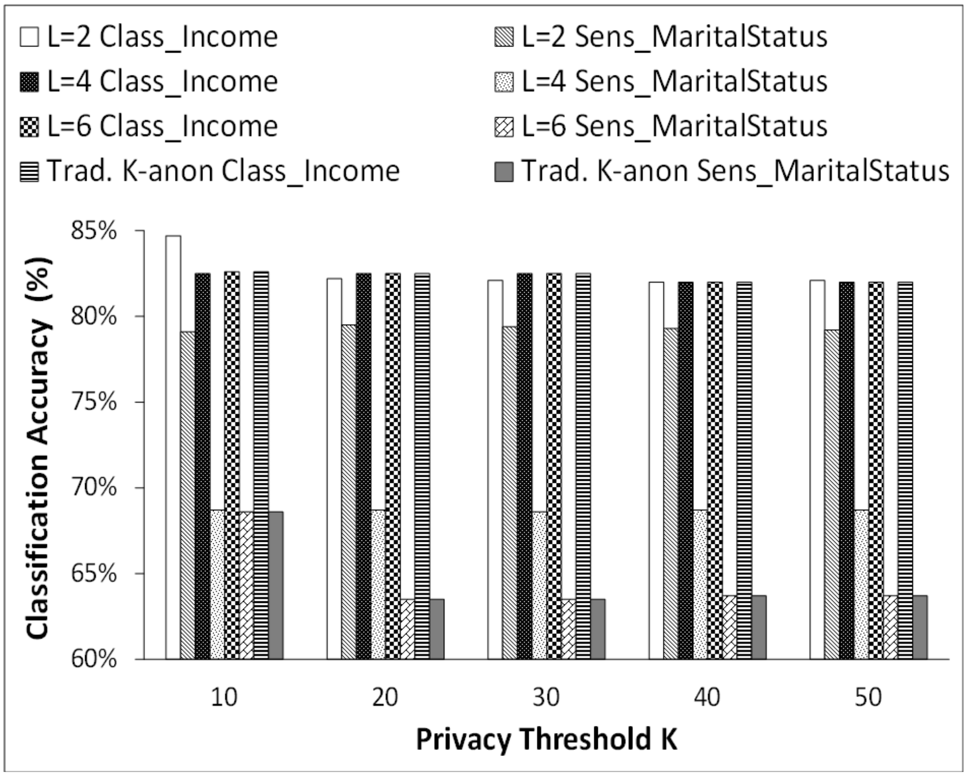

Figure 3: CA on Income and MaritalStatus for General Analysis

Figure 4 depicts the classification accuracy CA for classification analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that CA on the class attribute $Income$ *generally* decreases as L increases, but not monotonically with the increase in K. For example, CA on $Income$ increases by 3.1% when K increases from 10 to 20 for $L = 4$ and $L = 6$. In comparison to CA on the sensitive attribute $MaritalStatus$ *generally* decreases as L increases, but not monotonically with the increase in K. For example, CA on $MaritalStatus$ increases slightly by 0.6% when K increases from 30 to 50 for $L = 4$ and $L = 6$. The CA of LKC-privacy equals to the CA of traditional K-anonymity for both class attribute $Income$ and sensitive attribute $MaritalStatus$ when $L = 4$ and $L = 6$, but the CA remains unchanged with respect to the change of confidence threshold $10\% \leq C \leq 50\%$.

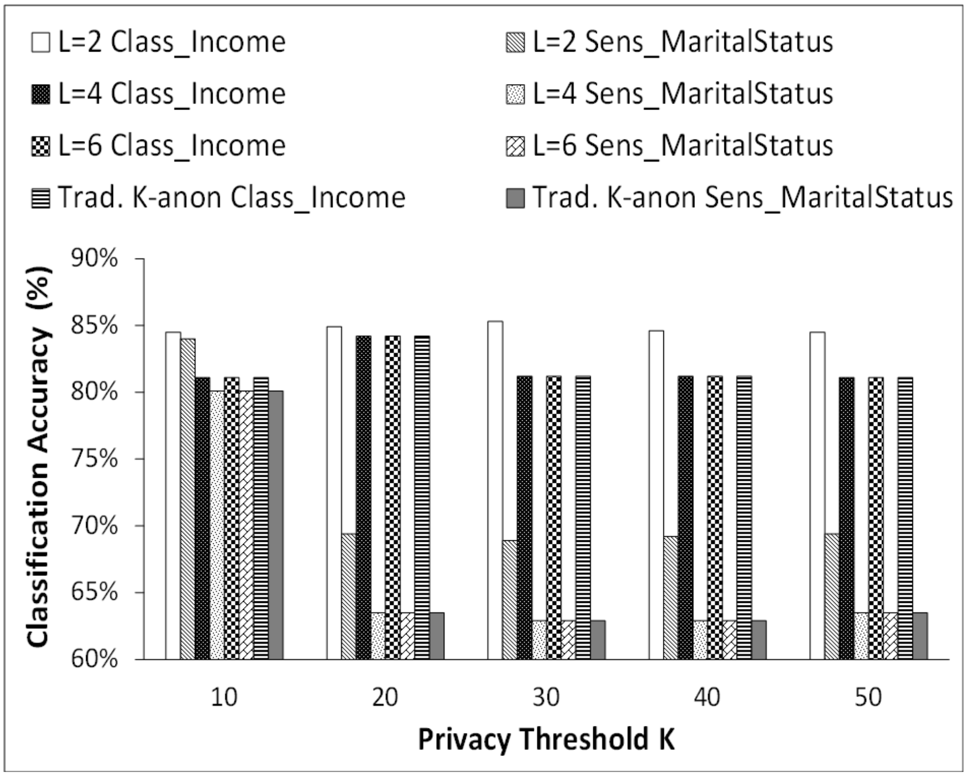

Figure 4: CA on Income and MaritalStatus for Classification Analysis

5.2 Cost of Distortion

Suppose the sensitivity of the dataset $SD = 3$, the price per record $Pr_{rec} = \$0.2$, the cost of lawsuit $Cost_{lwst} = \$1,000$, and the size of dataset $Size_{ds} = 45,222$.

Figure 5 depicts the cost of distortion CoD for general data analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that CoD *generally* increases as K or L increases, but sometimes has a fall when $K = 10$ and $L = 6$ and then increases gradually as K increases. This anti-monotonic property of the greedy algorithm helps in identifying the sub-optimal solution. The CoD of LKC-privacy equals to the CoD of traditional K-anonymity when $L = 6$. CoD is insensitive to change of confidence threshold $10\% \leq C \leq 50\%$.

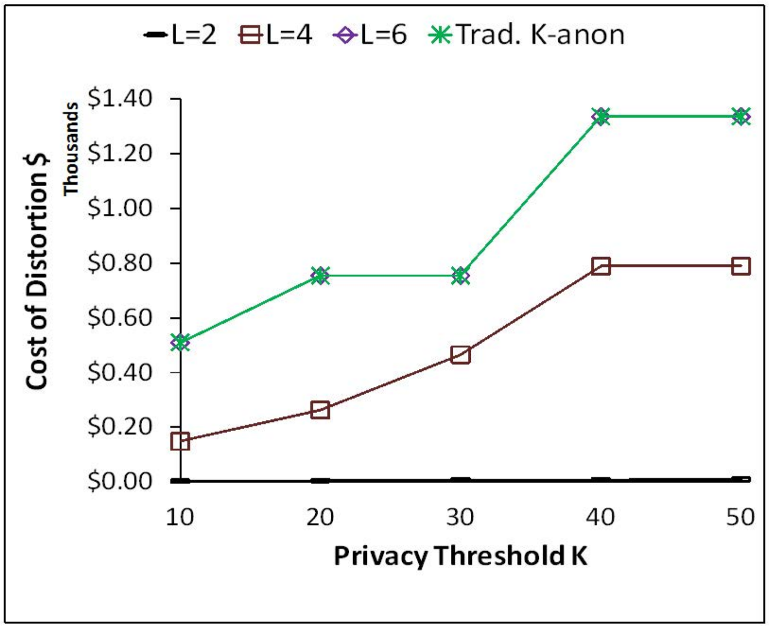

Figure 5: Cost of Distortion for General Analysis

5.3 Cost in terms of Classification Quality

Suppose the sensitivity of the dataset $SD = 3$, the price per record $Pr_{rec} = \$0.2$, the cost of lawsuit $Cost_{lwst} = \$1,000$, and the size of dataset $Size_{ds} = 45,222$. *Baseline Accuracy* (BA) as calculated on raw data without anonymization is 85.3%.

Figure 6 depicts the cost in terms of classification quality CQC for classification analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that CQC *generally* increases as L increases, but it is not consistent with the increase in K for specific level L. For example, CQC decreases by $\$841.12$ when K increases from 10 to 20 for $L = 4$ and $L = 6$. This fall happens because the cost of anonymization in terms of classification accuracy $(BA - CA)$ reduces from 4.2% to 1.1% and it aids in finding the sub-optimal solution. The CQC of LKC-privacy equals to the CQC of traditional K-anonymity when $L = 4$ and $L = 6$. CQC is insensitive to the change of confidence threshold $10\% \leq C \leq 50\%$.

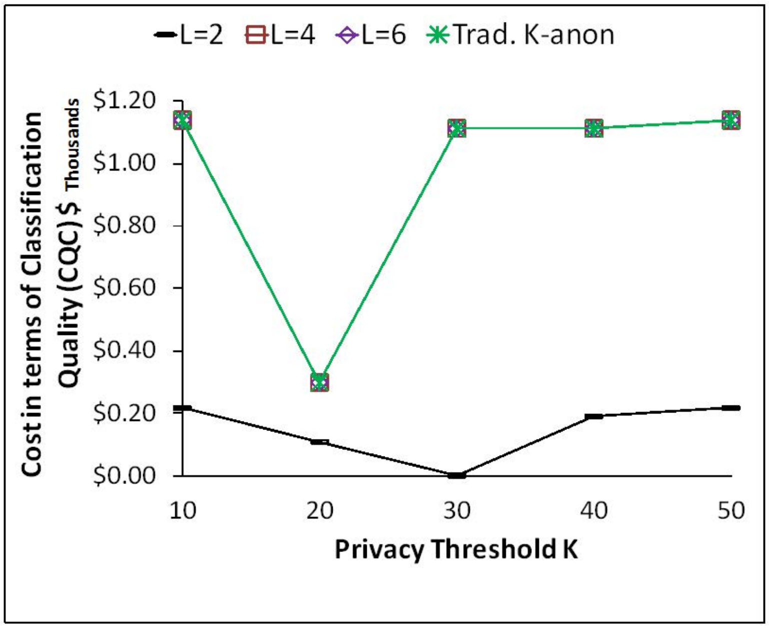

Figure 6: Cost in terms of Classification Quality for Classification Analysis

Figure 7 depicts the cost in terms of classification quality $Cost\_CQ$ using *DiffGen* for classification analysis with the specified parameters for privacy budget $\epsilon = 0.5, 1$ and specialization levels $3 \leq h \leq 19$. We use 30162 records of the real-life *adult* dataset to build the classifier and then measure the accuracy on the remaining 15060 records. We use 10-fold cross-validation to estimate the average accuracy. We observe that $Cost\_CQ$ *generally* decreases, as specialization level h increases, except when privacy budget $\epsilon = 0.5$ and specialization level h increase from 15 to 19. This is because Laplace noise overpower when specialization level h get increase from certain threshold. When average accuracy increases, $Cost\_CQ$ decreases. For example, average accuracy increases by 1.11% and $Cost\_CQ$ decreases by $\$301.18$ when specialization level h increases from 15 to 19 for privacy budget $\epsilon = 1$.

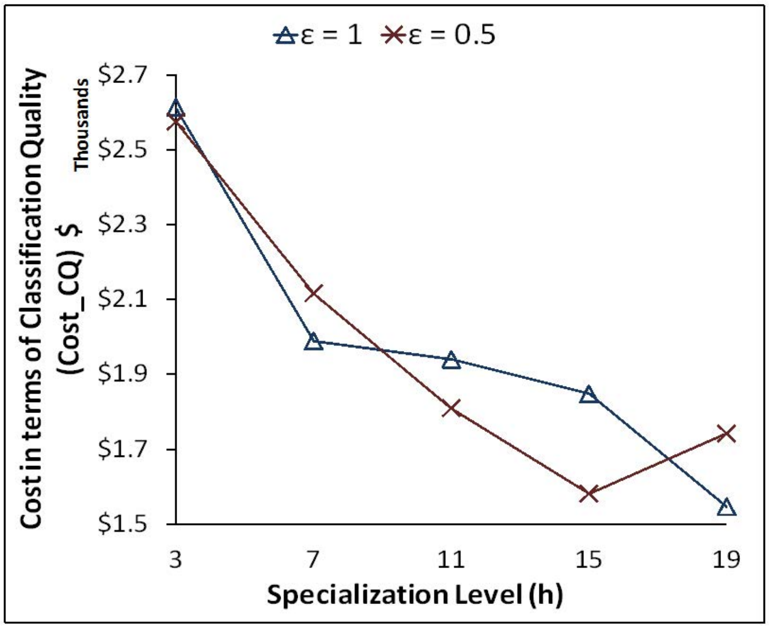

Figure 7: DiffGen Privacy Cost_CQ for Classification Analysis

5.4 Probability of Attack

Figure 8 depicts the probability of attack $Prob_{atk}$ for sensitive value $Married\text{-}civ\text{-}spouse$ in case of general data analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that $Prob_{atk}$ *generally* decreases as L increases, but not monotonically with the increase in K. The lowest value 71.27% of privacy attack inferred when $L = 6$ for $K = 20$ and $K = 30$, because of larger equivalence group. It provides trade-off in increasing the level of privacy protection and reducing the chances of privacy attack. The $Prob_{atk}$ of LKC-privacy equals to the $Prob_{atk}$ of traditional K-anonymity when $L = 6$. $Prob_{atk}$ is insensitive to change of confidence threshold $10\% \leq C \leq 50\%$.

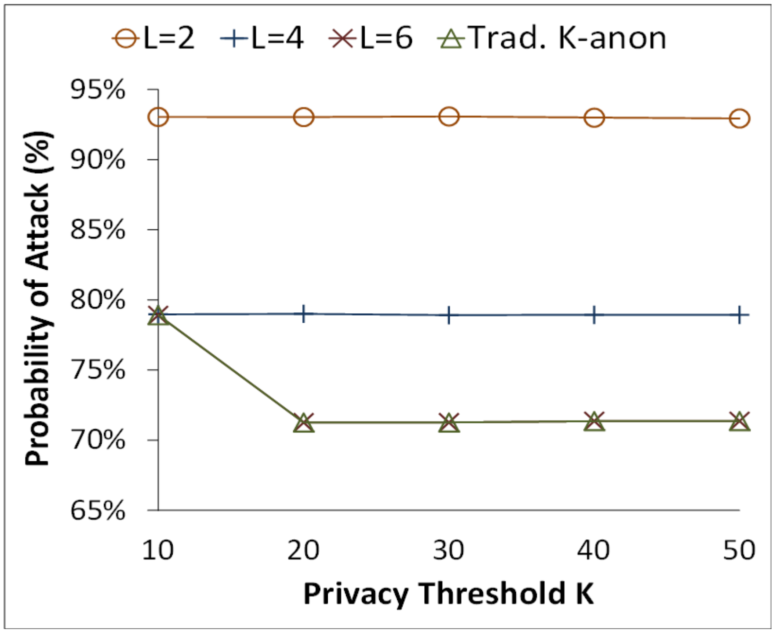

Figure 8: Probability of Attack for General Analysis

Figure 9 depicts the probability of attack $Prob_{atk}$ for sensitive value $Married$-civ-$spouse$ in case of classification analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that $Prob_{atk}$ *generally* decreases as increase in K for specific level L, but not monotonically. The lowest value 70.00% of privacy attack inferred when $L = 4$ and $L = 6$ for $K = 30$ and $K = 40$. It provides trade-off in increasing the level of privacy protection and reducing chances of privacy attack. The $Prob_{atk}$ of LKC-privacy equals to the $Prob_{atk}$ of traditional K-anonymity when $L = 4$ and $L = 6$. $Prob_{atk}$ is insensitive to change of confidence threshold $10\% \leq C \leq 50\%$.

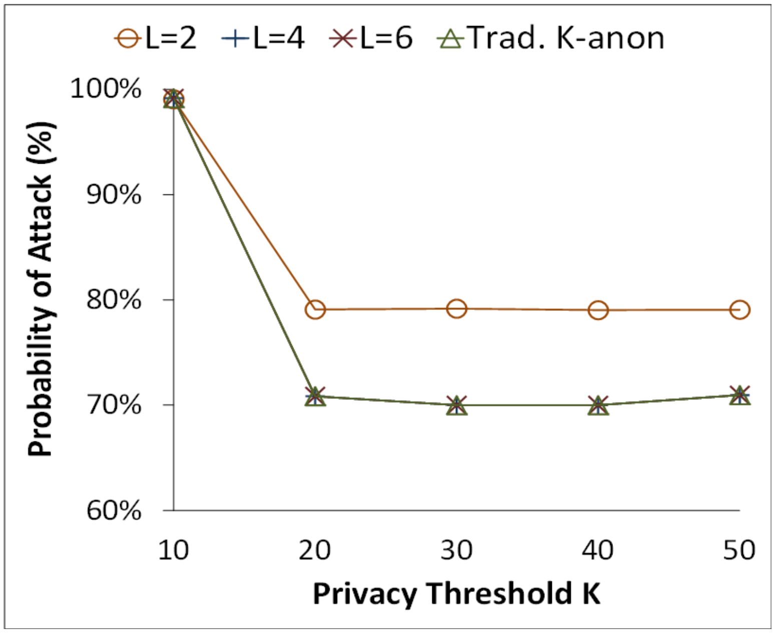

Figure 9: Probability of Attack for Classification Analysis

5.5 Likelihood of Privacy Breach

Suppose the adversary has a background knowledge about victim. He knows victim's age is in between 46 to 50, sex is $Male$, $education$-num is ≥ 13, $native$-$country$ is $Canada$, and $salary$ is $> 50,000$.

Figure 10 depicts the likelihood of privacy breach L_{PB} when adversary apply his background knowledge on sensitive value $Married$-civ-$spouse$ in case of general data analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that L_{PB} changes as increase in adversary knowledge L but it remains the same with the increase in K except for some irregularities. For example, L_{PB} increases from 87.04% to 88.49%, as L increases from 2 to 4 but remains the same with increase in K for specific level L. When L increases from 4 to 6 and K increases from 10 to 20 there is a fall of 5.06% in L_{PB}. This anti-monotonic property of the greedy algorithm helps in identifying the sub-optimal solution. It provides trade-off in privacy-preserving as higher value of L_{PB} results in more potential damage cost. The

L_{PB} of LKC-privacy equals to the L_{PB} of traditional K-anonymity when $L = 6$. L_{PB} is insensitive to change of confidence threshold $10\% \leq C \leq 50\%$.

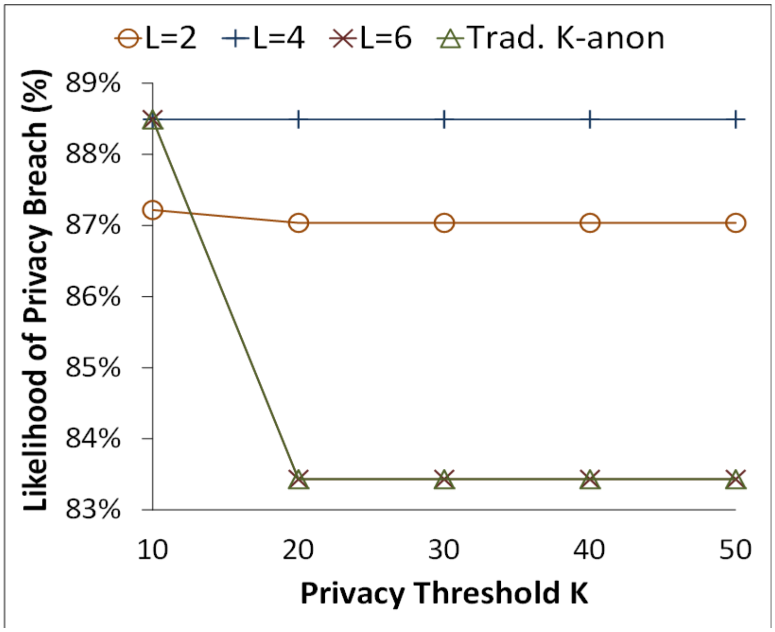

Figure 10: Likelihood of Privacy Breach for General Analysis

Figure 11 depicts the likelihood of privacy breach L_{PB} when adversary apply his background knowledge on sensitive value $Married\text{-}civ\text{-}spouse$ in case of classification analysis with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that L_{PB} increases, as increase in L or K, but not monotonically with the increase in K. For example, L_{PB} drop by 3.47%, as increase in K from 10 to 20 when $L = 4$ and $L = 6$. This anti-monotonic property of the greedy algorithm helps in identifying the sub-optimal solution. It provides trade-off in privacy-preserving as higher value of L_{PB} results in more potential damage cost. The L_{PB} of LKC-privacy equals to the L_{PB} of traditional K-anonymity when $L = 4$ and $L = 6$. L_{PB} is insensitive to change of confidence threshold $10\% \leq C \leq 50\%$.

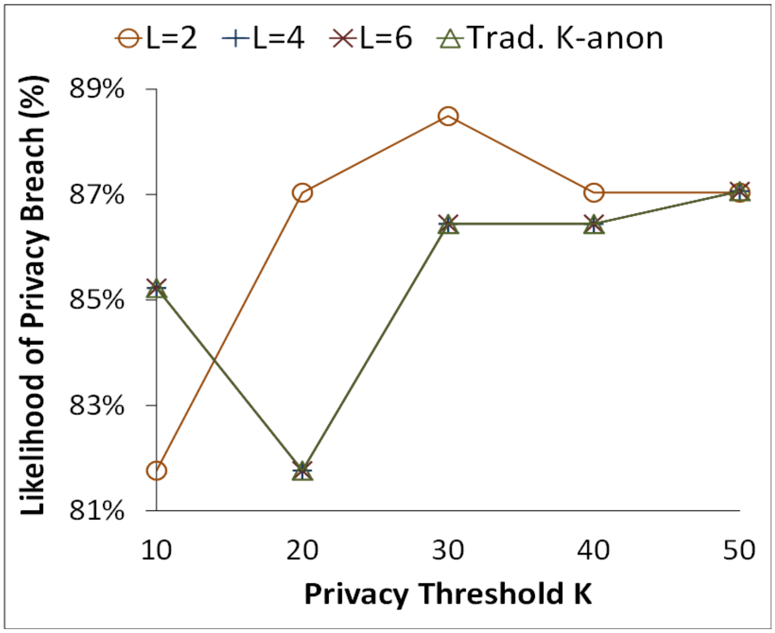

Figure 11: Likelihood of Privacy Breach for Classification Analysis

5.6 Net Cost-Benefit Analysis

Suppose the sensitivity of the dataset $SD = 3$, the price per record $Pr_{rec} = \$0.2$, the cost of lawsuit $Cost_{lwst} = \$1,000$, the fixed operating expenses $F_{oe} = \$100$, and the size of dataset $Size_{ds} = 45,222$. *Baseline Accuracy* (BA) as calculated on raw data without anonymization is 85.3%.

Figure 12 depicts the net cost-benefit analysis for general data analysis case to estimate the Rev_{ga}, PDC, NV_{ga}, and Opt_{cost} with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that Rev_{ga} *generally* decreases as K or L increases. PDC *generally* decreases as K or L increases, but not monotonically with the increase in K. For example, PDC increases slightly by \$13.3 when K increases from 30 to 50 for $L = 6$. Minimum PDC is desired subject to the utility of the data so PDC exhibits point of trade-off in privacy preservation. NV_{ga} estimates the impact of PDC on Rev_{ga}. Maximum NV_{ga} returns the Opt_{cost} \$$12,187$ for $K = 20$ and $K = 30$ when $L = 6$. Rev_{ga} and PDC of LKC-privacy equals to the Rev_{ga} and PDC of traditional K-anonymity when $L = 6$. Rev_{ga} and PDC remain unchanged with respect to the change of confidence threshold $10\% \leq C \leq 50\%$.

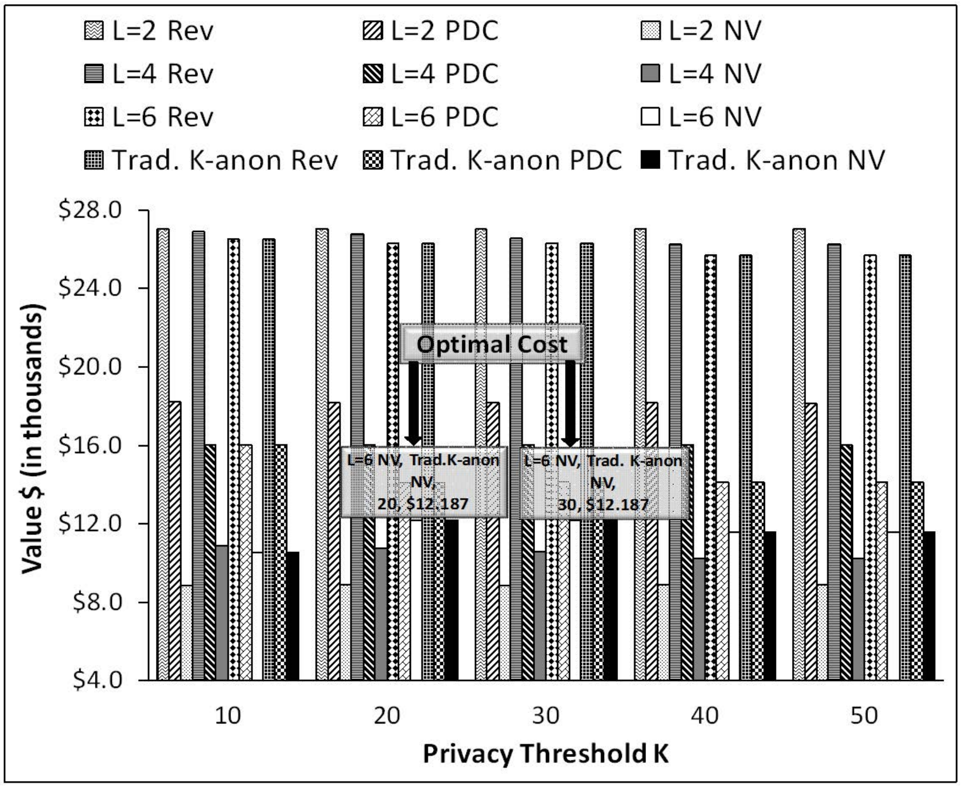

Figure 12: General Analysis Optimal Cost

Figure 13 depicts the net cost-benefit analysis for classification analysis case to estimate the Rev_{ca}, PDC, NV_{ca}, and Opt_{cost} with privacy threshold $10 \leq K \leq 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that Rev_{ca} *generally* decreases as L increases, but it is not consistent with the increase in K for specific level L. For example, Rev_{ca} increases by $\$841.12$ when K increases from 10 to 20 for $L = 4$ and $L = 6$. PDC *generally* increases as L increases, but not monotonically with the increase in K. For example, PDC decreases significantly by $\$5,160.52$ when K increases from 10 to 20 for $L = 4$ and $L = 6$. Minimum PDC is desired subject to the utility of the data so PDC exhibits point of trade-off in privacy preservation. NV_{ca} estimates the impact of PDC on Rev_{ca}. Maximum NV_{ca} returns the Opt_{cost} $\$12,934$ for $K = 20$ when $L = 4$

and $L = 6$. Rev_{ca} and PDC of LKC-privacy equals to the Rev_{ca} and PDC of traditional K-anonymity when $L = 4$ and $L = 6$. Rev_{ca} and PDC remain unchanged with respect to the change of confidence threshold $10\% \leq C \leq 50\%$.

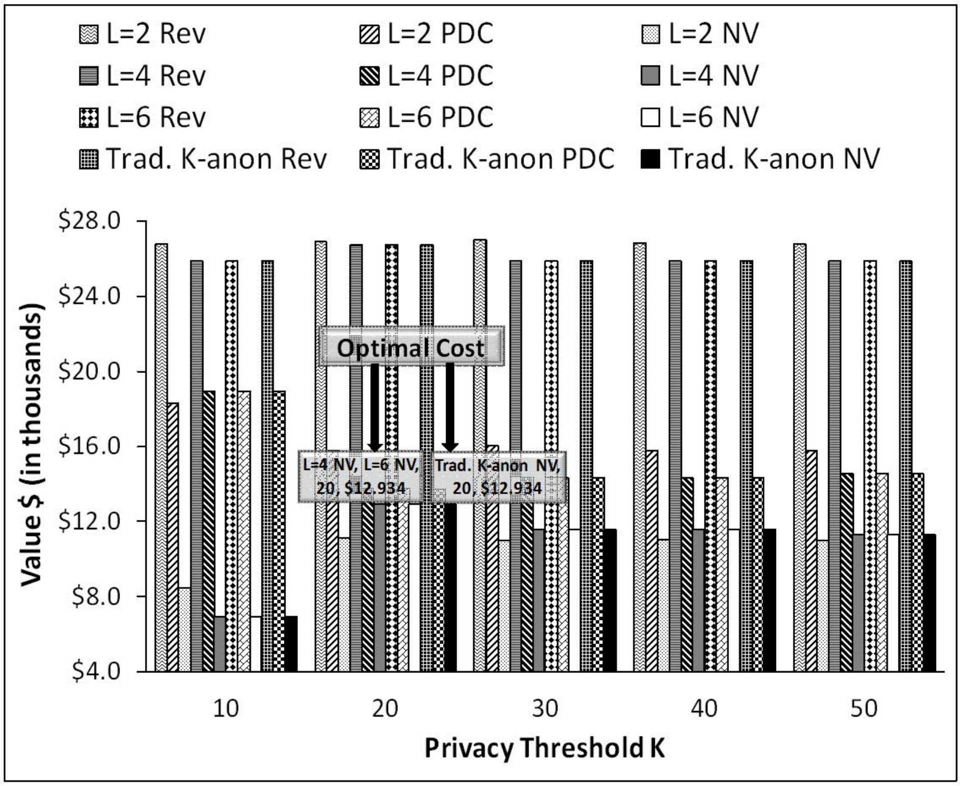

Figure 13: Classification Analysis Optimal Cost

Figure 14 depicts the revenue of using *DiffGen* for classification analysis with the specified parameters for privacy budget $\epsilon = 0.5, 1$ and specialization levels $3 \leq h \leq 19$. We observe that the revenue *generally* increases, as increase in specialization level h, for the specified privacy budget ϵ, except has a fall when $\epsilon = 0.5$ and specialization level h increases from 15 to 19. This is because Laplace noise overpower when specialization level h get increase from certain threshold. Maximum revenue $\$25,486.61$ is achieved, when $\epsilon = 1$ and $h = 19$. *DiffGen* guard against data breach by protecting against adversaries with background knowledge. We learned that by applying adversary knowledge mentioned in Section 5.5 on the anonymized *DiffGen* data have no considerable privacy breach. So, it reflects no potential damage cost.

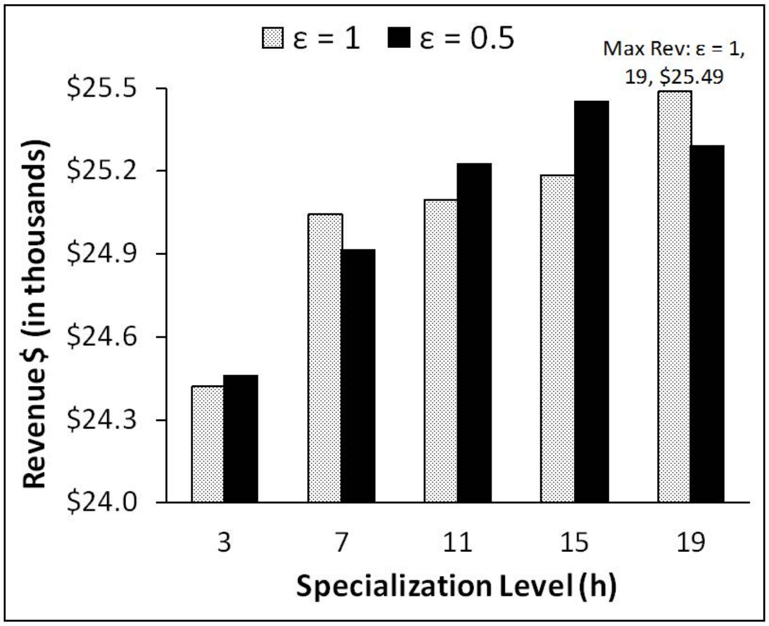

Figure 14: DiffGen Privacy Revenue

5.7 Summary of Empirical Study

We evaluate our proposed method to search for optimal value by performing experiments on a real-life dataset. The HIC can compare costs and benefits by choosing different privacy models, namely, K*-anonymity*, LKC*-privacy*, and ϵ*-differential privacy*. We learned that the revenue and potential damage cost are *generally* high when the privacy protection level is low by applying the cost factors on privacy models K-anonymity and LKC-privacy. The optimal cost happens at the maximum of net value NV; however, the maximum NV would not be found at lower privacy protection level. The maximum net value NV can be identified by gradually increasing privacy protection level and evaluating associated cost factors. For the case of classification analysis, by applying the cost factors on K-anonymity, revenue is not at the best with privacy level set as low, although the potential damage cost is high. When we apply cost factors on LKC-privacy model, neither revenue nor potential damage cost is at the best with privacy level set as low, this is due to the fact heuristic information gain used for classification and greedy algorithm search for sub-optimal solution with flexibility of adjusting adversary knowledge L. With the increase in privacy level some time results in better classification structure. Costs and benefits would vary with the change of parameters K and L, but confidence C does not produce change on outcome. When apply cost factors on *DiffGen* revenue *generally* increases, as increase in specialization level for the specified privacy budget, except has drop when specialization level increase from certain threshold. We also learned that by applying adversary knowledge on the anonymized *DiffGen* data have no considerable privacy breach. So, it is considered as safe from potential damage cost.

Chapter 6

Conclusion and Future Work

In this chapter we conclude the thesis and provide some possible research directions that can be conducted as a future work.

We propose an analytical cost model which can benefit health information custodians (HICs) in making better decisions while sharing health records for secondary use and commercial purposes. Our model provides trade-off in terms of monetary value between preserving privacy and extracting useful patterns or trends for both general data analysis and classification analysis. Our proposed solution discusses the relevant quantitative and qualitative cost factors associated with revenue earnings and potential damages. We present an attack model based on the well-known $C4.5$ classification model, then use precision and recall to evaluate the probability of attack, and measure the likelihood of privacy breach against sensitive value of the victim by applying adversary background knowledge on the anonymized data.

Our cost-benefit model and the factors employed in finding the trade-off between privacy and utility will be applicable to other privacy-preserving data publishing scenarios. This work sheds light for future research that studies the trade-off between privacy protection and information utility with different perturbative and anonymization techniques for other types of data, such as transaction [CMF$^+$11], trajectory [FCDX09], and social network data [ZP11].

Bibliography

[AAC$^+$11] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: On the trade-off between utility and information leakage. *CoRR*, abs/1103.5188, 2011.

[AFT06] Alessandro Acquisti, Allan Friedman, and Rahul Telang. Is there a cost to privacy breaches? an event study. In *Proceedings of the 27th International Conference on Information Systems*, 2006.

[AFWM10] Ahmed AL Faresi, Duminda Wijesekera, and Khaled Moidu. A comprehensive privacy-aware authorization framework founded on HIPAA privacy rules. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 637–646, 2010.

[Agg05] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB)*, pages 901–909, 2005.

[BEP00] David Baumer, Julia Brande Earp, and Fay Cobb Payton. Privacy of medical records: IT implications of HIPAA. *SIGCAS Computers and Society*, 30(4):40–47, 2000.

[BH03] Kevin Beaver and Rebecca Herold. *The Practical Guide to HIPAA Privacy and Security Compliance*. Auerbach, 2003.

[BL11] Paige Backman and Karen Levin. Privacy breaches - impact, notification and strategic plans. 2011.

[BM05] Christine Bycroft and Katherine Merrett. Experience of using a post randomisation method at the office for national statistics. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, 2005.

[BRL12] Ann Bevitt, Karin Retzer, and Joanna Lopatowska. Dealing with data breaches in europe and beyond. 2012.

[CBC12] CBC Canada. Privacy breach lawsuit launched against western health. 2012. http://www.cbc.ca/news/canada/newfoundland-labrador/story/2012/08/17/nl-privacy-breach-lawsuit-launced-817.html.

[CMF$^+$11] Rui Chen, Noman Mohammed, Benjamin C.M. Fung, Bipin C. Desai, and Li Xiong. Publishing set-valued data via differential privacy. *In Proceedings of the VLDB Endowment*, 4(11):1087–1098, 2011.

[CU08] Robert Cooter and Thomas Ulen. *Law & Economics*. Pearson Education, 5th edition, 2008.

[DD05] Jean-Pierre Danthine and John B. Donaldson. *Intermediate Financial Theory*. 2005.

[DeG13] John DeGaspari. Hospice of north idaho settles HIPAA security case for $50,000. 2013. http://www.healthcare-informatics.com/news-item/hhs-announces-first-hipaa-breach-settlement-involving-less-500-patients.

[Dep13] Department of Health and Human Services. Modifications to the HIPAA privacy, security, enforcement, and breach notification rules under the HITECH Act and the GINA Act; other modifications to the HIPAA rules. (78 FR 5565):5565–5702, 2013.

[DFMS02] Josep Domingo-Ferrer and Josep M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[DFOTMS02] Josep Domingo-Ferrer, Anna Oganian, Àngel Torres, and Josep M. Mateo-Sanz. On the security of microaggregation with individual ranking: analytical attacks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):477–491, 2002.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography (TCC)*, pages 265–284. Springer-Verlag, 2006.

[DR82] Tore Dalenius and Steven P. Reiss. Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6(1):73 – 85, 1982.

[Dwo06] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture notes in computer science*, pages 1–12. Springer-Verlag, 2006.

[Dwo08] Cynthia Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC)*, pages 1–19, Berlin, Heidelberg, 2008. Springer-Verlag.

[EG95] Edwin J. Elton and Martin Jay Gruber. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, 5th edition, 1995.

[EH13] Jessica Ellis and Bronwyn Harris. What is a civil suit? 2013. http://www.wisegeek.com/what-is-a-civil-suit.htm.

[FCDX09] Benjamin C.M. Fung, Ming Cao, Bipin C. Desai, and Heng Xu. Privacy protection for RFID data. In *Proceedings of the 24th ACM SIGAPP Symposium on Applied Computing (SAC)*, pages 1528–1535, 2009.

[FWCY10] Benjamin C.M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14:1–14:53, 2010.

[FWFY10] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC Data Mining and Knowledge Discovery. Taylor & Francis, 2010.

[FWY07] Benjamin C.M. Fung, Ke Wang, and Philip S. Yu. Anonymizing classification data for privacy preservation. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):711–725, 2007.

[Geh10] Johannes Gehrke. Programming with differential privacy: Technical perspective. *Communications of the ACM*, 53(9):88–88, 2010.

[Ghu10] Karminder Ghuman. *Management: Concepts, Practice & Cases*. Tata McGraw-Hill Education, 2010.

[GRS09] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 351–360, 2009.

[HDFF$^+$12] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul de Wolf. *Statistical Disclosure Control*. John Wiley & Sons, 2012.

[HJM07] Bijit Hore, Ravi Chandra Jammalamadaka, and Sharad Mehrotra. Flexible anonymization for privacy preserving data publishing: A systematic search based approach. In *Proceedings of the 7th SIAM International Conference on Data Mining (SDM)*, 2007.

[Inf13] Information Security Media Group. Global payments breach tab: $94 million. 2013. http://www.bankinfosecurity.com/global-payments-breach-tab-94-million-a-5415/op-1.

[JPH12] Nicola Jentzsch, Sören Preibusch, and Andreas Harasser. Study on monetising privacy: An economic model for pricing personal information. European Network and Information Security Agency (ENISA), 2012.

[KCG11] Paul H. Keckley, Sheryl Coughlin, and Shiraz Gupta. Privacy and security in health care: A fresh look. 2011.

[KF09] Edward C. S. Ku and Yi Wen Fan. The decision making in selecting online travel agencies: An application of analytic hierarchy process. *Journal of Travel and Tourism Marketing*, 26(5-6):482–493, 2009.

[Kif09] Daniel Kifer. Attacks on privacy and definetti's theorem. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 127–138, 2009.

[Kim86] Jay Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods*, pages 303–308. American Statistical Association, 1986.

[KM11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 193–204, 2011.

[Koc11] Richard Koch. *The 80/20 Principle: The Secret to Achieving More with Less*. Random House Digital, Inc., 2011.

[KT05] Hian Chye Koh and Gerald Tan. Data mining applications in healthcare. *Healthcare Information Management*, 19(2):64–72, 2005.

[KWG97] Peter Kooiman, Leon Willenborg, and Jose Gouweleeuw. *PRAM: A Method for Disclosure Limitation of Microdata*. Research paper. CBS, 1997.

[Lic12] Lauren B. Licastro. HIPAA/HITECH enforcement action alert. 2012.

[Lit93] Roderick J.A. Little. Statistical analysis of masked data. *Journal of Official Statistics*, 9(2):407–426, 1993.

[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)*, pages 517–526, 2009.

[LS08] Grigorios Loukides and Jianhua Shao. Data utility and privacy protection trade-off in k-anonymisation. In *Proceedings of the International Workshop on Privacy and Anonymity in Information Society (PAIS)*, pages 36–45, 2008.

[MCFY11] Noman Mohammed, Rui Chen, Benjamin C.M. Fung, and Philip S. Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)*, pages 493–501, 2011.

[MFHL09] Noman Mohammed, Benjamin C.M. Fung, Patrick C.K. Hung, and Cheuk-Kwong Lee. Anonymizing healthcare data: A case study on the blood transfusion service. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)*, pages 1285–1294, 2009.

[Min13] Mind Tools Ltd. Decision making techniques. 2013. http://www.mindtools.com/dectree.html.

[Moo96] Richard A. Moore. Controlled data-swapping techniques for masking public use microdata sets. Technical report, Statistical Research Division, U.S. Bureau of the Census, 1996.

[NCGC09] John M. Neclerio, Kathleen Cheney, C.Mitchell Goldman, and Lisa W. Clark. Adopting electronic medical records: What do the new federal incentives mean to your individual physician practice? *Journal of Medical Practice Management*, 25(1):44–8, 2009.

[NHT08] Jordi Nin, Javier Herranz, and Vicenc Torra. How to group attributes in multivariate microaggregation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 16:121–138, 2008.

[Off09] Office of the Privacy Commissioner for Personal Data. Review of the personal data (privacy) ordinance. 2009.

[Off12] Office for Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with the HIPAA privacy rule. 2012.

[PMJ02] Kathryn A Phillips, Tara Maddala, and F.Reed Johnson. Measuring preferences for health care interventions using conjoint analysis: An application to HIV testing. *Health Services Research*, 37(6):1681–1705, 2002.

[Pon12] Ponemon Institute LLC. 3rd annual benchmark study on patient privacy and data security. 2012. http://www.ponemon.org/library/third-annual-patient-privacy-data-security-study.

[Pon13] Ponemon Institute LLC. 2012 most trusted companies for privacy: Study of consumers in the united states. 2013. http://www.ponemon.org/library/2012-most-trusted-companies-for-privacy-1.

[Qui93] John Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[RB12] Brad M. Rostolsky and Nancy E. Bonifant. Massachusetts attorney general strikes: South shore hospital settles data breach allegations for $750,000. 2012. http://www.lexology.com/library/detail.aspx?g=feab1157-555a-4653-b21c-46457d71159f.

[RHA12] Sasha Romanosky, David A. Hoffman, and Alessandro Acquisti. Empirical analysis of data breach litigation. 2012.

[Saa08] Thomas L. Saaty. Decision making with the analytic hierarchy process. *International Journal Services Sciences*, 1(1):83–98, 2008.

[Sch12] Mathew J. Schwartz. Zappos hack exposes passwords. 2012. http://www.informationweek.com/security/attacks/zappos-hack-exposes-passwords/232400441.

[Slo92] Roman Slowinski. *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory*, chapter The discernibility matrices and functions in information systems, pages 331–362. 1992.

[SMOW94] Chris Skinner, Catherine Marsh, Stan Openshaw, and Colin Wymer. Disclosure control for census microdata. *Journal of Official Statistics*, 10(1):31–51, 1994.

[Sra10] Michal Sramka. A privacy attack that removes the majority of the noise from perturbed data. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2010.

[SS98] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, 1998.

[Swe02] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[Tak99] Akimichi Takemura. Local recoding by maximum weight matching for disclosure control of microdata sets. In *ITME Discussion Paper, No.11*, 1999.

[WF10] Raymond Chi-Wing Wong and Ada Wai-Chee Fu. *Privacy-Preserving Data Publishing: An Overview*. Morgan & Claypool, 2010.

[Wit07] Normann Witzleb. Monetary remedies for breach of confidence in privacy cases. *Legal Studies*, 27(3):430–464, 2007.

[WW98] AG de Waal and Leon C.R.J. Willenborg. Optimal local suppression in microdata. *Journal of Official Statistics*, 14(4):421–435, 1998.

[WW99] Ton de Waal and Leon Willenborg. Information loss through global recoding and local suppression. 14, 1999.

[YC11] Huimin Ye and Elizabeth S Chen. Attribute utility motivated k-anonymization of datasets to support the heterogeneous needs of biomedical researchers. *In Proceedings of the AMIA Annual Symposium*, pages 1573–82, 2011.

[YS08] Abdulsalam Yassine and Shervin Shirmohammadi. Privacy and the market for private data: A negotiation model to capitalize on private data. In *Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 669–678. IEEE Computer Society, 2008.

[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.

[ZP11] Bin Zhou and Jian Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and Information Systems*, 28(1):47–77, 2011.