Optimizing the Privacy Risk - Utility Framework in Data Publication

By

Weiyi Xia

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

May, 2017

Nashville, Tennessee

Approved:

Bradley Malin, Ph.D.

Daniel Fabbri, Ph.D.

Yevgeniy Vorobeychik, Ph.D.

Eric Johnson, Ph.D.

Murat Kantarcioglu, Ph.D.

ACKNOWLEDGMENTS

First and foremost, I wish to thank my advisor, Dr. Bradley Malin at Vanderbilt University. Dr. Malin is a great mentor who guided me into the wonderful world of scientific research and taught me skills that helped me in countless endeavors in the academic world, like writing and presentation. Most importantly, he helped me to appreciate scientific research with his enthusiasm and diligence. I am very grateful to have an advisor who is encouraging. I am very thankful to Dr. Malin for all the insightful discussions and great ideas coming out from the countless meetings we had through out of years. I am also very grateful to Dr. Malin for always being there whenever I had questions or encountered obstacles in my research. Dr. Malin has also connected me to so many other experts and valuable research resources in our domain.

I would like to thank Dr. Yevgeniy Vorobeychik at Vanderbilt University for providing guidance for me with his profound vision and knowledge in computational economics and game theory, many insightful discussions and suggestions and lots of great ideas for my projects. Dr. Vorobeychik has helped me so much in my publications and the work that made this dissertation possible. I am grateful to Dr. Murat Kantarcioglu at the Univerisity of Texas at Dallas. As an expert in the domain of data privacy, Dr. Kantarciglu's has kindly given me tremendous academical support to the work that built this dissertation with his vision, great insights, sound advice and concrete suggestions. I wish to thank Dr. M. Eric Johnson and Dr. Daniel Fabbri at Vanderbilt University for accepting to be part of my Ph.D. committee and their insightful advice and suggestions.

I would like to thank Dr. Ellen Clayton at Vanderbilt University, for all the inspiration and encouragement she gave me , especially her help in opening my eyes and mind to a much bigger picture of the world in which my research work exists.

I wish to thank Dr. Raymond Heatherly for his help and teachings when he was a postdoc in the Health Information Privacy Laboratory (HIPLAB). In addition to his technolog-

ical and scientific input, Dr. Healtherly sat in the audience of my practice presentations on numerous occasions ever since I joined HIPLAB and provided me with extraordinary help in improving my academic writing and presentations. I also would like to extend my gratitude to Dr. Jiuyong Li, at University of South Australia, Xiaofeng Ding, at Huazhong University of Science and Technology in China and Dr. Joshua C. Denny at Vanderbilt University for their support in my projects. My sincere thanks go to my collaborator and friend Zhiyu Wan, a fellow Ph.D. student at HIPLAB, for being extremely nice, optimistic, patient, pleasant and resourceful when we were doing projects together. I greatly appreciate the discussions we had in various projects.

I thank the wonderful fellow HIPLABers and HIPLAB Alumni: You Chen, Muqun (Rachel) Li, Wen Zhang, Zhijun Yin, Wei Xie, Lina Sulieman, Adarsh Subbaswamy and Grayson Ruhl, all of whom are amazing in creating such a lively and creative reseasrch environment where good works and breakthroughs keep emerging. The creativity, knowledgement and intelligence of all of them have contributed to this dissertation.

There are dozens of people outside of HIPLAB have also helped me through out my Ph.D. study. I thank one of my best friends Yuan Liu for engaging in discussions with me about my work and being such a great study buddy who used to read, study and discuss machine learning books with me. I thank Dr. Douglas Fisher at Vanderbilt University for teaching me Machine learning and answering my questions in that area. I thank Dr. Julie Johnson for helping me with my teaching assistant job during my first year at Vanderbilt University.

I thank my parents for their love, support and investment in my education all these years. Finally many thanks to my husband Steve L. Nyemba for opening my eyes to the importance of organization, and taking ownship of my work and developing an intuitive understanding and passion of my domain. I cherish his love, unique wisdom and exceptional perspectives in things very much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In the past decade, we have witnessed a rapid growth in the quantity, quality, and diversity of personal data we shed through our daily activities. These data are collected by a wide range of organizations to assist in the optimization and refinement of the services they provide [2, 3].

At the same time, it is increasingly recognized that personal data has tremendous value in supporting a variety of endeavors beyond its initial purpose; e.g., ensuring transparency in activities, transparency in operation and basic discovery-driven research [4]. For instance, the personal health information in electronic medical record (EMR) systems can enable predictive modeling [5, 6], learning health systems [7], novel association studies [8, 9, 10], as well as the discovery of personalized treatment regiments [11, 12].

The secondary usage of data often leads to the need for broad access and data sharing. For instance, federal grant policies (e.g., [13]) may require sharing patient-level data for information reuse (e.g., [14]) and transparency (e.g., [15, 16]). Despite the recognized value of personal data, organizations worry about how best to protect the privacy rights of their constituents while maximizing the benefits [17].

Data privacy is an overloaded term that takes on many different forms [18]. One concern centers around private information disclosure risk of a de-identified personal dataset (i.e., a dataset in which each record is associated with one individual and explicit identifiers or identifying information of the individual (such as name and phone number) are removed). A malicious adversary might use the information gathered from the de-identified dataset with or without external data resources to infer private information of the subjects of the records in the de-identified dataset[19, 20, 21]. For example, an adversary can use de-identified genome sequencing data with auxiliary information gathered from free, pub-

licly available genetic genealogy databases from the Internet to infer the last surname of the data subject [21]. Further, the combination of the surname and other information provided with the sequencing data (e.g., race and U.S. state of residence), can lead to the identity of the data subject in some cases [21].

The majority of the prior investigations on privacy disclosure in data publishing from the computer science community focus on formal protection models (e.g., k-anonymity or ε-differential privacy). These formal models protect against a predefined possible disclosure attack conducted by a motivated adversary equipped with necessary auxiliary information. Since converting the dataset to satisfy these models often introduces noise or generalization (e.g., replacement of a 5-digit ZIP code with a 3-digit ZIP code) to the dataset, they affect the utility of the dataset. Optimization algorithms have been developed to find ways to manipulate the data to satisfy a formal model while minimizing the information loss (or maximizing the data utility).

For instance, k-anonymity [22] requires that each record in the dataset be in an equivalence group of size $\geq k$. In this case, an equivalence group is a group of records with equivalent values on a set of predefined attributes that can be used to link to an individual (e.g., demographic information) called a quasi-identifier. For example, in the de-identified dataset shown in Figure 1.1, there are 3 equivalence groups, each indicated by a different color, when the attribute set $\{gender, race, age\}$ is considered as the quasi-identifier. In this scenario, the underlying assumption is that the adversary knows that an individual is in the de-identified dataset and will link that individual to the group of records in the dataset with a matched value on the quasi-identifier. Therefore, a dataset that is k-anonymous guarantees that when such an attack happens, no individual will be linked to less than k records.

Another privacy protection model is ε-differential privacy [23]. This model is defined as a condition on the release mechanism of the dataset. In particular, a randomized algorithm A is ε-differentially private if for all datasets D_1 and D_2 that differ on a single element (i.e., data of one person), and all $S \subseteq Range(A)$, $P_r[A(D_1) \in S] \leq e^{\varepsilon} \times P_r[A(D_2) \in S]$, where

Age	Gender	ZIP	Diagnosis
33	F	372**	hypertension
33	F	372**	diabetes
33	F	372**	hypertension
55	M	37509	diabetes
55	M	37509	hypertension
70	M	4****	heart attack

Figure 1.1: An example of equivalence groups in a de-identified dataset (each color represents a equivalence group).

$Range(A)$ is the output range of the algorithm A. Differential privacy assumes that the adversary knows information about all of the individuals in the dataset except one individual and his goal is to learn the information about the last individual. Thus the release mechanism that satisfies ε-differential privacy prevents such an adversary from learning information about the last individual with certainty higher than what is determined by the parameter ε. The implementation of a differentially private release mechanism often requires introduction of noise to the released result of the algorithm A [24]. Thus it can influences the integrity of the data and render it unsuitable for certain applications, such as a medical study that requires the data to be precise [25].

While the attack scenarios for which the formal disclosure protection models are designed are possible, it does not mean they are probable. In other words, the aforementioned assumptions about the adversary's prior knowledge and motivation might be unlikely. Moreover, laws and regulations do not require perfect protection, but rather that data be shared in a manner that makes it difficult to ascertain an individual's identity. For example, the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [26] in the United States allows using an "Expert Determination" (section 164.514(b)(1)) method to meet the de-identification standard. The "Expert Determination"

methods states that "A covered entity[1] may determine that health information is not individually identifiable health information only if (1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination" [26]. Therefore, if organizations can demonstrate that the type of attacks and adversaries the formal protection models assume are of sufficiently low probability, they can be afforded an opportunity to achieve more data utility by replacing the implementation of formal protection models with risk management strategies.

1.1 Thesis Goal

The focus of this thesis is to develop a risk management framework towards an optimal balance between disclosure risk and data utility in data publishing. Instead of making an assumption about the adversarial scenario, this framework evaluates disclosure risk by explicitly taking into account the probability that the aforementioned assumptions about the adversary hold true in a particular data publishing case: 1) the adversary is motivated to conduct an attack; and 2) the adversary has the necessary prior knowledge; 3) the adversary has access to the external resources needed for the attack.

In order to calculate the probability that these assumptions hold true, a variety of parameters need to be evaluated, such as the adversary's gain from such an attack, the cost of access external resources, the penalty defined in the deterrents that are put in place (e.g., data use agreements, the time and effort to gather the external information necessary to

[1] HIPAA defines a covered entity as 1) a health care provider that conducts certain standard administrative and financial transactions in electronic form; 2) a health care clearinghouse; or 3) a health plan.

compromise the data, or penalties for misusing data), and the rate of detection of attacks. However, the value of these parameters may be uncertain, such that the model should be capable of providing a risk assessment under uncertainty.

To generate an optimal data publishing plan, three critical questions should be considered. First, how can we formally represent the disclosure risk without relying on a set of predefined assumptions? There is a solution [27] which determines whether or not the adversary will conduct an attack by a single payoff value - which is the sum of the adversary's potential gain and cost of committing an attack. Yet, this model fails to consider that, in the attack process, the adversary may choose to abort the attack in the process.

Second, how can we evaluate the values of the parameters that are incorporated into the disclosure risk framework? This is a challenging task due to several reasons: 1) There is little historical data on re-identification attacks [28]. This may be an indication of the rareness of such events. Alternatively, these events may have happened, but only behind closed doors; 2) There are various external resources, and it is often difficult to determine the probability that a record is linked to an individual in an external resource; 3) the deterrents do not always impose a penalty in monetary terms, and thus require extra analysis to convert it into a cost value.

Third, how can we find the optimal data publishing strategies based on the disclosure risk? Some of the existing solutions perturb the dataset to satisfy a formal mathematically provable constraint, such as k-anonymity [29] or ε-differential privacy [23]. Some of these solutions perturb the data to maximize the payoff of the data publisher by considering the adversary as the opponent of the data publisher in a game theoretic framework [27]. However, these methods usually do not provide a series of solutions with a range of disclosure risk and data utility levels. This is partially due to the computational complexity of the problem. Thus, this dissertation introduces efficient and scalable algorithms to search for data publishing strategies on a Risk-Utility (RU) frontier.

Figure 1.2 provides an overview of the three aims of this dissertation in the big picture

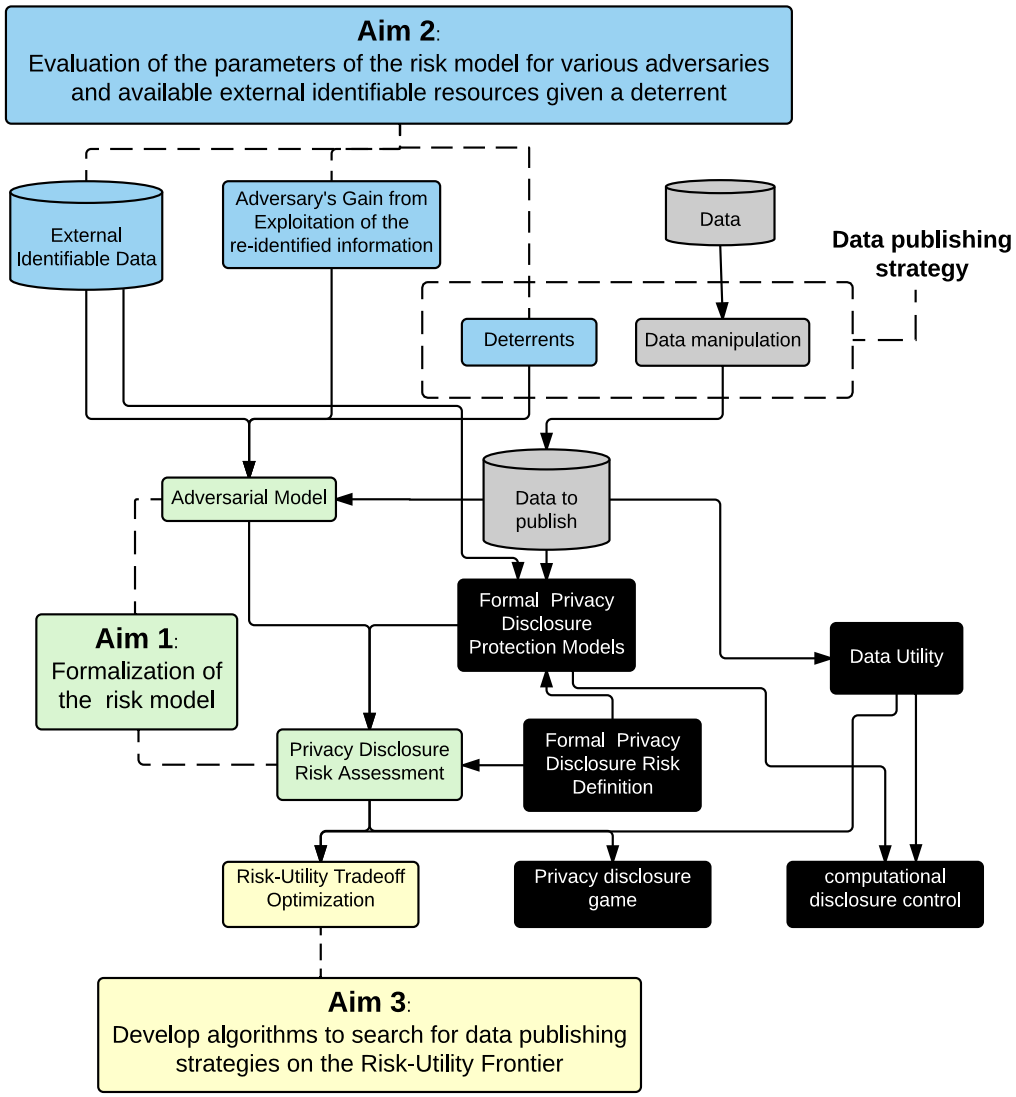


Figure 1.2: The domain of privacy disclosure control in publishing individual data and the aims of this dissertation.

of privacy disclosure control in publishing individual data. In the dissertation, we assume that a data publishing strategy is composed of a deterrent and data manipulation strategy. The modified dataset based on the manipulation strategy will be published. The data manipulation process often causes utility loss to the dataset. The goal of privacy disclosure control in publishing individual data is to find an data publishing strategy that yields an optimal outcome in terms of a set of predefined objectives, such as minimizing privacy disclosure risk and/or maximizing data utility. A data publishing strategy, as denoted in Figure 1.2 is composed of deterrents and data manipulation strategies that is used to perturb the data. Each of the three aims of this dissertation, along with the elements in the domain associated with it, are grouped by color in the figure. In general, the three aims fall into two primary subjects in the domain. The first subject is the privacy disclosure risk. In particular, this dissertation proposes a new approach to formalize a privacy disclosure risk model based on an adversarial model that takes into account external identifiable datasets, an adversary's gain and the available deterrents. The second subject is the optimization algorithms for an data publishing strategy. The subjects colored in black including data utility, computational disclosure control, and privacy disclosure games are those that are considered beyond the scope of this dissertation.

Private information disclosure can be categorized into three classes. The first class argues that privacy is compromised when a record is linked to an individual from who it was derived (often referred to as *identity disclosure* or *re-identification*) [30]. The second class is the inference of a sensitive value associated with the corresponding individual (often referred to as *attribute disclosure*) [31]. The third class is the ability to detect if someone is a member of a dataset (often referred to as the presence/absence problem) [32, 33], or the degree to which viewing an individual's contribution to a dataset permits an adversary to gain knowledge about them (the basis of models like ε-differential privacy) [34, 35, 36].

In this thesis, we focus on the identity disclosure because all other privacy disclosures require the adversary to link the data to the identity of a single individual or a group of

individuals in order to commit the specific attack or cause any harm. Moreover, existing privacy regulations are centered around the notion of anonymity. There are various regulations that encourage organizations to suppress identifying information from personal data prior to its dissemination. Several examples of regulations with explicit identity protections include HIPAA [26] in the United States and the Data Protection Directive [37] in the European Union. In particular, we assume that the data to be published is composed of a set of tuples in the form of a relational table. Each tuple contains a set of attributes of an individual. In this setting, identity disclosure means that the identity of the subject of a tuple in the published dataset, or the identities of a group of individuals that are associated with certain sensitive information provided in the dataset, are revealed unintentionally. The revelation is often not explicit, but achieved through some inference.

1.2 Problem Statement

It has been demonstrated that de-identified personal data can still be linked to, or reveal sensitive information about, the corresponding individual by adversaries [38, 21, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51]. Therefore, disclosure control methods are required for publishing personal data. In the meantime, disclosure risk is the probability that an adversary driven by economic gain given limited resources takes a series of steps to achieve a successful attack. Possible disclosures are not necessarily probable, and thus a demonstration of possible disclosure does not necessarily indicate the disclosure risk level. The majority of existing investigations in privacy disclosure control in publishing personal data only focus on providing protection methods without consideration for the probability of a privacy disclosure, i.e., the risk. Depending on the situation, applying these methods can lead to overprotection, which causes unnecessary data distortion and harms the legitimate usage of the data, as well as underprotection, which could harm the data subjects and/or data publisher.

The goal of this thesis is to build a framework to reason about disclosure risk given the

dataset and the context in which it is published including the adversary's decision making process, the adversary's gain from a successful attack, the adversary's cost for accomplishing the attack process and external identifiable data and develop methods to find data publishing solutions that provide desirable tradeoff between identity disclosure risk and data utility.

1.3 Specific Aims

There are three specific aims of this dissertation.

1.3.1 Specific Aim 1. Develop an accurate and efficient model to quantify the identity disclosure risk of individual-level records in a de-identified dataset which accounts for the re-identification attack process.

Multiple factors in the context in which the dataset is made available influence how and whether an adversary will attempt to re-identify the individuals to whom the records correspond. These include the potential gain the adversaries will obtain by exploiting the re-identified information in a way that benefits them, the potential loss the adversaries will face if the attack gets caught by the some authority or the publisher of the data if deterrence mechanisms are set in place (e.g., execution of a data use agreement (DUA)), the potentially available external data resources containing the identifiable information, and the cost of accessing these resources. We propose a model that explicitly captures these factors in a Markov decision process (MDP) and assess the adversary's motivation to initiate a re-identification attack and overcome the challenges in each step until reaching the end goal. Based on the adversary's decisions, we quantify the identity disclosure risk of the dataset as the probability that the adversary eventually reaches a successful re-identification and causes harm. The MDP model can grow exponentially with the number of state variables, and thus we propose to develop scalable algorithms to solve the model.

1.3.2 Specific Aim 2. Develop methods for evaluating the parameters of the identity disclosure risk model for various adversaries and available external identifiable resources given a deterrence mechanism.

To use an identity disclosure risk model in practice, there are at least two parameters that need to be evaluated: 1) the penalty a deterrence mechanism enforces and 2) the adversary's gain in committing a re-identification attack. We first aim to develop methods to measure the cost of penalty in an existing deterrence mechanism (e.g., time-based deterrence adopted by various organizations, such as the National Institutes of Health (NIH) database of Genotypes and Phenotypes (dbGaP) [52] [2] and Wellcome Trust Case Control Consortium (WTCCC) [53]). Time-based deterrence mechanisms do not impose a direct fine on the adversary, but instead bar the adversary from accessing the datasets for a limited period of time. Secondly, there may exist different types of adversaries, some of which may be extremely aggressive, but at the same time extremely unlikely, while others may be rational decision makers and at the same time extremely likely. Thus, we futher aim to profile the different types of adversaries in terms of their exploitations and gains from the re-identified information.

1.3.3 Specific Aim 3. Develop algorithms to search for data publishing solutions on the Risk-Utility frontier

The data publisher can combine different data manipulation strategies and deterrence mechanisms to form a data publishing solution. Data manipulation often leads to less data utility, so we aim to develop methods to automatically discover optimal data publishing solutions in the format of a series of data manipulation strategies that yields a de-identified dataset on Risk-Utility frontier; i.e., the solutions for which there are no other solutions

[2] The National Center for Biotechnology Information (NCBI) created dbGaP to serve as a platform for sharing data from large scale cohort and clinical studies initiatives of genome-wide association studies (GWAS) to enable investigator access to data from these initiatives at NIH and beyond.

with better utility and less risk than. There are two challenges we need to overcome to reach this goal: 1) formalize the solution space in a manner that efficient search algorithms can be built upon and 2) develop efficient and scalable dual-objective optimization algorithms (e.g., possibly via heuristic based search) to find optimal solutions in an extremely large solution space.

1.4 Contributions

- *We propose a novel re-identification risk framework, which formalizes incentive and deterrence mechanisms in a real world environment where the de-identified dataset is released.* This framework explicitly models the adversary as an optimal planning agent using a factored Markov decision process (FMDP). Given that the state space of the FMDP grows rapidly, we introduce a two-level linear programming algorithm to efficiently solve it. We conduct a case study in which an adversary has the option of leveraging a public voter registry in a specific U.S. state to attack de-identified records. The results illustrate how traditional beliefs about re-identification risk can be underprotective and overprotective of the data. Moreover, we conduct a detailed sensitivity analysis that demonstrates how changes in costs of each stage of attack, penalty of a violation, and violation detection rate influence when the adversary will abort an attack. The results indicate that the adversary can be sufficiently deterred with a small amount of data manipulation, provided appropriate detection and penalization policies are in place.

- *We investigate if a real world penalization mechanism that is assumed to work is actually feasible.* In particular, this penalization mechanism imposes a temporal penalty, which suspends an adversary who violates the terms in the data use agreement (which prohibits privacy violations, such as re-identification) from accessing any dataset from the system for a period of time. The temporal penalty is assumed to provide deterrence for the adversaries based on the assumption that the value of

the data for academic research declines over time. Therefore, we propose a novel approach to investigate the feasibility and effectiveness of this penalty by examining the change of the value of the data over time using linear regression. It is a challenging task to define the value of the data. Thus, in this dissertation, we use a proxy in the form of the impact of the publications that relies on the data. We enriched a dataset of publications authored by authorized investigators of data in dbGaP by adding the Journal Citation Reports (JCR) journal impact information, the dbGaP dataset made available date, and the publication made available date. We analyzed the data using several linear regression based models. The results demonstrate that there is no evidence to suggest that such a temporal policy provides the anticipated protection.

- *We develop a de-identification policy discovery platform that selects high performance de-identification policies using the tradeoff between risk and utility as the criteria.* We formally define the de-identification policy frontier discovery (DPFD) problem. Given the extremely large search space structured as a lattice, we developed a set of novel heuristic-based algorithms to construct a high quality frontier more efficiently than baseline algorithms. We conduct a extensive empirical analysis using the Adult dataset with simulated ZIP code information from 10 US states , North Carolina voter registration list and US census 2010 data to evaluate our algorithms. The result demonstrates that the heuristic algorithms outperforms the random search strategy. Moreover, we demonstrate that our approach consistently discovered frontier policies that provide more utility and less risk than a commonly adopted health data de-identification policy (in the form of HIPAA Safe Harbor).

1.5 Dissertation Outline

The reminder of this dissertation is organized as follows. Chapter 2 reviews relevant literature and highlights their limitations. Chapter 3 introduces a novel process based model

to quantify re-identification risk of de-identified personal information in a particular context in which the data is published. Chapter 4 describes the statistical analysis of a temporal penalty mechanism in a real world data setting using dbGaP data. Chapter 5 introduces the Risk-Utility frontier search problem in the de-identification solution space and several heuristic based algorithms to tackle this problem. Chapter 6 concludes this dissertation and highlights the opportunities for future research.

Chapter 2

Related Work

The problem of how to mitigate identity disclosure while keeping the data useful for a secondary purpose when publishing de-identified datasets is an essential part of the more general privacy preserving data publishing challenges. There is existing work in the computer science community on computational disclosure control via formal protection models, as well as in the statistics community on identity disclosure risk assessment. However, none of these investigations provide a framework to reason about the optimal disclosure control mechanisms under various adversarial assumptions.

In this section, we first review the big picture of privacy in data publishing. We then examine the area of computational disclosure control and statistical disclosure risk assessment with a particular focus on the identity disclosure issue. In addition, we survey the research fields related to the methods we propose including: Markov decision processes, the economics of identity disclosure, game theory, and multi-objective optimization.

2.1 Privacy in Data Publishing

There are many different definitions on what constitutes a privacy violation when publishing data that contains personal information. These views argue that privacy can be compromised when a record is linked to the individual from whom it was derived (often referred to as *identity disclosure*) [30], the inference of a sensitive value associated with the corresponding individual (often referred to as *attribute disclosure*) [31], the ability to detect if someone is a member of a dataset (often referred to as the presence/absence problem) [32, 33], or the degree to which viewing an individual's contribution to a dataset permits an adversary to gain knowledge about them (the basis of models like ε-differential privacy)

[34, 35, 36].

In this dissertation, we focus on the identity disclosure problem because this is the primary focus of current regulation. Specifically, various laws state that data are sufficiently protected when it is "difficult" to ascertain an individual's identity [54]. For example, the European Union's Data Protection Directive refers to such data as "anonymised" [37] and the U.S. Health Insurance Portability and Accountability Act (HIPAA) calls that data de-identified [55] (the convention we use henceforth). In so doing, these laws aim to prevent identity disclosure, which transpires when a recipient of the data links it with some resource containing explicit identifiers (e.g., a voter registration list [56, 19]).

2.2 Computational Disclosure Control

The majority of existing methods for addressing the issue of privacy disclosure in data publishing falls into the category of computational disclosure control, which was first proposed by Sweeney [57]. A disclosure is defined as an unintended release of explicit or inferable information about individuals who are the subjects of the person specific data. Computational disclosure control is rooted in a mathematical representation of the privacy disclosure problem. Assuming that the information in the published dataset is intended to release, the only concerns are the facts that might be inferred from the information in the published dataset and, perhaps, with other auxiliary information. As such, computational disclosure control is centered around an inference problem with two inputs: 1) the published person specific data and 2) the auxiliary information.

The goal of computational disclosure control is to provide a mathematical guarantee on the limitations of the inferable information (i.e., the unintended release) that can be obtained by solving the inference problem (or the certainty of the inference). There exist different techniques, such as generalization, suppression and noise addition, that can help to reduce what can be inferred from a published dataset. These techniques introduce different types of distortion to the data and therefore affect the data utility. The challenge is to find

a minimum distortion of data which ensures that the inferable information (or the certainty of the inference) is below a threshold.

Computational disclosure control first focused on the inference of the identity of subjects in a de-identified dataset, (i.e., identity disclosure), though over the years, it has expanded to other types of disclosures. Here, we focus on identity disclosure because it is the most relevant to the proposed study. There exist different definitions of inference with respect to identity, each of which is referred to as a protection model. These models also often assume that the de-identified dataset and the external resource are all drawn from the same population.

These models also assume that the data holder can identify a set of attributes, called quasi-identifiers, that can be used to infer the identity of an individual. A common example of a quasi-identifier is a set of variables related to an individual's demographics. Attributes that are in the quasi-identifier usually also appear in an external information that communicates the individual's identity. Thus, a linkage of the de-identified and the external datasets on the quasi-identifier can map a record in the de-identified dataset to a set of individual identities. These linkages may or may not be correct, since there might be individuals in the population who are not in the de-identified dataset.

A protection model defines assertions that can be made upon these linkages. For example, the k-map [57] model maintains the property that each record in the de-identified dataset must be linked to at least k individuals given the external resources. Other models, such as non-map and wrong-map [57], define different assertions on the invariance of the datasets. Given these models, optimization needs to be performed to turn the de-identified dataset into a form for which this assertions can be made upon with minimal distortion. However, enforcing the assertions defined by such protection models on the de-identified dataset are deemed impractical, thus there rarely methods developed for the optimization. Instead, more strict, but computationally feasible, models are considered in practice. These models are designed with the necessary conditions that if ensured, the invariant defined

by the original model will hold true. One broadly studied model is k-anonymity [38]. A dataset is said to satisfy k-anonymity when every value of the quasi-identifier occurs at least k times in the dataset. The group of records with the same value on the quasi-identifier is an equivalence group [58].

Formal protection models focus on providing a mathematically guaranteed protection level based on the chosen value of the parameters, such as k in k-map, regardless of the actual risk of a re-identification attack[1]. Thus, it is difficult for the data holders to know what specific attacks these models protect their data from and whether or not the damage caused to the data is justified by the reduced risk from applying the protection model. On the other hand, the data holders also do not know whether or not the protection is sufficient. The essential rule is that k needs to be large enough that it is very difficult for any potential adversary to go through each potential individual to unambiguously find out which individual is the actual subject of the tuple from the dataset.

To justify that a selected k is large enough to provide sufficient protection, the adversary's cost of contacting each individual, the adversary's motivation and gain need to be taken into account. However these models do not provide a principled method for reasoning about these elements. Thus, computation disclosure control can either underprotect or overprotect the dataset in practice.

Moreover, the formal protection models often face tremendous resistance in deploying in practical systems, even though the protection ensured is mathematically sound. Let us take k-anonymity as an example. Given the same k value, it provides at least the same level of protection as k-map. However, the extra distortion introduced by k-anonymity in comparison to k-map can vary to a large extent depending on the population from which the de-identified dataset and the external datasets are sampled from. Only if the set of individuals in both the de-identified dataset and the external dataset are the same, the extra

[1] Informally, the risk of re-identification attack is considered as the product of the probability that a successful re-identification happens and the harm it brings to the data subject and/or the organization that releases the data.

distortion is 0.

In practice, however the probability of this situation can be very low. In situations in which the population is extremely large in comparison to the size of the de-identified dataset, the extra distortion is very high. For example, if the de-identified dataset includes a group of 5000 patients from Vanderbilt University Hospital and the external dataset is the Davidson county voter registration list (where the hospital resides), the population set that covers both is at least the size of the Davidson county population, which is over 600,000 people. In this case, it is possible to ensure that each tuple in the de-identified dataset links to at least 1000 individuals in the population to meet the requirement of k-map when $k = 1000$. However ensuring the same k for k-anonymity can render the dataset completely useless for many applications, especially for analysis on the dataset that requires a high level of integrity of the data. By applying models such as k-anonymity, we face the risk of overprotecting the dataset substantially. Thus, computational disclosure control is impractical in practice without a principled method to evaluate the risk before, as well as after, converting the de-identified dataset into a form that satisfies the protection model.

2.3 Methods for Data Protection

Under a computational disclosure control framework, many techniques have been developed to convert the dataset to reduce the probability (or certainty) of unintended disclosures, such as generalization, suppression, and noise addition. In this dissertation, we focus on applying generalization and suppression to the de-identified dataset to mitigate disclosure risk. Since there are different generalization strategies, here, we review the most relevant to our work. [59] introduced the concept of full-domain generalization in which all values of each attribute are generalized to the same level of the domain generalization hierarch (DGH). This paper also proposed greedy heuristics to generate k-anonymous solutions. [29] showed the generalization space maps to a partially ordered lattice and introduced a binary search method, which guarantees the solution is optimal according to a

certain cost metric.

By relaxing the constraint of mapping the entire domain to the same level of the DGH, [60] defined the generalization solution space as all arbitrary partitions on the ordered set of values in a single attribute's domain (e.g., age 14 is reported as 10-15 while age 16 is reported as 16-30). Given that the size of the search space is exponential in the size of a QI's domain, exhaustive search strategies are impractical. Thus, [60] proposed a genetic algorithm, to perform a partial search of the generalization space. This work is also notable because it represents QIs and their generalization as bit-strings, a model we adopt in this dissertation.

Since genetic algorithms do not provide a guarantee about the optimality of the solution and are often associated with long runtimes, [61] restructured the space of [60] to a tree and provided a systematic search algorithm using pruning and rearrangement to find an optimal k-anonymization in a practical amount of time. [62] further expanded the solution space to permit arbitrary partitioning of each attribute domain without forcing a total order on it. Based on this partitioning, they proposed a novel method to create a partition enumeration tree and search algorithms to efficiently discover the optimal anonymization solution. In all of the generalization strategies mentioned above, each attribute is generalized independently (i.e., a single-dimension attribute domain). Thus, any specific value in each domain is generalized in the same way in every tuple of the dataset.

[63] extended this model to a multi-dimensional space by generalizing values of tuples in the dataset. A greedy partitioning strategy was introduced to discover a k-anonymization solution. While flexible (e.g., females with age 14 are expressed as [female, 14] while males with age 14 are expressed as [male, 10-14]), the expansion of the search space provides more generalization options. As a result, the generalized dataset can be difficult to interpret because the same value of a QI attribute can be mapped to different values in the same generalized dataset.

In this dissertation, we use full-domain generalization and full-subtree generalization

and adopt the lattice structure to represent the generalization hierarchy. However, the optimization algorithms for k-anonymity given these generalization strategies can not be directly used to find a solution on the Risk-Utility (R-U) frontier, which is the set of policies for which there is no other policy with both better utility and less risk. The search for the k-anonymous solution which minimizes the total quantity of generalization is an NP-hard problem [64, 65]. Given the definition of k-anonymity, $1/k$ can be considered as an disclosure risk limit in a particular situation, in which the adversary has access to an external dataset which covers the exact same set of population as the de-identified dataset and the amount of generalization is a particular case for data utility metric. Thus, searching for an optimal generalization solution that satisfies k-anonymity can be considered as a special case of the broader problem of searching for generalization solutions that are on the risk-utility frontier given an arbitrary risk metric and utility metric. Thus, the frontier search problem is at least at the same complexity level as k-anonymization. As a result, we will need to develop heuristic search strategies to find the frontier solutions efficiently.

2.4 Identity Disclosure Control Using Risk Management

It has been suggested that assessing disclosure risk requires a holistic modeling of different types of adversaries [66, 67]. Such models should account for the motivation, means, opportunity cost, consequence of attempt, and likelihood of success. However, such investigations have not provided a formal approach to risk quantification that accounts for the elements in the data environment. Rather, existing disclosure risk measures mainly focus on the uniqueness of records in the dataset and in the population. For instance, three popular disclosure risk metrics (i.e., prosecutor, journalist and marketer) [68] assume that the adversary is always motivated to attack and the extra information required for re-identification is always available. As a consequence, the risk level is only dependent on the data itself.

The risk model we propose, on the other hand, explicitly formalizes elements beyond the de-identified dataset itself, which influence the adversary's decision making. We note

that there have been several investigations in applying game theoretic frameworks to analyze the adversary's best course of action and the corresponding disclosure risk [27, 69]. For instance, the adversary in [27, 70] is formalized as an opponent of the data publisher in a Stackelberg game. To maximize payout, the adversary decides if they should attack by comparing the potential gain against the cost of committing an attack. Yet, this model oversimplifies the adversary's decision process of gathering, linking, and exploiting data. Moreover, in their formalization, there were no explicitly modeled penalties for detecting the misuse of the data.

Moreover, the existing disclosure risk management methods did not give an frontier of solutions with a range of disclosure risk levels and data utility levels to guide the data holder's decision making.

2.5 Adversarial Modeling and MDPs

As mentioned earlier, we propose to build the adversarial model of the disclosure attack for risk evaluation based on the assumptions that 1) the adversary is rational and optimal planner and 2) the adversary makes a series of decisions in the attack process. In particular, we propose to use a Markov decision process (MDP) [71, 72] to represent the adversary's decision making process. An MDP is designed to model sequential decision making in which there is a reward associated with taking an action at each state, while the outcome of the action can be random. In technical terms, an MDP is a discrete time schochastic control process. At each step the decision maker is at a state, and the decision maker will make a decision to choose an action from a finite set of actions. There is an amount of reward granted to the decision maker given the action taken and the current state.

The next state a random state generated based on a probability distribution over a set of possible states given the current state and the action. In our model, the adversary is a decision maker, the state where the decision maker is at, is the state of the attack process (e.g,, the adversary has gained access to the external resource). At each step, the adversary

makes a decision on the next action. For example, whether or not to contact an individual that is linked to a tuple in the de-identified dataset. The adversary's action can bring some reward, which can be negative. For example, if the adversary's action is to access an external dataset, the value of the reward is the negative value of the cost of the external dataset. Moreover, the next state is not completely determined by the adversary's action. For example, when the adversary chooses to exploit an individual, he is uncertain about whether or not he will be detected and pulished. Also, before the adversary chooses to access the external dataset, he may not be certain about the number of identified individuals to which the de-identified record may be related. Therefore, our adversary's decision making naturally fits into the representation of the MDP framework.

To the best of our knowledge, MDPs have not been used to model adversaries in the privacy preserving data publication setting. Yet it has proven to be a useful tool in modeling adversary's optimal planning in security problems. This is because the MDP representation captures an adversary's uncertainty on the outcome of a security related action [73], similar to the disclosure adversary. An important difference between our adversarial model and the one in Letchford and Vorobeychik [73] is that, in the security scenario, the adversary terminates once he is caught, whereas in our model, the adversary may only pay a fine and continue to attack.

In particular, we propose to use a factored MDP to represent the adversary's decision making process. A factored MDP is an MDP model in which the state is represented by an assignment of the state variables. The computational challenge we face is how to solve the factored MDP (i.e., how to compute the optimal action at each state) [74]. Standard methods to solve a MDP include linear programming and dynamic programming. However, these methods have scalability issues, especially when dealing with a factored MDP because the state space of which grows exponentially with the number of state variables. Approximation algorithms have been designed to solve large scale MDPs including approximate dynamic programming [75, 76, 77, 78], neuro-dynamic programming [79] and

approximate linear programming [80, 81, 82, 83]. There also exist algorithms that leverage the structure of the MDP [84, 85, 86, 87, 88] to construct an approximate solution without enumerating the state space. For MDPs that can be decomposed to a series of weakly coupled tasks, algorithms have been designed to construct a global solution from the solutions of the small tasks [89, 90].

2.6 Game Model vs Multi-objective Optimization

Applying game theoretic models [27, 91, 70] to assess disclosure risk and generate optimal data sharing strategies for the data holder is an emerging research area. Wan [27] formalized the problem of releasing personal data in the form of a Stackelberg (leader-follower game) game, in which the data publisher is the leader and the data recipient (the potential adversary) is the follower. It is assumed that both the data publisher and the data recipient are economically driven with the goal of the maximizing the payoff. The data publisher's strategy set is the different ways of obfuscate the data, the adversary's strategy set is whether or not to re-identify a record. The data publisher's optimal strategy is obtained by solving the Nash equilibrium of the game.

The Risk-Utility frontier search method we introduce, which is essentially multi-objective optimization shares common roots with game theoretic model based methods. Yet there are significant differerences in terms of the goals, computational challenges and applications.

Both the game model and the frontier search algorithm consider the adversary as a rational and economically driven agent who makes decisions based on the payoff of each alternative action. Therefore, both of the two frameworks require the analysis of the adversary's gain from re-identifying a record, external resource cost, computational cost, and potential penalty cost from the authorities. The game model introduced in [27] did not emphasize constructing an elaborate adversary process model for identity disclosure attack, but instead, relied on an ultra simplified one-step decision model. We believe that the process-based adversarial model introduced in this dissertation can help advance privacy

game models.

Despite what is in common between the privacy game and the Risk-Utility frontier search, they are two irreplaceable approaches designed for finding the optimal data releasing strategy. In general, the solution to a game is a strategy that optimizes the goal of the game. Essentially it is a single objective optimization, while the solution to a multi-objective optimization is a Pareto frontier composed of the set of non-dominated strategies[2]. The implication is that, in order to adopt the game model based method, both the value of publishing the personal data and the loss brought about by a successful re-identification attack for the data publisher need to be measured in the same term, such as a monetary value.

This condition does not always hold true. For example, in some cases, the value of the data is measured by the utility of the data for certain type of scientific research, the value of which can not be measured in monetary terms, while the loss is a fine from the regulator for failing to protect subject's privacy. In this example, the game model is not applicable. On the other hand, the Risk-Utility frontier search method can provide guidance in making a decision on the data releasing strategy.

Another exclusive application for the Risk-Utility search method is when the data publisher can make a decision based on how the utility increases with the increase of the risk in a acceptable range. In other words, by using the Risk-Utility frontier search method, the data publisher is awarded the opportunity to trade a significant gain in data utility with a small loss in privacy (or vice versa).

2.7 The Economics of Identity Disclosure Attack

It is essential for the identity disclosure risk modeling proposed in this work to understand and analyze what types of exploitation an adversary can conduct and the amount of

[2]Given a strategy s, if there does not exist any other strategy s' is better than s in all dimensions, then s is a non-dominated strategy

gain can be obtained from such exploitations with the personal information inferred from the de-identified dataset (i.e., the economics of identity disclosure attack). However, there is only a limited amount of investigation into this issue.

Sweeney [57] enumerated a few entities that can potentially make a profit by exploiting personal health information [92]. For instance, a bank can make decisions on whether or not to call in outstanding loans based on whether or not the individual has a severe medical condition, such as cancer [93]. Alternatively, companies can make employment decisions about employees based on their medical records [94]. And insurers can sell personal health information to lenders, employers, or marketers [95]. However, the amount of profit gained from these exploits are difficult to measure without internal information from these entities.

In the meantime, there do exist some studies that can provide insights into the economics of an identity disclosure attack indirectly. For example, one study on the economics of financial and medical identity theft [96] looked at the financial flows and business models of the possible exploitation of personal medical data with a particular focus on medical identity theft. This study pointed out several ways to obtain a financial gain using personal health information, such as selling this information to an individual without insurance coverage or a wanted criminal who needs to obtain access to medical care, especially expensive procedures (e.g., organ transplant), or using this information to fake an identity to obtain prescription drugs and medical equipment to resell into black markets. Yet these types of exploits require personal health information that includes secret identity information which is normally not available in publicly available data resources that can be used in identity disclosure attacks against protected personal data, such as an SSN or an insurance ID. Thus, these exploits of personal health information are unlikely to happen with the information an adversary can gather by committing an identity disclosure attack against personal data published for secondary usage. However, the same methods that have been used in conducting medical identity theft study can be adopted to study the ways in which the information about an individual gathered from the identity disclosure attack can be exploited and the

subsequent potential financial gain from the exploitations.

An indirect way to estimate the financial gain from the exploitation of personal information discovered via an identity disclosure attack is the price of this information at which the subject or the owner of the information is willing to sell at. There exist studies on the value that individuals assign to their personal information, privacy market, and pricing personal information [97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108]. A study [97] through online experimens based on behavioral economics and decision research found that the value an individual assigns to her private information forms a non-normal distribution that is dependent on the context in which, and how, the transaction is made. In particular, the individual values the private information differently depending on, not only the endowment, but also the order in which different privacy options are presented. Moreover, the value of the private information is highly affected by factors that are not supposed to affect decision making.

With respect to privacy market, a personal identifiable information market is proposed in [105], in which the value of the personal identifiable information is decided by the auction between the information aggregators and the data subjects. Moreover, different private information markets have been recognized [103], such as a market in which data aggregators buy and sell data to other organizations and a market in which the data subjects trade their information for free services. If the market is known beforehand, the value of the information can be evaluated in the framework of the particular market. The economics of privacy studies have recognized the relationships between personal private information and the dynamic pricing (in other words, price discrimination) [109, 110]. Therefore, a possible way to evaluate the value of the personal information is to measure the profits brought up by the dynamic pricing strategy based on the consumers' private information.

Another alternative way to evaluate the adversary's gain from exploiting the information discovered from the identity is applying similar methods in measuring the cost of breaching a security system. Different security system offers a different cost to break

(CTB) [111]. Economic approaches have been proposed to measure the CTB of a system, such as offering a reward to the first exploitation of the vulnerability of the system and use the lower bound of the reward as the CTB [112].

2.8 Sampling and Prior Probability

An essential issue of assessing identity disclosure risk when releasing a de-identified dataset is to estimate that the prior probability the subject of the record in the dataset is also in the available external identifiable datasets. To address this issue, it is critical to estimate the population statistics from which the published personal dataset and the external dataset are sampled. Models have been proposed to estimate the number of population uniques (i.e., the number of people in the population with an unique value on the quasi-identifier) using sample data based on the assumption that the size of the equivalence group in the population is a realization of a superpopulation distribution, such as a Poisson-gamma model [113], Argus method [114], log-linear models [115, 116], neighborhood regression model [117], and a smoothing model using a local neighborhood [118]. These methods cannot be directly used to compute a prior probability, but the population statistics derived from a sample set based on these models can be adopted in further computing the prior probability.

Chapter 3

Theoretical Modeling of Re-identification Risk

3.1 Introduction

Many formal privacy protection models have been developed for publishing personal data. However, these formal models are based on simplistic adversarial frameworks, which can lead to both under- and over-protection. For instance, such models often assume that an adversary attacks a protected record exactly once. Moreover, these models protected against possible attacks, but it is unclear if they are probable. This is important because laws and regulations do not require perfect protection, but rather that data be shared in a manner that makes it difficult to ascertain an individual's identity. Organizations are thus afforded an opportunity to achieve data protection using risk management techniques. However, there does not exist a principled method to evaluate re-identification risk while accounting for the elements beyond the scope of the data itself that contribute to re-identification risks, such as deterrence mechanisms (e.g., data user agreements, the time and effort to gather the external in- formation necessary to compromise the data, or penalties for misusing data), that influence an adversary and the adversary's behavioral pattern. Thus, in this chapter, we introduce a principled approach to assess re-identification risk that that incorporates data- and penalty-based disincentives and the adversary's behavioral pattern.

3.2 Re-identification Risk Quantification Framework

Our framework quantifies the re-identification risk of publishing each record in a de-identified dataset. We assume the dataset is composed of person-level records in a relational form. We define re-identification risk as the composite of the probability that an adversary

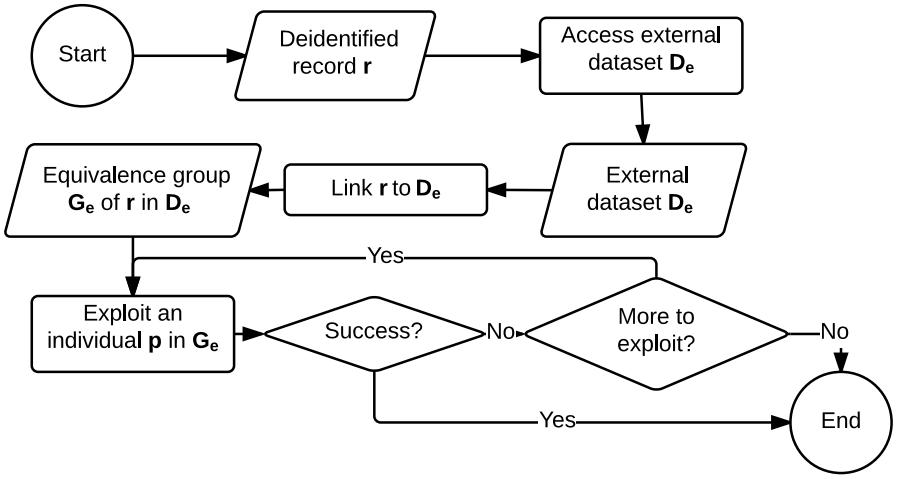


Figure 3.1: The re-identification attack process.

re-identifies a record and the harm it causes:

$$risk = P_{reid} \times L_{reid} \tag{3.1}$$

where P_{reid} is the re-identification probability and L_{reid} is the associated publisher loss. We assume L_{reid} is a predefined input, and focus on P_{reid}.

The re-identification probability is derived from the adversary's sequential decision process, outlined in Figure 3.1. The adversary begins with a de-identified record r. The adversary's first decision is to access an external table D_e or not. His second decision is whether to conduct a linkage attack, which yields an equivalence group of records G_e. This corresponds to the set of individuals with the same value as the target's published quasi-identifier (QI). At this point, each individual $\alpha \in G_e$ has a probability that they actually correspond to the targeted record r. This translates into a probability that an attack (e.g., confirmation of the patient's identity) on α will be successful. If the attack fails, the adversary can choose to exploit another individual from G_e. This process can repeat until the adversary decides to stop or he exhausts all of the records in G_e.

There are several notable aspects of this attack process. First, it should be recognized that this is a stochastic process. For example, the adversary may not know if the individual

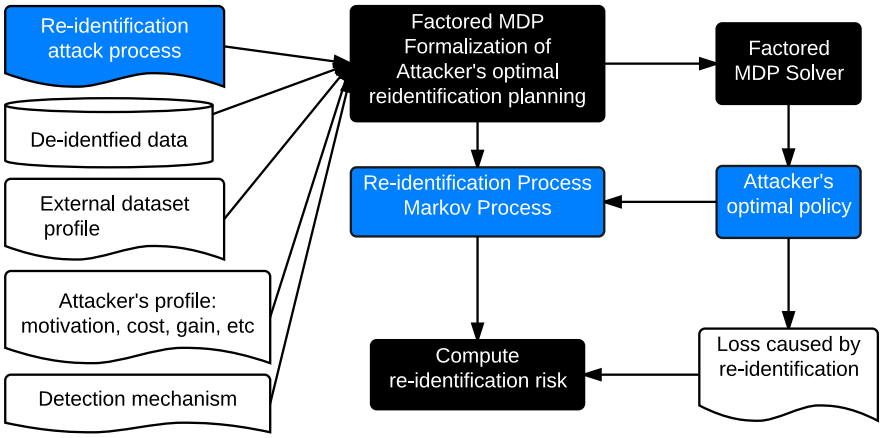


Figure 3.2: A general architecture of the re-identification risk quantification framework.

to whom the target record corresponds is in D_e. Therefore, the outcome of accessing the external dataset is uncertain. Furthermore, the result of exploiting an individual is stochastic, with outcomes ranging from success to failure to being detected and punished. A second notable aspect of the attack process is that there is a cost and a reward associated with each action, for instance, the reward for a success, the cost of accessing D_e and the penalty if an attack is detected.

More precisely, we model the adversary as a planner using a factored Markov decision process (FMDP) [85]. In a FMDP, a state of the world is characterized by a collection of random variables (or factors). The adversary is modeled as a rational agent computing an optimal policy; i.e., an optimal action to choose in each state of the FMDP. Given such a policy, we can compute risk according to Equation 3.1.

In Figure 3.2, we show the general architecture of the re-identification risk quantification framework. The framework is composed of three modules (the black rectangles in Figure 3.2): 1) the FMDP formalization of the adversary's decision process, 2) the FMDP solver, and 3) the re-identification risk computation module. The FMDP formalization module takes four inputs: i) the attack decision process, ii) the de-identified data, iii) the external dataset profile, and iv) the adversary's profile. The factored MDP model is then solved by the FMDP solver module to determine the adversary's optimal policy. Finally,

Table 3.1: The state variables of the FMDP model.

Variable	**Explanation**
X_t, binary	If **T**, attack is terminated
X_d, binary	If **T**, exploit of an individual is detected
X_p, integer	Number of previous exploits penalized
X_s, binary	If **T**, target record r is successfully re-identified
X_a, binary	If **T**, external dataset D_e has been accessed
X_l, binary	If **T**, target record r has been linked to the external dataset D_e
X_g, integer	The size of the equivalence group of target record r in external dataset D_e
X_r, integer	The remaining number of unexploited individuals in the equivalence group for record r in external dataset D_e

the risk computation module computes the quantified risk value given optimal attack policy and associated probability of successful re-identification attack. In the following sections, we dive into the details of each of the three modules.

3.3 Re-Identification as an FMDP

The FMDP model is a 4-tuple (X, A, R, P), where $X = \{X_0, ..., X_m\}$ is a finite set of random variables, each with a finite domain. In this model, A is a finite set of actions; R is the reward function $R(x, a)$, representing the reward for each action a taken in state $X = x$; and P is a Markovian transition function $P(X_i' | X_i^{parent}, a)$, which represents the probability distribution of the state variable X_i' in the next state given the value of a subset of state variables X_i^{parent} and action a (X_i^{parent} is the set of variables that X_i' is dependent on given the action is a). We denote the value of a state variable X_i in state x as $x[X_i]$. We assume that the FMDP has an infinite horizon, and time is exponentially discounted with a discount factor γ.

State variables

As summarized in Table 3.1, the FMDP model is based on eight state variables. Here, we take a moment to provide intuition into each of these variables. First, X_t is a binary variable that represents the termination of an attack. When $X_t = T$ (true), the corresponding state is absorbing, effectively ending the decision process. Next, we assume the existence

Table 3.2: The actions of the FMDP model.

Action	**Explanation**
$terminate$	Abort the attack
$access$	Access the external dataset D_e
$link$	Link the de-identified record r_i to the external dataset D_e
$exploit$	Exploit a random individual in the equivalence group of record r in the external dataset D_e

of an attack detection mechanism, and the state of detection is indicated by a binary variable X_d. The following variable, X_s, indicates whether the exploit is successful (in which case the adversary obtains a positive reward). The next two variables are associated with data manipulation. X_a is a binary indicator of whether the external dataset D_e has been accessed, while X_l is a binary indicator of whether it has been linked to the published target record r. X_p maintains the number of times the exploitation has been detected and penalized. The final two variables, X_g and X_r keep track of the size of the equivalence group and the remaining unexploited individuals in the group. Thus, as the adversary attempts (unsuccessful) attacks on matched records, X_r decreases while X_g remains constant. This is because the original group size associated with linking is fixed. To keep our presentation compact, we represent each state x as a vector $[x_0,...,x_m]$ in the FMDP model, where x_i denotes the value of the i^{th} variable in the list $[X_t, X_d, X_p, X_s, X_a, X_l, X_g, X_r]$.

Action set

There are four classes of actions in our system, which are summarized in Table 3.2. The adversary has the option of aborting the attack at any time by choosing the $terminate$ action. The other three actions represent the adversary's operation in three different phases of the attack. The *access* action represents the accessing of the external dataset D_e. The $link$ action represents the linking of the de-identified record r to the external dataset D_e. The $exploit$ action represents a potentially harmful exploitation of an individual that is deemed to be related to the record r under attack. The particular type of exploitation may differ under various circumstances. For example, if the adversary's goal is to demonstrate the vulnerability of the system, the exploit may be to prove they can contact the individual and

confirm the record is really associated with them [119]. Or, the adversary's goal may be to conduct direct marketing to the individual based on the sensitive information in the record (e.g., for a particular pharmaceutical). Regardless, an exploit is assumed to be successful if it is conducted against the individual to whom the record corresponds.

Reward

Reward functions are determined by several factors: the cost of taking an action, the loss to the adversary from detection (both negative rewards), and the gain from a successful attack. We formally define the reward function as:

$$R(x,a) = R_g(x[X_s], x[X_t]) + R_p(x[X_d], x[X_p]) - C_a \tag{3.2}$$

where C_a is the cost of action a (denoted by C_d, C_c, and C_e for *access*, *link*, and *exploit* actions, respectively). $C_a = 0$ for the *terminate* action. $R_g(x[X_s], x[X_t])$ represents the gain from a successful exploitation. $R_g(x[X_s], x[X_t]) = G$, if $X_s = T$ and $X_t = F$ and 0 otherwise.

We assume there is a maximum number of times, n_f, that the adversary will be subject to a penalty (e.g., a fine for law or contract violation) if he is detected. Note that this permits an analysis on the special case of $n_f = 1$, where the adversary is only penalized once. This is notable because it represents the real scenario where a data user is penalized for violating a contract, but is not prevented from continuing to exploit the data they have already received. We denote the cost related to the fine as $R_p(x[X_d], x[X_p])$. $R_p(x[X_d], x[X_p]) = -C_p$, if $X_d = T \wedge X_p < n_f$ and 0 otherwise.

State transition dynamics

We use a dynamic Bayesian network (DBN) $\tau_a = \langle G_a, P_a \rangle$ for each action a (except action *terminate*), as shown in Figure 3.3, to represent the transition function $P(X_i | X_i^{parent}, a)$. We denote the current state and the next state as x and x', respectively. If the action is to *terminate*, $x'[X_t] = T$.

If the action is to access, as the DBN shows in Figure 3.3(a), there are 3 state variables

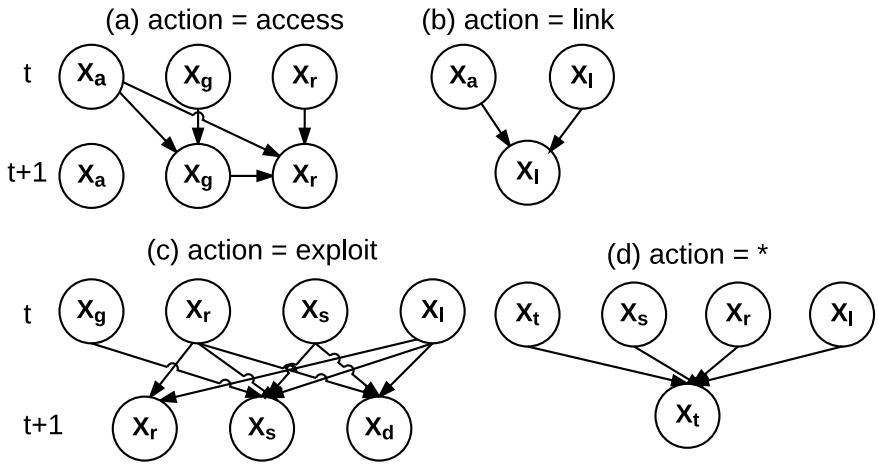


Figure 3.3: The dynamic Bayesian network (DBN) for each action of our FMDP model.

that may change in the following step: X_a, X_g and X_r. We highlight that if the external dataset D_e has not yet been accessed (i.e., $x[X_a] = F$), the adversary's belief of the equivalence group size in the next state $x'[X_g]$ is a probability distribution over a set of values, represented as $P(G_{r,D_e})$. Our experiments simulate $P(G_{r,D_e})$ under different levels of certainty and its influence on the adversary's behavior and re-identification risk.

In Figure 3.3(b), the *link* action sets $x'[X_l] = T$ when $x[X_a] = T$ (i.e., external dataset is available for linkage).

The prerequisite condition for the *exploit* action is $x[X_l] = T$, $x[X_s] = F$, and $x[X_r] > 0$. In other words, we can only exploit a record if 1) the equivalence group is non-empty, 2) the dataset has been linked to the record, and 3) the record has not already been re-identified. In this case, the number of remaining candidates in the equivalence group is decremented ($x'[X_r] = x[X_r] - 1$).

Moreover, the probability that the exploited individual is associated with the record is the probability of selecting an individual at random from the set of individuals in the population (with the same quasi-identifier) who have not been exploited. The number of individuals with the same quasi-identifier in the population who have not been exploited is the sum of the number of individuals outside (i.e., $\frac{1-prior_{r,D_e}}{prior_{r,D_e}} \times x[X_g]$) and inside (i.e.,$x[X_r]$) the external dataset d_e. Thus, the success probability of an exploitation can be formally

represented as:

$$\begin{aligned} &P_{suc}(x[X_g], x[X_r], prior_{r,D_e}) \\ &= \left(\frac{1 - prior_{r,D_e}}{prior_{r,D_e}} \times x[X_g] + x[X_r]\right)^{-1} \end{aligned} \tag{3.3}$$

where $prior_{r,D_e}$ is the probability that the individual corresponding to the data is in the external dataset D_e.

$P(x'[X_d] = T)$ (i.e., the probability of being caught) denoted as P_{det} can be modeled in a number of ways. Since the probability an exploit is detected is very likely to increase with repeated attempts due to various factors (e.g., increased vigilance), we model the detection probability using a sigmoid function:

$$P_{det} = (1 + e^{-(h_0 + h_1 \times (x[X_g] - x[X_r]))})^{-1} \tag{3.4}$$

where $x[X_g] - x[X_r]$ corresponds to the number of exploit attempts the adversary has committed against records in the equivalence group. Note that this formulation allows us to model the special case, where the probability of detection does not increase over time by setting $h_1 = 0$.

Finally, regardless of the action, the transition of variable X_t is determined as follows (see Figure 3.3(d)): $x'[X_t] = T$ if $x[X_t] = T \vee x[X_s] = T \vee (x[X_l] = T \wedge x[X_r] = 0)$.

3.4 Algorithms

Solving the MDP

Solving an infinite-horizon discounted MDP amounts to computing an optimal policy, $\pi(x)$, which prescribes an optimal action to take in each state [74]. Equivalently, it suffices to compute a value function, $V(x)$, which is the optimal discounted sum of rewards of an optimal policy.

A number of methods exist for solving an MDP. Linear programming (LP) is one such

method, which computes the value function, $V(x)$, for every state x. An important limitation of the standard methods, including LP, is scalability. In particular, if we do not take advantage of problem structure, the runtime is polynomial in the number of states, which itself grows exponentially in the number of state variables. Approaches exist that leverage the structure of the factored MDP, but they are approximate, and require the pre-specification of a fixed set of basis functions over the state space. Next, we present a special-purpose method, which we call Two-Level LP, that takes advantage of our problem structure (including the factored state) and reports an exact answer.

Two-level Linear Programming

We designed the Two-level LP algorithm under the principle of removing all the "well-known" parts from the FMDP structure to save space and runtime. The algorithm constructs a two-level structure from the state space. The states in the FMDP model form a *sink cluster* sub-structure, which satisfies the following properties: a) there is no outbound and b) there is only one inbound state (i.e., x_{start} has only one inbound edge). Based on the property of the FMDP, each sink cluster can be solved independently. The bottom-level of the Two-Level LP algorithm solves a LP and stores the value of the state x_{start} for each sink cluster. The top-level algorithm then constructs and solves a LP of the entire state space by replacing each sink state with its corresponding x_{start} and assigns the pre-computed $V(x_{start})$ to it.

Specifically, each sink cluster contains the descendant states of a state x_{start} in which the adversary has taken the action of access and link, but has not yet started exploitation, i.e., $x_{start} = [F,F,0,F,T,T,s_i,s_i]$, $s_i \in (0, \max(G_{r,D_e}))$. Given two different group sizes s_1 and s_2, the two sink clusters with $x_{start} = [F,F,0,F,T,T,s_1,s_1]$ and $x_{start} = [F,F,0,F,T,T,s_2,s_2]$ do not overlap because the value X_g remains constant when $X_a = T$ and $X_l = T$. The resulting values of all the x_{start} states are used in the top-level LP to solve the values for the remaining states, such as the state in which the adversary is attempting to access the external dataset (i.e., $x = [F,F,0,F,F,F,0,0)]$).

We make two performance improvements for Two-Level LP. First, we introduce a pruning strategy which leverages the fact that the value of the starting states for each cluster (i.e., $V(x_{start} = [F,F,0,F,T,T,s,s])$) decreases as the size of the equivalence group $X_g = s$ increases. We omit the proof of this property due to brevity.

Thus, we sort the sink cluster by the value of $x_{start}[X_g]$ in ascending order. Specifically, if $V(x_{start}) = 0$ given $x_{start}[X_g] = s$, then all of the sink clusters with $x_{start}[X_g] > s$ will be pruned. Second, we use a result caching strategy. In doing so, the result from the bottom-level LP is cached and reused with multiple records. This happens when there is an overlap in the adversary's belief of the probability distribution interval of the equivalence group size X_g.

Computing Re-identification Probability

The re-identification probability P_{reid} is the sum of the probability of reaching each of the states with $x[X_s] = T$ and $x[X_t] = F$ in 1 to t_{max} time steps. Formally, P_{reid} is computed as:

$$P_{reid} = \sum_{t=0}^{t=t_{max}} \sum_{x \in x_{suc}} M^t[x_0, x] \tag{3.5}$$

In equation 3.5, $x_0 = [F,F,0,F,F,F,0,0]$ represents the state in which the adversary has not accessed the external dataset yet, x_{suc} is the set of states with $x[X_s] = T$ and $x[X_t] = F$, x is an arbitrary state. M is the state transition $N \times N$ matrix of a Markov chain, where N is the number of states.

The state transition matrix M is obtained by replacing the action a in the transition dynamics function of the FMDP with $policy(x)$, i.e., $P(X_i'|X_i^{parent}, policy(x))$. However, there is one exception. Given the current state is x_0, in the FMDP model, $x'[X_g]$ is a probability distribution over a range of values due to the uncertainty of the adversary's belief, while, in the risk computation Markov chain, $P(x'[X_g] = g_{r,D_e}) = 1$, g_{r,D_e} is the actual group size in D_e. This is because the Markov chain already embeds the adversary's optimal policy, and consequently the adversary's belief in the group size no longer matters. Instead, what matters is the actual group size. We assume that g_{r,D_e} is an input to the risk framework.

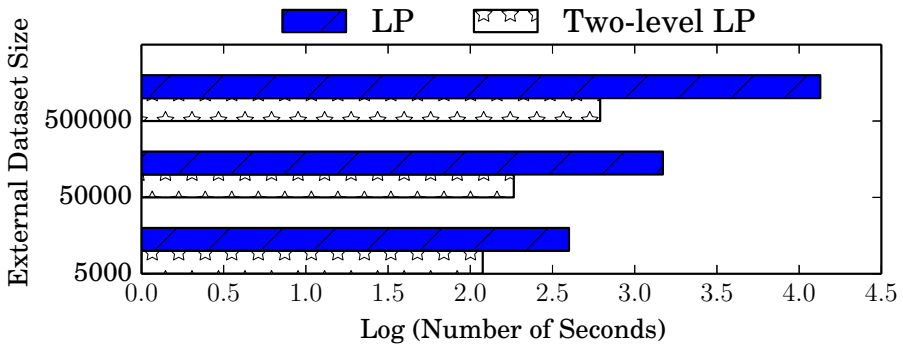


Figure 3.4: Runtime (log_{10}) of the FMDP solving algorithms for a dataset of 5000 de-identified records.

The value t_{max} is the maximum number of time steps it takes for all the states to transit into a state where $X_t = T$ (i.e., an absorbing sink state). Formally: $\exists t_{max} > 0 \ \forall x_i, x_j \in x_i, i \in [0,N], \text{if } x_j[X_t] \neq 0, M^{t_{max}}[x_i, x_j] = 0$.

3.5 Experiments

Dataset

Our experiments make use of three resources. First, we use the freely available North Carolina voter registration (NCVR) list [120] as the identified external dataset. This dataset consists of 6,018,999 records without missing values over 18 fields. These records include explicit identifiers (e.g., personal name and phone number), as well as quasi-identifiers (e.g., age, gender, race, and ethnic group). For the purposes of this study, we restricted the dataset to a set of four quasi-identifying attributes, {*Age*, *Race*, *Gender*, *5-Digit ZIP Code*}.

Second, we use the Adult dataset from the UCI Machine Learning Repository, as the de-identified dataset. This dataset consists of 32,561 records with 14 fields each, based on a sample of the U.S. Census, without missing values. This dataset contains *Age*, *Race*, and *Gender*, but not *5-Digit ZIP Code*. As such, for each record in the Adult dataset, we synthesize and append a 5-digit NC ZIP code based on the population distribution in the

US Census Bureau's 2010 Census Tables PCT12A-G. We also replaced a topcoded age value $[90+]$ by a random value in the range of $[90, 120]$.

Third, we assume that both the de-identified and identified datasets are sampled from the entire population of NC. In this case, it should be noted that the total size of the NC population, according to the census is 9,553,967.

Equivalence Group Size Distribution

In the experiments, the probability distribution of the value X_g after the adversary takes the action to access the dataset $P(G_{r,D_e})$ is derived from the adversary's knowledge about the external dataset D_e or the population statistics. Here, we consider two scenarios. In the first scenario the adversary knows the target's equivalence group size when starting the attack. Specifically, $P(G_{r,D_e} = g_{r,D_e}) = 1$. We refer to this scenario as the *known group* model.

However, the adversary may not have such knowledge before accessing D_e. In this case, we assume the adversary knows only the total size of the external dataset, n, and the probability density of the target's record, i.e., the joint probability of the target's quasi-identifying values $P(r[QI])$, in the population. Assuming that the external dataset is sampled uniformly at random from the population, $P(G_{r,D_e})$ can be represented as a binomial distribution defined in Equation 3.6:

$$P(G_{r,D_e} = k) = B(k, n, P(r[QI])) \tag{3.6}$$

We refer to this mechanism as the *unknown group* model.

3.6 Results

Performance Analysis

We evaluated the runtime of the framework with 5000 randomly selected Adult records. In this analysis, we consider the *unknown group* scenario with a $P(G_{r,D_e})$ computed as Equation 3.6 in which the size of the external dataset is set to 3 different values: 5K,

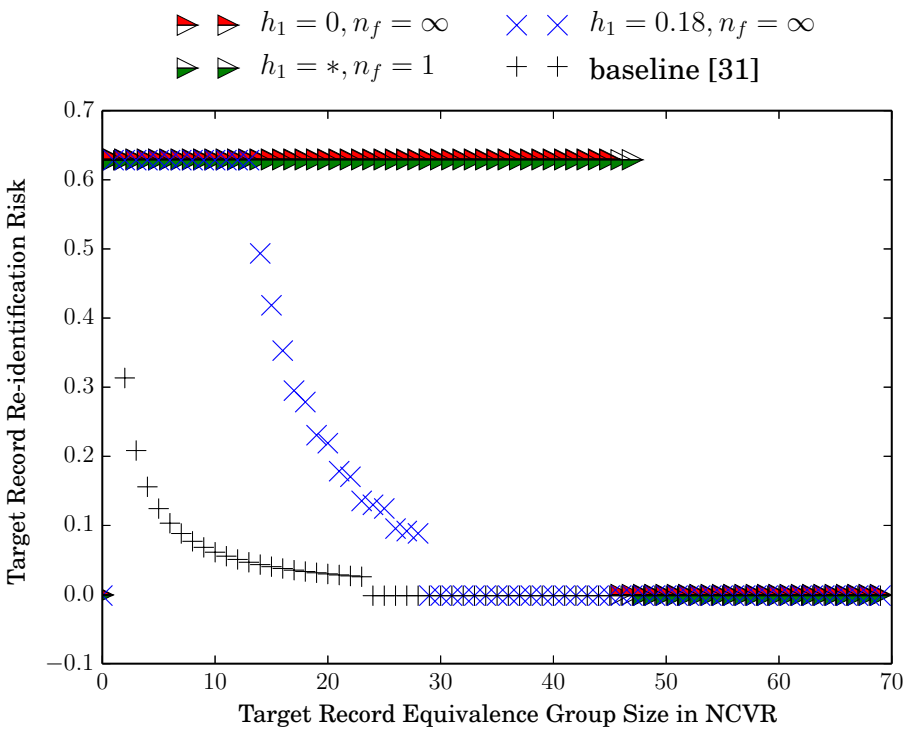


Figure 3.5: The equivalence group size for the target record in the NCVR dataset and the re-identification risk under the *known group scenario*.

50K, and 500K. We present the result of the *unknown group* scenario because the *known group* scenario yield FMDP with constant size state space, while *unknown group* scenario leads to increasing state space when the external dataset size n increases simply because of the interval of $P(G_{r,D_e})$ increase with n. The detection and penalty mechanism is set to $h_0 = -4.59$, $h_1 = *, n_f = 1$ (i.e., penalize only once and the probability of detection is 0.01 based on equation 3.4). The other parameters of the model were set to $prior_{r,D_e} = 0.63$, $C_d = 100$, $C_e = 10$, $G = 8000$ and $C_p = 10000$.

The algorithms were implemented in Python and all experiments were run on an Ubuntu server with 24 Intel(R) Xeon(R) CPUs at 2.4 GHz and 64 GB of RAM. The LP solver was implemented in the IBM ILOG CPLEX optimizer.

Figure 3.4 reports the runtime for the LP and Two-level LP algorithms. It can be seen that, as expected, the Two-level LP is always faster than the standard LP algorithm. The

difference in speed is accentuated as the size of the external dataset grows. By the time there are 500K records in the external dataset, the runtime of the Two-level LP is approximately 21x faster (616 seconds vs. 13,444 seconds).

Case study

To perform a case study, we assume the Adult and NCVR records are random samples of the NC population. Thus, the prior probability that the individual corresponding to an Adult record is in the NCVR is the sample ratio, or $prior_{r,D_e} = 6,018,999/9,553,967 = 0.63$. The NCVR data is free; however, considering the effort to obtain it, we set the cost of accessing the external dataset C_d to \$100.

The cost to exploit, gain and penalty values were set to $C_e = \$10$, $G = \$8000$ and $C_p = \$10000$ for each record, respectively. We acknowledge these values may vary in practice. The goal is to simulate a case in which the adversary will attack at least a subset of the records. This allows us to examine how different deterrence mechanisms and uncertainty about the equivalence in the external datasets affects the adversary's behavior and the re-identification risk. Therefore, these parameters are selected from a range in which the adversary will attack some of the records.

Known Group Model

We compare the *known group* model to the risk model in [27]. This is the only available model for re-identification based on an adversary's optimal decision. In the baseline, the adversary makes a single decision on when to attack based on the total payoff:

$$Payoff_{baseline} = G * (\frac{prior_{r,D_e}}{G_{r,D_e}}) - p_{det} * C_p - C_d - C_l - C_e \tag{3.7}$$

If $Payoff_{baseline} > 0$, the adversary exploits a random individual and the risk of re-identification is $prior_{r,D_e}/G_{r,D_e}$. Otherwise, the risk is 0. The FMDP is configured under three detection and penalty settings: a) a constant detection probability with repeated penalties (i.e., $h_1 = 0$ and $n_f = \infty$), b) a one-time penalty (i.e., $h_1 = *$ and $n_f = 1$)[1] and c) an increasing rate of

[1]The $*$ indicates that h_1 can be anything because only a single penalty is assigned.

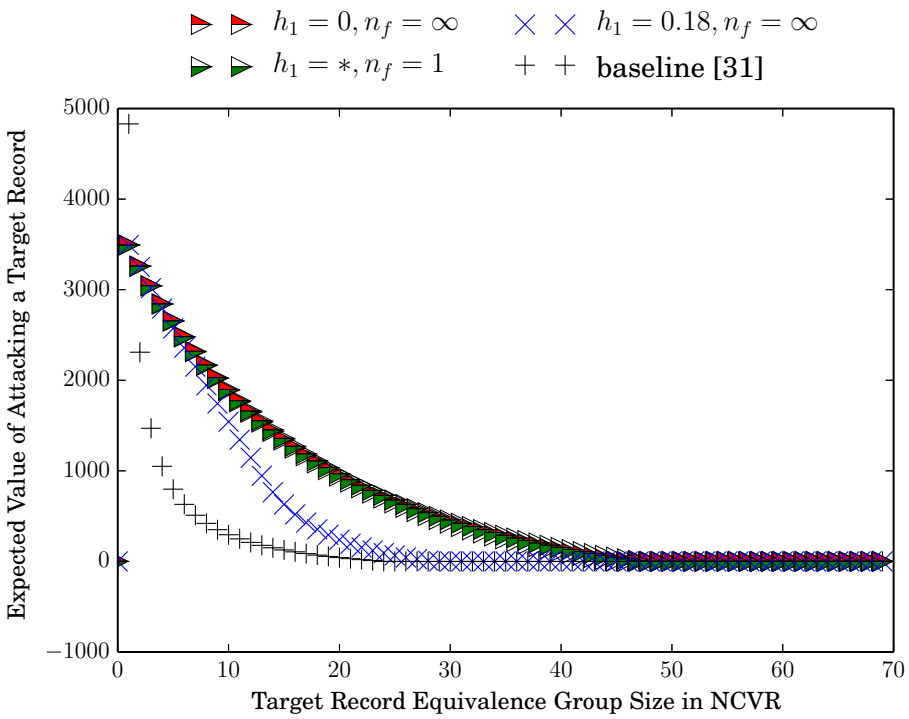


Figure 3.6: The equivalence group size of of the target record in the NCVR dataset and the adversary's expected payoff under the *known group scenario*.

detection with repeated penalties (i.e., $h_1 = 0.18$ and $n_f = \infty$). In each setting, we set $h_0 = -4.59$. This yields a 0.01 detection rate for the first exploit, an increase to 0.012 for the next exploit, and so on.

The results are illustrated in Figure 3.5. There are three notable findings to highlight.

Finding 1: The baseline risk never exceeds the FMDP models. This is because the baseline assumes that the adversary can only select one random individual, which is suboptimal. Thus, as can be seen in Figure 3.6, the baseline adversary's expected value drops at a faster rate than the adversary who acts according to the FMDP. Moreover, the adversary's success rate is also lower for the baseline. This is because the adversary only exploits one random individual from the equivalence group. This indicates that the baseline model often underestimates the re-identification risk.

Finding 2: When the detection probability is constant (i.e., $h_1 = 0$) and there is no upper bound on the number of times a penalty is levied on the adversary (i.e., $n_f = \infty$),

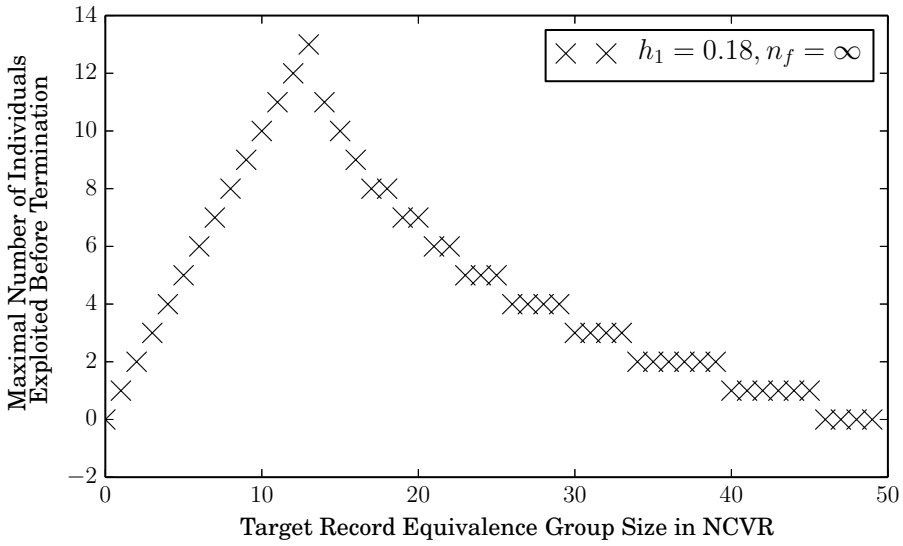


Figure 3.7: The size of the equivalence group of the target record in the NCVR dataset and the actual number of individuals the adversary exploits before terminating under the *known group* scenario.

the adversary either exploits 1) all records in the equivalence group or 2) no records.

Finding 3: When the probability of detection grows with repeated attempts (i.e., $h_1 > 0$) or there is an upper bound on the number of times a penalty is levied on the adversary (i.e., n_f is a finite value), the adversary exploits a subset of the equivalence group. In the scenario represented by Finding 2, the adversary chooses not to issue an attack when the equivalence group size is $\geq$ a threshold k, but the adversary exploits all the individuals in the equivalence group otherwise. Thus, the re-identification risk is either equal to the prior probability $prior_{r,D_e}$ or 0. This is because when the optimal action is to attack one individual in the NCVR equivalence group the subsequent optimal action is always to continue to exploit each of the remaining individuals provided that each exploitation has the same probability of being detected and the adversary will always be fined if detected.

In the scenario of Finding 3, the adversary may terminate the attack before exhausting the candidates in the equivalence group. Thus, the risk can be any value between 0 and the prior probability $prior_{r,D_e}$. This is due to two possible reasons. First, if $h_1 > 0$, both the

likelihood of detection and a successful re-identification are increasing when more individuals are exploited. Thus, the adversary stops when the increment in the expected penalty exceeds the increment in the expected payout, which can happen before the adversary exhausts all the candidates. Second, if n_f is finite, and the adversary was not detected in the previous exploitations, the expected payout can decrease when the number of the remaining cadidates reduces.

Similar to Finding 2, if the group size is $< k$, the adversary exploits all the individuals in the equivalence group. By contrast, if the group size is $\geq k'$, the adversary stops issuing an attack. For the group size in the range of (k, k'), the adversary's optimal action is to stop before reaching the last candidate in the group. The actual number of candidates the adversary will exploit before termination is shown in Figure 3.7. In this case, $k = 14$ and $k' = 29$.

These two findings are contradictory to what is expected by the baseline model. In particular, the records with equivalence group size $< k$ all have the same level of risk according to the FMDP model, while the records with smaller equivalence groups have more risk than those with larger equivalence groups according to the baseline model. The indication of this finding from the data protection perspective is that applying mechanisms, such as generalization, to increase the equivalence group size can only effectively reduce risk if the equivalence group size $\geq k$. In other words, increasing the equivalence group size to any value $< k$ will only harm the utility of the data without reducing the risk.

Unknown group Model

The FMDP enables us to evaluate risk when the adversary is uncertain in the equivalence group size; i.e., the *unknown group* scenario. We assume that the adversary's belief of the group size is as in equation 3.6 with $n = 6018999$, with $p_{r[QI]}$ equal to the probability density of the corresponding target record in the NC census population. The other parameters are the same as defined in the *known group* model. Our result illustrates the following findings.

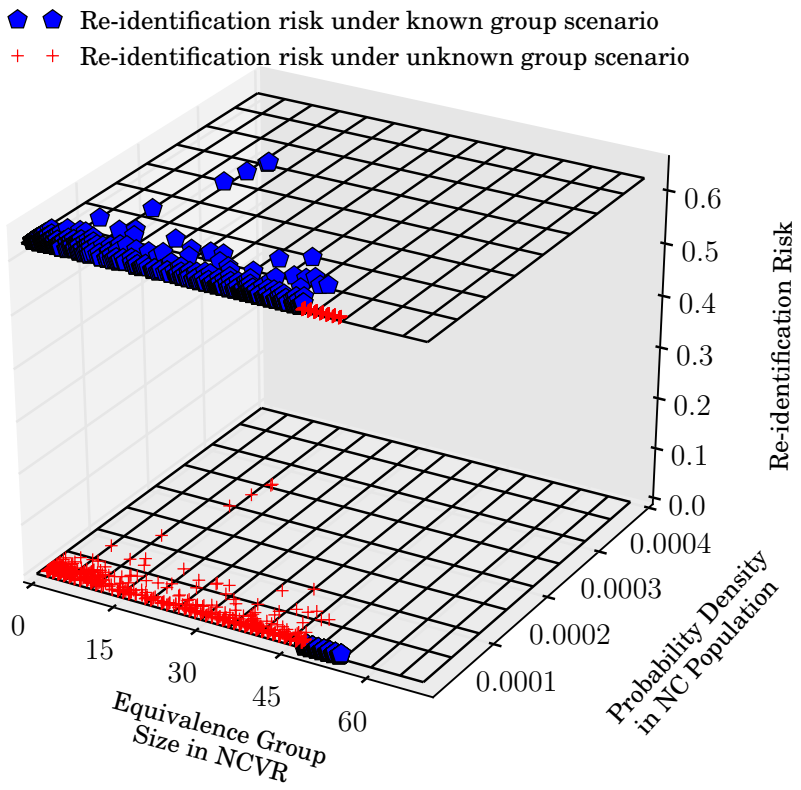


Figure 3.8: The equivalence group size, population probability density and the re-identification risk of the record with inconsistent risk values in the *known and unknown group* scenarios ($n_f = 1$).

Finding 4: The unknown group scenario can yield *lower risk* than the known group scenario.

Finding 5: The unknown group scenario can yield *higher risk* than the known group scenario. These findings illustrate that uncertainty in the group size can change the action of the adversary. To make this observation more concrete, Figure 3.8 depicts the risk for the 1920 records that have exhibited different risk scores. 1118 of these records (or 58%) have a risk of 0.63 under the *known group* scenario and 0 under the *unknown group* scenario. The remaining 802 (or 42%) records have the exact opposite result. The former is due to the fact that the adversary underestimates the payoff by using the probability distribution of the equivalence group size. As a result, the adversary does not access D_e, while the actual group size is $<$ the threshold $k = 48$ and in the *known group* scenario the adversary will access D_e and attack. The latter is, on the other hand, due to adversary's overestimation of the expected payoff based on their inaccurate belief about the equivalence group size. These cases are counterintuitive because one may argue that even if the adversary decides to access D_e, he or she will not exploit and there is no risk because the actual equivalence group size is $\geq k = 48$. However, this is not always true because after the adversary obtains D_e, the cost C_d (i.e., the cost of accessing the external dataset) became a sunk cost. As a consequence, the payoff is computed without considering C_d and the threshold the adversary can tolerate increases from 48 to 51. If the actual equivalence group size is between the two thresholds, the adversary with less knowledge (i.e., in the *unknown group* scenario) may be able to cause greater risk, even though the adversary does not necessarily obtain a higher payoff than the *known group* adversary.

Records resulting in different risk levels in the *known* and *unknown group* scenarios are not very common in this experiment setting. A majority of the records lead to the same risk (94%, or 30641 in total). The is due primarily to the fact that this analysis is dominated by records whose corresponding equivalence group size is larger than 55. Specifically, 64%, or 20891 in total, satisfy this situation. This is notable because, even if a positive payoff

expected from $P(G_{r,D_e})$ leads the *unknown group* adversary to access the external dataset, the adversary never chooses to exploit such records, yielding a risk of 0.

Sensitivity Analysis

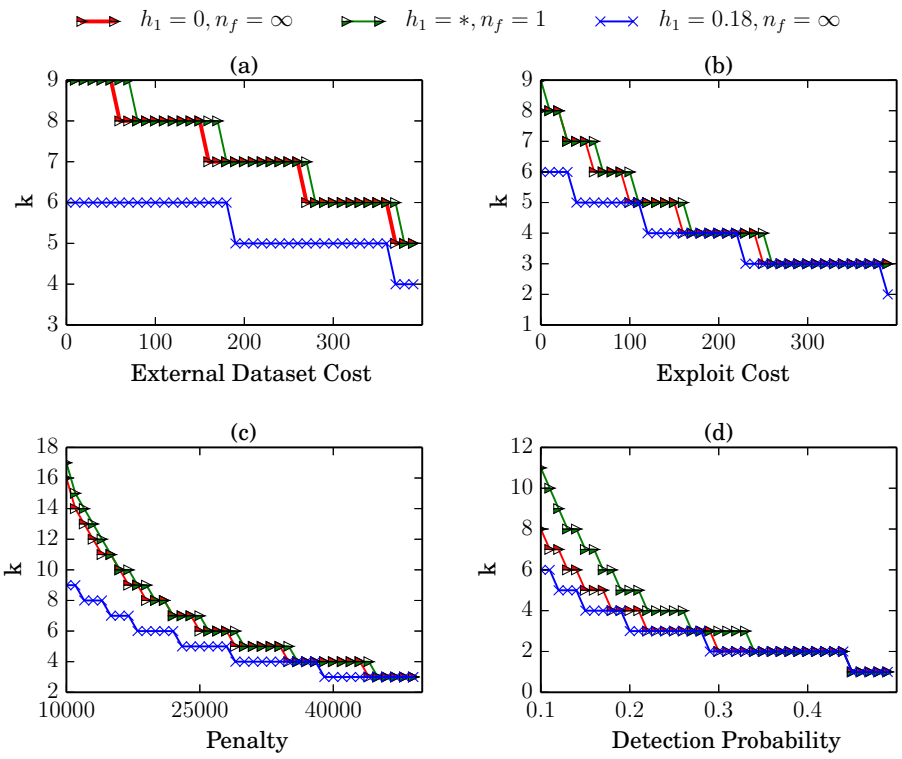


Figure 3.9: Sensitivity analysis on group size threshold (k) as a function of (a) external dataset cost C_a; (b) exploit cost C_e; (c) *Penalty*; and (d) detection probability P_{det}.

In this section, we investigate how the deterrence parameters influence the threshold k of the size of the equivalence group when the adversary walks away, such that the risk is 0 when equivalence group size is $\geq k$ under the three scenarios studied above. For this analysis, we assume a *known group* scenario and both the de-identified and external datasets cover the entire population, such that $prior_{r,D_e} = 1$. We vary 1) the cost to access data, 2) the cost to exploit the targeted individual, 3) the penalty levied when re-identification attempts are detected, and 4) the detection probability. In the analysis, we vary one variable at a time while holding all other variables constant to: $C_p = \$20000$, $G = \$1000$, $C_e = \$10$, $C_l = \$0$, $C_a = \$100$.

The result is unsurprising, but notable. Specifically, as illustrated in Figure 3.9, as

the deterrence mechanism is ramped up, the expected payout is lower and the adversary tolerates less risk. For example, when the penalty is set to $10,000, the adversary always attacks when the group size is smaller than 9 individuals. By the time the penalty is raised to $50000, the adversary will only risk an attack if there is one individual in the group. This result clearly indicates that penalties and costs for access to data can quickly deter an adversary from committing an attack.

3.7 Discussion and Conclusions

This research provides a formal process-based approach to characterize the privacy risks for published data and opens a novel direction in the field of data privacy. It also introduces a scalable algorithm based on linear programming to solve the attacker's optimal planning problem. A core contribution of this approach is that it accounts for deterrence mechanisms beyond data manipulation methods. We demonstrated the feasibility through a case study in a real world scenario, where an adversary uses a publicly available population registry (with over 6,000,000 individuals) to attack a record subject to a data obfuscation mechanism.

Our results reveal that a broadly accepted adversarial model in which the adversary will randomly choose one individual that matches the record to attack can be suboptimal, and an adversary may try and exploit every individual in the corresponding equivalence class. In addition to penalization mechanism, our result demonstrated that the adversary's optimal decision depends on the information about the external resources they may use (e.g., voter registration lists) before they access them to mount an attack. This work provides strong evidence that the risk to such systems in the real world is heavily dependent on the amount of effort an adversary needs to exert and the expected payout they can receive based on their attack. This investigation further provides intuition into how data perturbation techniques can be complemented by alternative disincentive strategies (e.g., charging for access to data or levying fines for malicious behavior) to lower the risk inherent in data sharing.

Our approach has several limitations which can provide directions for future research in

this area. First, if such a risk estimation procedure is to be put into practice, policy makers will need information about the nature of deterrence mechanisms, the existence and costs of external data resources, as well as the adversary's potential gain. Moreover, our work shows that knowing the prior probability that the corresponding target is in an external resource is critical to the model. Our model assumes that the external dataset is a random sample from a large population that also covers the protected data. Such information is not always readily available to the data publisher when evaluating risk. In the event the publisher believes they could underestimate such parameters, they may lobby for larger fines on misuse, thus deterring users with legitimate interests from accessing their resource. Thus, a future direction for research is in the development of approaches to estimate such parameters of the attack process. This may be possible, for example, by building a model for the detection rate based on existing detection mechanisms.

Second, there are limitations in the scope of the adversary's goals. Consider, the process model assumed an adversary targets only one record in the protected dataset at a time. It also assumes that the adversary has access to only one external resource to mount an attack. Perhaps more significantly, we assume that the success of an exploit will be confirmed. Yet, certain adversaries may be interested in multiple records in the protected data (or even the entire dataset) and may have access to multiple resources. Removing any of these assumptions will lead to an increase in the complexity of the adversary's decision problem. We note that the process model can be extended to account for these scenarios by introducing more state variables and actions. However, this will lead to an explosion in the state space. Therefore, a future direction of research is to generalize the FMDP model while improving the scalability of the solver algorithm.

Finally, our empirical analysis was conducted on a specific type of data, namely the demographic information within the publicly available population registry. Such a process-based approach to privacy risk assessments is applicable to other types of data where the attack is not a linkage-based exploit, but focuses rather on other definitions of privacy, such

as inferential disclosure. The adaptation of such a technique will depend on the extent to which the adversary's process for realizing their exploit can be represented.

Chapter 4

A Feasibility Assessment for Temporal Penalties in Data Sharing

4.1 Introduction

Sharing de-identified personal data can cause potential harm to the sharer and the data subjects if the data is not handled properly (e.g., being used in a way that is not supposed to or being re-identified), therefore oftentimes the data users need to sign a data use certification (DUC) or data use agreement (DUA) with the data publisher before the data access can be authorized. A DUC or DUA usually includes a term that specifies that the data user should not try to re-identify any data subject. In order to future prevent privacy and other violations from happening, certain penalties need to be enforced on the data users that violate the terms in the DUC or DUA.

Some of these penalties have a straightforward quantifiable consequence for the data users. For instance, a fine will cause immediate monetary loss to the data user. Other penalties, however, have a less quantifiable impact on the data user. A special case of the latter is the temporal penalties that have been adopted by the NIH genomic data sharing platform dbGaP and the Wellcome Trust Case Control Consortium (WTCCC) [53]. The temporal penalties basically mean that the user will be suspended from accessing the database, conducting research, publishing papers or writing grant proposals using the data from the database for a period of time. The length of the period can be influenced by the types or severities of the particular violation. The assumption behind the temporal penalties is that the value of the data for academic research decreases over time and being revoked access will have a negative impact on the user's grant funding applications in the future. However, whether or not the value of the data for academic research actually declines over time has not been investigated. To the best of our knowledge, this dissertation is the first

one to look into this problem.

To ascertain the relationships between the value of the dataset a data user can obtain and the length of the time that has passed since the release of the data, we conducted an extensive regression analysis on the set of publications that use dbGaP data. We consider the impact of the publication of a data user as a measurement of the value of the data that is obtained by the data user. We admit that there are other values that a data user can obtain from the data, for instance, the researching funding obtained by writing a proposal using the data. However, the value beyond the impact of the publication is outside the scope of this dissertation, since our goal here is to gain perspective on the change of the value of the data as time passes from a particular perspective.

In particular, we collected information from journal impact reports (JCR) on the impact factor and eigenfactor scores of a set of publications that use dbGaP data in the scholarly journals enlisted in the JCR annual report from 2007 to 2014. Moreover, we collected the manuscript's first received date which indicates the finish date of the work and the published date of the publication. In addition, we associated each publication with the date when the data used in the publication is made available. We fit linear regression models to the impact factor as a function of the length of the period between the data made available and the date when the dataset is released for the set of publications under a series of constraints. The results strongly suggest that the impact of the publication and how soon it comes out after the data is released may not be correlated.

4.2 Preliminaries

4.2.1 dbGaP

dbGaP is a public central repository created by NCBI at NIH to host individual-level phenotype, exposure, genotype, and sequence data, and the associations between them [52]. The data in dbGaP are organized as studies. Each study is assigned a unique identifier with

a prefix "phs" which means phenotype and a six digit number, a version number (*.v#*) and a participant set number (*.p#*). The version number update indicates a change in the phenotype variables collected, while the participant set number update indicates a change in the set of participants.

4.2.1.1 Embargo

Each dbGaP study with a particular version number and participant set number has a embargo release date. Before the embargo release date, only the investigators who contribute to the study that generate the data (i.e., the primary investigators) have the right to seek publications on the data, other investigators can request and download the data, but they are not allowed to publish.

4.2.1.2 Requested data

Although the summary level data, including the study meta data, the association analysis result, and the summary statistics of the phenotype variables are accessible for all users, the individual-level data, including individual-level phenotype and genotype information are only accessible for users that are granted access to. In particular, an Principle Investigator (PI) can request for access to individual-level data of a study in dbGaP, at which time the NIH institute that sponsors the study in dbGaP (i.e., the appropriate NIH Data Access Committee (DAC)) will make a decision on whether or not to grant access. Once the access is granted, the PI can download the de-identified individual-level data file. The authorized users and the institution of a dataset are obliged to comply with the Data Use Certification (DUC) document and responsible research use and data handling of the genomic datasets as defined in the NIH Genomic Data Sharing (GDS) Policy [121]. NCBI only releases de-identified individual-level data. However, as we has been discussed at length earlier in this dissertation, de-identification does not eliminate the possibility of linking the data to a specific individual to violate their privacy [122]. Therefore, the DUC usually states that

the data users agree to not use the datasets to re-identify individuals from whom data were collected. For example, the DUC of the GAIN: International Multi-Center ADHD Genetics Project[1] specifies that "Approved Users agree not to use the requested datasets, either alone or in concert with any other information, to identify or contact individual participants from whom data and/or DNA samples were collected.".

4.2.1.3 Temporal Penalty

According to the GDS Policy, if a data user violates the terms of conditions for secondary research use, NIH will take actions against the user as specified in the DUC. However, to the best of our knowledge, neither the GDS policy nor other NIH policies define the specific penalties that will be enforced on a data user given an DUC violation. As such, the penalty is up to the DUC to specify or to the DUA to decide case by case. One type of penalty that is specified in DUCs is revoking user's access to the dbGaP. For example, the DUC of the GAIN: International Multi-Center ADHD Genetics states that if the user violates the terms in DUC, the DAC may revoke the user's access to all NIH genomic datasets. How long the users, who violated DUC terms, are revoked from the system is, on the hand, recorded in the dbGaP compliance violation report [2]. There were only 27 reported DUC compliance violation incidences. A brief summary of the each incidence, the policy expectations violated, and the action taken and/or preventative measures implemented are reported. An example of such a incidence is, in 2009, an approved user of the dbGaP study: phs000021: Genome-Wide Association study of Schizophrenia, conducted research that was not stated in the data access request. This was a violation because the DUC requires approved users only use the data for the purpose that is described in the approved data access request. After this incidence happened, the DAC revoked the users' access to all NIH genomic datasets for three months. This report shows that the users who violated the DUC are usually suspended from accessing data on dbGaP for a period ranging

[1] https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000016.v2.p2
[2] https://gds.nih.gov/20ComplianceStatistics_dbGap.html

from three to six months. We refer to the suspending of a user's access to the data for a period of time as ***temporal penalty***.

4.2.2 Journal Impact Factor and Eigenfactor Score

The impact factor of journals was first computed in 1969 to rank journals sampled from Science Citation Index (SCI) [123] by the founder of Science CItation Index, Dr. Eugene Garfield. Since 1975 the Intellectual Property and Science of business of Thomson Reuters has been publishing annual Journal Citation Reports (JCR). The JCR 2016 edition includes 11,365 journals from 81 countries in 234 disciplines within the sciences and social sciences.

The JCR impact factor is the ratio between the total number of citations in the current year to the papers published in the previous two years and the total number of papers published in the previous two years. For example, the 2015 impact factor of a journal is the ratio between the total citations in 2015 to the papers published in that journal during 2013 and 2014 and the number of papers published during 2013 and 2014. The citation data for computing the metrics in JCR is from the Thomson Reuters Web of Science.

Journal impact factor has its limitations, since there are factors unrelated to the influence and impact of the journal that affect its value (e.g., the average number of items in the reference list of a published paper in that journal and the type of the articles that are published in the journal). Moreover, the journal impact factor is not suitable for comparing journals across disciplines because the maximal impact factor in different disciplines vary to a large extent.

An alternative measure for journal influence is importance is the eigenfactor score[3][124] reported in JCR since the 2007 edition. The goal of creating the eigenfactor score is to derive a metric that reflects the volume of citations, as well as the quality of the citing journals. To reach this goal, the eigenfactor score approach first built the entire citation network from the same citation data as used in computing journal impact factor. Each node in the citation

[3] http://www.eigenfactor.org/about.php

network represents a journal, while the weight of each directed link represents the number of citations from one node to the other. Based on the citation network, an iterative algorithm similar to Google's PageRank [125] algorithm is used to compute the eigenfactor score for each journal.

To mitigate bias by each of the journal metrics, in this dissertation, we use each of JCR journal impact factor and eigenfactor score of each publication to represent the importance and influence of the publications in a journal in our analysis, respectively.

4.3 Methods

4.3.1 Materials

4.3.1.1 Dataset of Publications involving analysis of dbGaP data

To assess the impact of dbGaP data on facilitating additional biomedical research studies, the librarians at NCBI composed a dataset by gathering information on publications involving analysis of dbGaP dataset authored by approved users of dbGaP data [126] in 2013. This dataset contains information available for each publication on MEDLINE (Medical Literature Analysis and Retrieval System Online, or MEDLARS Online), which is a database of citations for literatures in the domain of life sciences and biomedical information, including title, authors, year, journal name, citation, PMID (the identification number on PubMed, which is the online search engine for MEDLINE), and PMCID (the identification number on PubMed Central, which is an archive of free-access biomedical and life sciences journal literature at the U.S.National Institutes of Healths National Library of Medicine (NLM)).

In addition, the librarians also manually labeled the following fields for each publication:

1. The accession numbers of the dbGaP data cited in the publication (or guessed based on text),

2. The study category (such as methods/computational biology, microbiome, or musculoskeletal),

3. The studied disease or trait (such as bone density, type 2 diabetes, and schizophrenia),

4. Primary study or not,

5. Validation/replication study or not,

6. Methods or not,

7. Controls or not,

8. Population/variation or not.

The NCBI librarians were able to identify approximately 1205 publications between 2007 (the year dbGaP was launched) and the end of 2013 (the year this database was created) that describe studies involving analysis of dbGaP data. Among these publications, 885 did not include the dbGaP accession number in the manuscript. Instead the dbGaP data is referred to by the dataset name.

When a dbGaP dataset is referred to by name, it can be ambiguous. An example of this is the publication with PMID 21796100 and title "The Neuropeptide Galanin and Variants in the GalR1 Gene are Associated with Nicotine Dependence". The dataset that is stated in the manuscript by name is Collaborative Study on the Genetics of Alcoholism (COGA) from dbGaP; however, this name matches two dbGaP accession number including: CIDR: Collaborative Study on the Genetics of Alcoholism Case Control Study with dbGaP Study Accession number phs000125.v1.p1 and The Collaborative Study on the Genetics of Alcoholism (COGA) with the dbGaP Study Accession: phs000763.v1.p1. The librarians at NCBI were able to determine the actual dataset used in this paper is phs000125. Further investigation revealed that the dbGaP study phs000763.v1.p1 was released on January 22, 2015 while this paper is published in published on July 27, 2011, thus the most likely dataset is used in the publication is phs000125.v1.p1.

A set of the publications listed involves data from multiple studies on dbGaP. For example, the publication with PMID 22073273 titled GALC Deletions Increase the Risk of Primary Open-Angle Glaucoma: The Role of Mendelian Variants in Complex Disease used data from two studies in dbGaP: Whole Genome Association Study of Bipolar Disorder (phs000017.v2.p1) and CIDR: Genome Wide Association Study in Familial Parkinson Disease (PD) (phs000126.v1.p1).

The field "*primary study or not*" indicates whether or not the publication is describing the original study that created the data by genotyping study subjects and deposited the genomic data in dbGaP. If the study described in the publication is the source of the dbGaP data, this usually means that the study genotyped the DNA sample of the subjects, and this publication is considered as a primary study. Some of these publications recruited and sampled their own study subjects for genotyping, others genotyped subjects samples from existing study populations. For example, the publication with the PMID 21741921 and title "Gastrointestinal Micbobiome Signatures Of Pediatric Patients With Irritable Bowel Syndrome" is a primary study that recruited the subjects, gathered their data and deposited the data on dbGaP with name The Human Gut Microbiome and Recurrent Abdominal Pain in Children and the dbGaP Study Accession phs000265.v2.p1. On the other hand, the publication with PMID 22508271 and title Fasting Glucose GWAS Candidate Region Analysis across Ethnic Groups in the Multi-Ethnic Study of Atherosclerosis (MESA) is also a primary study involving dbGaP study Multi-Ethnic Study of Atherosclerosis (MESA) SHARe (dbGaP Study Accession: phs000209.v4.p1) that genotyped the subjects in the study but the subjects are not recruited by this study. Instead they are from an existing study population.

A publication that did not describe the process of subject recruitment, sampling or genotyping or sequencing might still be a primary study because the genomic data generation process could have been described in a separate publication for the same study. For example, the publication with PMID 22185703 title as "Morphometic analysis of TCGA

glioblastoma multiforme" is labeled as a primary study involving dbGaP study under accession number phs000178 and phs000489, while the process of generating the genomic data is not described in it.

4.3.1.2 The extension to the dataset

We constructed a relational table from the original dataset by mapping each publication to each dbGaP study accession number without the version number *.v#* and the participant set number *.p#* involved in this publication, since the version and participant set numbers are not provided in publications that cite the used dbGaP data by the name. Then we extended the original table by adding the following fields:

9. The dbgap data release date,
10. The dbgap data embargo release date,
11. The received day of the manuscript,
12. The published online date of the manuscript,
13. The published in print date of the manuscript,
14. The impact factor of the journal,
15. The eigenfactor score of the journal.

The dbgap release date and the embargo release date are directly obtained from NCBI (dbGaP release date 03-14-2016). There are a release date and embargo release date for a particular version (*.v#*) and participant set (*.p#*) of a study (*phs#*). We use the release date and embargo date of the first version and first participant set (.v1.p1).

The received date, the published online date and published date are extracted from the XML file of each publication from the PubMed and PMC databases on the NCBI entrez system. This information is not available for all the publications. The values of the received

date are missing for a large proportion of of publications with high impact factors, such as publications in Science, Nature Genetics, and Nature. A subset of journals are publishing in print only, such that the paper they publish do not have a published online date. The rest of the journals are either publishing in electronic format only, for which the published online date is the same as published date, or in both electronic and print format, for which the published online date might be different from the published in print date.

We also downloaded the Journal Citation Reports of the involved journals from 2006 and 2015. Based on how the JCR journal impact factor is computed, the journal impact factor of the next two years are based on the citation data of the papers published in the each year. Therefore, for each paper in the dataset, we assign the average journal impact factors in the two JCR releases after the year in which the paper is published as the journal impact factor for the paper. For example, if a paper is published in 2006, we assign a paper the mean of the journal impact factor of 2007 and 2008. Similarly, we assign a paper the mean of the journal eigenfactor score of the next 5 years after it is published, because the eigenfactor score of the next 5 years are affected by the citation number of the paper of the current year.

The main variables in the extended dataset for each publication are detailed in Table 4.1.

4.3.2 Data imputation

The XML files we downloaded from PubMed and PMC provide the manuscript received date for a subset of publications. PubMed and PMC data also provide a published in electronic form date for all the publications that are available in electronic format. The published date is available for all the publications in the XML files [4]. This published date can be either the electronic version published date or the print version published date. If a

[4]There are papers which only have the volume and issue numbers, the publication date is obtained from the journal's website based on the volume and issue numbers.

Table 4.1: The variables for each publication in the extended dataset

Variable	**Description**
pmid	PubMed identification
pmcid	PubMed Central identification number
primary study	A boolean variable that indicates whether or not this publication is the description of the study by the primary investigators of the corresponding dbGaP data
category	The category of the study described in the paper (e.g., Ethics, Digestive Disorders, Immune System, Mental Disorders, Behavior/ Cognition, Anthropometry, Cancer, Population Genetics, Respiratory & Environment)
disease/trait	The disease or trait studied
validation/replication	A boolean variable that indicates if the publication is the description of a validation/replication of another study
methods	A boolean variable that indicates if the publication is the description of a methods study
controls	A boolean variable that indicates if the related dbGaP data is used as a control set
pop/variation	A boolean variable that indicates if the related dbGaP data is population variation
journal title	The title of the journal
journal impact factor	The mean of the JCR journal impact factors of each of the two year after the year in which the paper is pulished
journal eigenfactor score	The mean of the JCR journal eigenfactor scores of each of the five year after the year in which the paper is published
dbGaP study	The accession number of the dbGaP dataset involved and the version number and the participant set number *phs*#.*v*#.*p*#
dbGaP study data release date	The release date of the dbGaP data
dbGaP study data embargo release date	The ending date of the embargo period
manuscript received date	The date when the manuscript is received
paper published date	The date when the paper is published in the journal
paper electronic version published date	The date when the electronic version of the paper is published in the Internet

journal is specified as published in print only, the published date is the published in print date. On the other hand, if a journal is not specified as in print only, it can be either electronic only or electronic and print, which is unspecified. We infer the published date as published in print date only if it is different from the published in electronic format date. The availability of these date information for publications in our sample set is summarized in Table 4.2.

Table 4.2: Summary of the Availability of Date Information in the Sample Publication Set

	Electronic Only	**Print Only**	**Electronic & Print**	**Total**
Received Date Available	98	18	176	292
Received Date Unavailable	6	47	60	113
Total	104	65	236	405

We imputed the missing values in the received date column. The imputation is based on the hypothesis that the there is a linear relationship between the manuscript received date and the published date including the published in electronic form date and the published in print form date. In particular, we sampled two datasets from all the publications in the original publication set. The first dataset contains all the publications with received date and published in electronic form date; the second dataset contains all the publications with received date and the published in print form date. The sample size of the first and the second dataset is 752 and 566, respectively (the numbers there are overlaps between them).

We use ordinary linear squares (OLS) and fit a linear regression model to each dataset. The results are depicted in Figures 4.1 and 4.2. Each date value is converted to an integer that corresponds to the number of days between the date and a fixed starting point, which in this case is 2007-01-01. The reason for selecting this date as starting point is because most of the publications in our dataset are published after this date.

The regression coefficients and measurements are shows in Tables 4.3 and 4.4, respectively. The results suggest that the *stderr* of the model of *received date* and *electronic version published date* is less than that of model of *received date* and *print version published date*. Therefore, we use the linear regression model that fit the data of the electronic

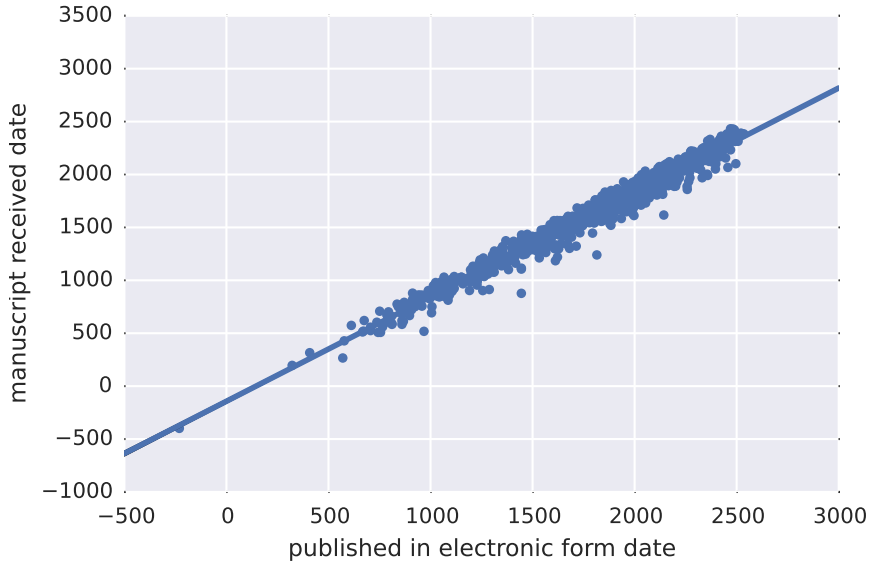


Figure 4.1: The scatterplot and the OLS fitted line of the published in electronic form date versus the manuscript received date.

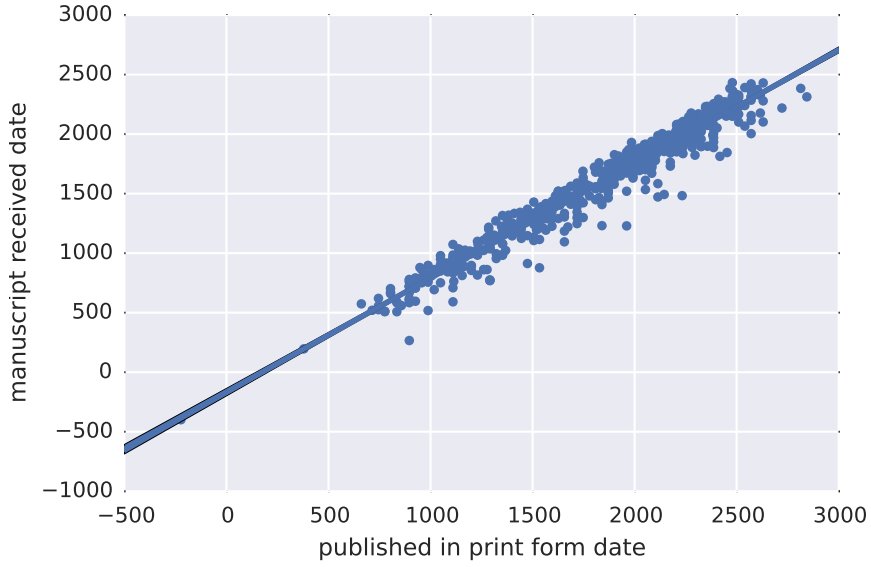


Figure 4.2: The scatterplot and the OLS fitted line of the published in print form date versus the manuscript received date.

Table 4.3: Parameter estimates of the OLS model for predicting received date

	$y = received\ date, x = print\ version\ published\ date$				
	$coeff$	$std\ err$	t	$P > \|t\|$	$95.0\% Conf. Int.$
Intercept	-167.2629	16.624	-10.061	0.000	[-199.916, -134.610]
print version published date	0.9580	0.009	106.425	0.000	[0.940, 0.976]
	$y = received\ date, x = electronic\ version\ published\ date$				
	$coeff$	$std\ err$	t	$P > \|t\|$	$95.0\% Conf. Int.$
Intercept	-142.7491	11.140	-12.814	0.000	[-164.619, -120.879]
electronic version published date	0.9867	0.006	159.517	0.000	[0.975, 0.999]

Table 4.4: The performance measures of the OLS model for predicting received date

	y = received date x=print version publish date	y=received date x=electronic version publish date
R-squared	0.953	0.971
Adj. R-squared	0.952	0.971
F-statistic	1.133e+04	2.545e+04
Prob (F-statistic)	0.00	0.00
Log-Likelihood	-3462	-4355.2

version published date and received date to predict the missing manuscript received date if a publication has electronic version. Otherwise, we use the model of *received date* and *print version published date* to predict the missing manuscript received.

4.3.3 Regression analysis

The goal of the regression analysis is to anwser the question of whether or not the value of the data from a dbGaP study data declines over time after the embargo release date of the dbGaP data. The value of data from a dbGaP study is measured by the influence of the publication that uses the data in terms of journal impact factor and journal eigenfactor score.

4.3.3.1 Linear mixed effects model

We use a linear mixed effects (LME) model to describe the relationship between the *journal impact factor/journal eigenfactor score* and the (*manuscript received date - dbGaP study embargo release date*). To simplify the presentation, we denote ***journal impact factor***

as ***jif***, ***journal eigenfactor score*** as ***jes***, (***manuscript received date - dbGaP study embargo release date***) as ***period***.

An LME model is a generalized linear regression model tha consists of fixed effects and random effects. The probability distribution of the response variable depends on both the variable with fixed effects and random effects. If the relationship between a variable and the response variable is fixed, this variable has a fixed effect on the response variable. The usual linear regression terms are the fixed effects. On the other hand, the random effects in the LME model are those variables that affect the response variable in a manner that can not be described by linear regression. A typical example is a dataset consisting of repeated measures to each individual, who are randomly sampled from a population. The individual variance has an effect on the response variable, but this effect is random instead of linear. The random effects usually represent the individual or the group where the data is from. The individual or group variance contribute to the variance in the response variable which is independent of the random error. The cases that belong to the same individual or group are dependent, such that an LME model captures this dependency, while a simple linear regression model does not.

We use the LME model because the cases in our dataset are grouped by the dbGaP study where the data used in a publication is from. The publications that use data from the one dbGaP study belong to a cluster. The publications in one cluster might have a different initial impact from publications that use data from another. Also, the rate at which the impact factor of the publications changes can also vary from one cluster to another. The size of the clusters vary to a large extent as shown in Figure 4.3, but the LME model does not require the clusters to have similar sizes. As shown in Figures 4.4 and 4.5, the fitted simple linear regression lines using ordinary least square (OLS) differ from one cluster to another for the top 5 clusters.

We fit two LME models: **Model 1** for *jif* and **Model 2** for *jes*. We denote the jif and *jes* of the j^{th} publication in cluster i as $jif_{i,j}$ and $jef_{i,j}$, respectively. m is the number of

clusters, n_i is the size of cluster, $period_{i,j}$ is the *period* of the j^{th} publication in cluster i. $\varepsilon_{i,j}$ is the random error. β_0 is the fixed effect parameter of the constant term 1, and β_1 is the fixed effect parameter of the term *period*. γ_0 is the random effect parameter of the constant term 1, and γ_1 is the random effect parameter of the term *period*. D is the covariance matrix of random effect γ_i.

Model 1

$$\begin{aligned} jif_{i,j} &= \beta_0 + \beta_1 \times period_{i,j} + \gamma_{i0} + \gamma_{i1} \times period_{i,j} + \varepsilon_{i,j} \\ \gamma_i &\sim N_2(0, D) \\ \varepsilon_i &\sim N(0, \sigma^2) \\ i &= 1, ..., m; j = 1, ..., n_i \end{aligned} \tag{4.1}$$

Model 2

$$\begin{aligned} efs_{i,j} &= \beta_0 + \beta_1 \times period_{i,j} + \gamma_{i0} + \gamma_{i1} \times period_{i,j} + \varepsilon_{i,j} \\ \gamma_i &\sim N_2(0, D) \\ \varepsilon_i &\sim N(0, \sigma^2) \\ i &= 1, ..., m; j = 1, ..., n_i \end{aligned} \tag{4.2}$$

4.3.3.2 Data used in the regression analysis

We use the extended dataset of publications to do the regression analysis after proper imputation as described in the previous section. The fields in the extended dataset is shown in Table 4.1. Each publication corresponds to a case in the extended dataset. We removed the publications that use data from multiple dbGaP studies because each of these publications correspond to multiple cases in the extended dataset. There are 532 publications in total, which use only one dbGaP study. We also removed publications that are received before the embargo release date of the dbGaP study. By doing so, the number of publications are reduced to 491. Finally, we excludes the publications in journals not listed in JCR. By doing so, the number of publications used in our analysis is further reduced to 439.

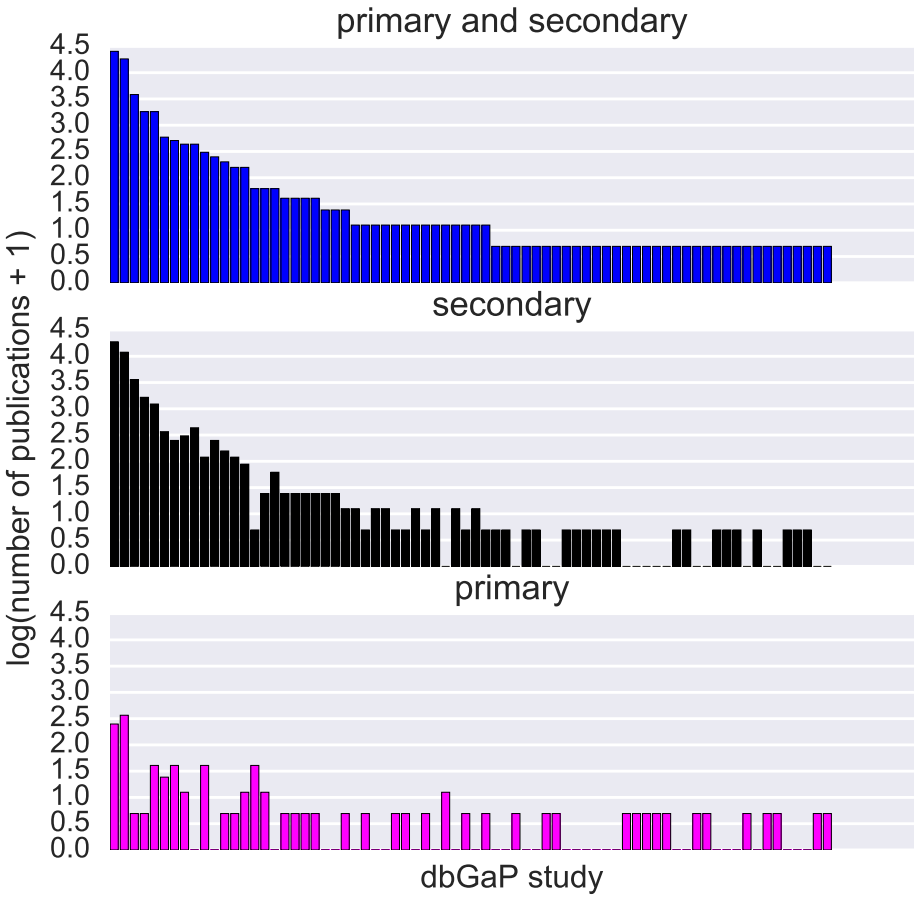


Figure 4.3: The number of secondary and primary publications that use data from each dbGaP study in the publication set. The top 5 dbGaP studies with most publications are: phs000178: The Cancer Genome Atlas (TCGA), phs000007: Framingham Cohort, phs000020: Major Depression: Stage 1 Genomewide Association in Population-Based Samples, phs000021: Genome-Wide Association Study of Schizophrenia, phs000017: Whole Genome Association Study of Bipolar Disorder.

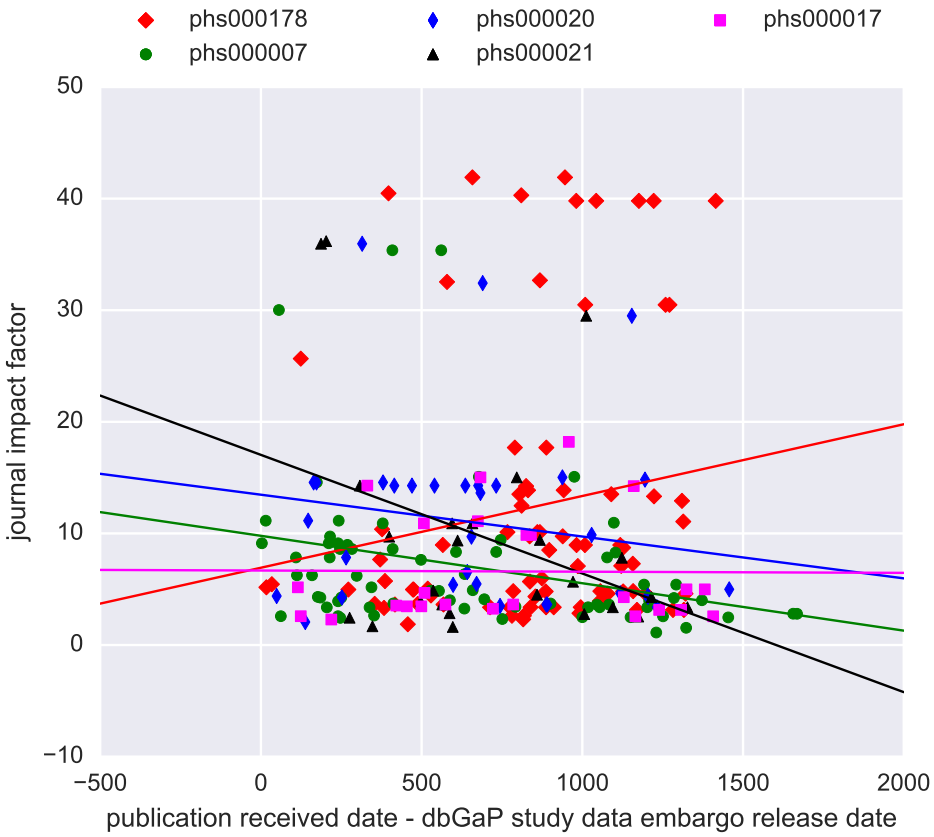


Figure 4.4: The impact factor versus the length of time between the publication received date and the related dbGaP study embargo release date of secondary and primary publications that use data from the top 5 dbGaP studies with most publications: phs000178: The Cancer Genome Atlas (TCGA), phs000007: Framingham Cohort, phs000020: Major Depression: Stage 1 Genomewide Association in Population-Based Samples, phs000021: Genome-Wide Association Study of Schizophrenia, phs000017: Whole Genome Association Study of Bipolar Disorder. The lines are the OLS lines for each cluster of publications grouped by the dbGaP study.

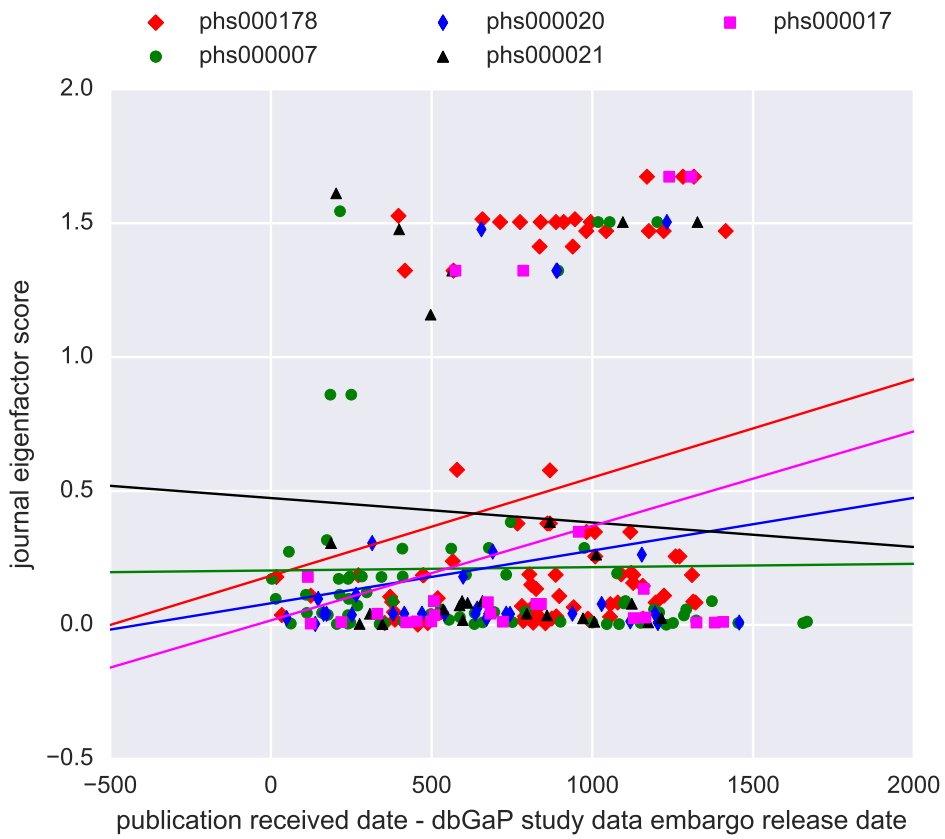


Figure 4.5: The eigenfactor score versus the length of time between the publication received date and the related dbGaP study embargo release date of secondary and primary publications that use data from the top 5 dbGaP studies with most publications: phs000178:The Cancer Genome Atlas (TCGA), phs000007: Framingham Cohort, phs000020: Major Depression: Stage 1 Genomewide Association in Population-Based Samples, phs000021: Genome-Wide Association Study of Schizophrenia. The lines are the OLS lines for each cluster of publications grouped by the dbGaP study.

Table 4.5: The summary of the sizes of clusters in the dataset of primary and secondary publications.

number of clusters	38
min. cluster size	2
max. cluster size	81
mean cluster size	10.7

Table 4.6: The summary of the sizes of clusters in the dataset of secondary publications.

number of clusters	30
min. cluster size	2
max. cluster size	71
mean cluster size	11.7

There are primary publications (79 in total) and secondary publications (360 in total) in our dataset. The summary statistics of the sizes of the clusters by dbGaP study are shown in Tables 4.5 and 4.6. The summary graph of the primary and secondary publications are shown in Figures 4.6 and 4.7. The primary investigators might have a competitively edge to publish prestigious papers before the secondary investigator because they created the dataset.

We conduct regression analysis on 1) the secondary publications and 2) the primary and secondary publications. For each analysis, we remove a case if the dbGaP study associated only has one case in the analyzed dataset.

4.4 Results

4.4.1 Model 1: $jif \sim period$

The results were obtained using Python 3 MixedLM package. The random effects parameter estimates and fixed effects parameter estimates of Model 1 for the dataset of all the primary and secondary publications are shown in Tables 4.7 and 4.8. The random effects parameter estimates and fixed effects parameter estimates of Model 1 for the dataset corresponding to all the secondary publications are shown in Tables 4.9 and 4.10.

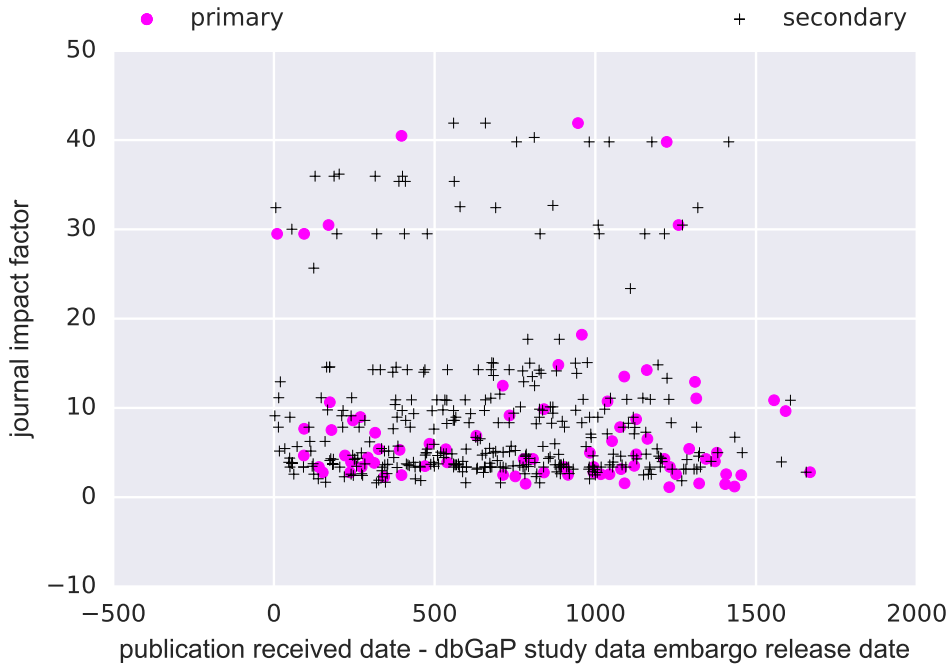


Figure 4.6: The impact factor versus the length of time between the publication received date and the related dbGaP study embargo release date of primary and secondary publications.

Figure 4.7: The eigenfactor score versus the length of time between the publication received date and the related dbGaP study embargo release date of primary and secondary publications.

Table 4.7: Random effects parameter estimates of **Model 1** fitted to the data of **primary and secondary publications**.

	$Coef$	$Std.Err.$
$D[0,0]$	9.519	1.288
$D[0,1]$	-0.010	0.001
$D[1,1]$	0.000	0.000

Table 4.8: Fixed effects parameter estimates of **Model 1** fitted to the data of **primary and secondary publications**.

	$Coef$	$Std.Err.$	z	$P > \|z\|$	[0.025 0.975]
β_0	9.468	1.209	7.834	0.000	[7.099, 11.837]
β_1	-0.002	0.002	-1.310	0.190	[-0.005, 0.001]

The random effects parameter estimate show that, for both datasets, the variance of the random effects on the variable *period* is rounded to 0. The indication is that the impact factor of publications belonging to different clusters (using data from different dbGaP studies), change at the same rate as the *period* changes. Therefore, we only need to focus on the slope itself to perform a hypothesis test to see if there is a significant linear relationship between *period* and jif. The null hypothesis is:

$$H_0 : \beta_1 = 0 \tag{4.3}$$

The alternative hypothesis is:

$$H_a : \beta_1 \neq 0 \tag{4.4}$$

The p-value is 0.190 for the dataset consisting of both primary and secondary publications, and 0.495 for the dataset consisting of only secondary publications. Therefore, there is no significant evidence to reject the null hypothesis to believe that there is a significant linear relationship between *period* and jif.

Table 4.9: Random effects parameter estimates of **Model 1** fitted to the data of **secondary publications**.

	$Coef$	$Std.Err.$
$D[0,0]$	19.251	2.057
$D[0,1]$	-0.023	0.002
$D[1,1]$	0.000	0.000

Table 4.10: Fixed effects parameter estimates of **Model 1** fitted to the data of **secondary publications**.

	$Coef$	$Std.Err.$	z	$P > \|z\|$	[0.025 0.975]
β_0	9.357	1.519	6.161	0.000	[6.380, 12.333]
β_1	-0.001	0.002	-0.682	0.495	[-0.005, 0.003]

4.4.2 Model 2: $jes \sim period$

The maximum likelihood optimization failed to converge for Model 2. Thus we constrained the LME model to only allow the random effect on the intercept by setting $D[1,1] = 0$ and $D[0,1] = 0$. The estimated model parameters for both primary and secondary publications are shown in Tables 4.11 and 4.12. The estimated parameters for primary and secondary publications are shown in Tables 4.13 and 4.14.

Using the same null hypothesis testing as in Model 1, the p-value is 0.067 for the dataset consisting of both primary and secondary publications, and 0.003 for the dataset consisting of only secondary publications. Therefore, there is insufficient evidence to reject the null hypothesis to believe that there is a significant linear relationship between *period* and *jes* for the combination of primary and secondary publications. However, there is a significant linear relationship between *period* and *jes* for the secondary publications only. Since our

Table 4.11: Random effects parameter estimates of **Model 2** fitted to the data of **primary and secondary publications**.

	$Coef$	$Std.Err.$
$D[0,0]$	0.035	0.040
$D[0,1]$	0	0
$D[1,1]$	0	0

Table 4.12: Fixed effects parameter estimates of **Model 2** fitted to the data of **primary and secondary publications**.

	$Coef$	$Std.Err.$	z	$P > \|z\|$	[0.025 0.975]
β_0	0.258	0.066	3.930	0.000	[0.129, 0.386]
β_1	0.000	0.000	1.835	0.067	[-0.000, 0.000]

Table 4.13: Random effects parameter estimates of **Model 2** fitted to the data of **secondary publications**.

	$Coef$	$Std.Err.$
$D[0,0]$	0.055	0.063
$D[0,1]$	0	0
$D[1,1]$	0	0

interest is whether or not the *jes* decreases when *period* increases, we formulate another null hypothesis and the corresponding alternative hypothesis:

$$H_0' : \beta_1 = \alpha, \alpha <= 0 \tag{4.5}$$

$$H_a' : \beta_1 ! = \alpha, \alpha <= 0 \tag{4.6}$$

For a non-positive value α, the p-value for the null hypothesis H_0' will be < 0.0015. Therefore, the null hypotheis is rejected for all the non negative α. As such, there is statistically significant evidence to believe that the *jes* does not decrease as *period* increases.

4.5 Discussion and Conclusions

The results for the combination of primary and secondary publications and all the secondary publications suggest that there is not sufficient evidence to believe that the journal

Table 4.14: Fixed effects parameter estimates of **Model 2** fitted to the data of **secondary publications**.

	$Coef$	$Std.Err.$	z	$P > \|z\|$	[0.025 0.975]
β_0	0.232	0.077	3.011	0.003	[0.081, 0.384]
β_1	0.000	0.000	2.984	0.003	[0.000, 0.000]

impact factor or the eigenfactor score of publications that use a dbGaP study dataset declines over time after the embargo release date of the dataset.

Still, this analysis has several limitations which lead to several future research directions. First, the data that is used in the regression analysis may be incomplete and inaccurate because it does not contain the publications after 2014 and the manual labeling process might have introduced some errors.

In the future, natural language processing algorithms could be developed to discover publications that use data from dbGaP studies and the particular dbGaP studies and the versions and participants set numbers associated with it. Still, this task is nontrivial because not all publications which use data from dbGaP refers to the dbGaP study identifier. We anticipate that the algorithms to tackle this task need to intelligently identify the context in which the dataset used in the analysis is mentioned. They also need to extract the text refer to the dataset from the context.

Our models are based on the premise that only data from one dbGaP study is used in each publication. However, this is not always the case. There are many instances in which data from multiple dbGaP studies are used. In fact, this is one of the main reasons for sharing these data in the first place (i.e., to allow new scientific discoveries by aggregation). This limitation is also responsible for the exclusion of around half of the publications from the original publication set from our analysis. To account for publications that aggregate data from multiple dbGaP studies, there needs to be an approach to estimate the contribution of each dataset to the total value of the publication.

Third, our LME models only considers the random effects caused the dbGaP study that is used in the publication. In other words, in our LME models, we only consider the dbGaP study as the grouping variable. However, there might exist other variables which also have random effects: such as the study category, the trait and disease studies, whether or not it is a primary study and whether or not it is a method publication. In the future, an more extensive analysis can be carried out by considering all these variables.

Fourth, we chose to use the embargo release date of the first version and first participant set of each dbGaP study as the date that the study is made available for all users without accounting for the particular version and participant set that are used. This is due to two reasons: 1) the version and participant set are not available for all publications in our data and 2) the difference between versions was expected to be insignificant. The problem with this strategy is that, for some dbGaP studies, the the difference between versions can be significant if they were two different studies. In this case, our analysis can lead to biased result by assuming the dbGaP study data is made available to all at the time it was first embargo released instead of the embargo release time of the particular version and participant set.

Finally, in the experiments we use the journal impact factor and journal eigenfactor score of the publication to represent the impact and importance of the publication. In the future, other different metrics can be also accounted for, such as other journal impact metrics or the citations to the publication itself.

Chapter 5

Search for Optimal Tradeoff between Re-identification Risk and Data Utility

5.1 Introduction

To achieve de-identification, many organizations often times follow a pre-defined de-identification policy by privacy laws and regulations. One de-identification policy that is broadly adopted by many healthcare organizations in US is the Safe Harbor model defined by U.S. Health Insurance Portability and Accountability Act (HIPAA), which specifies 18 rules, including suppression of explicit identifiers ((e.g., personal names) and generalization of "quasi-identifiers" which could enable linkage (e.g., dates of events, such as birth, are replaced with time periods no more specific than one year and ages over 89 years-old are recoded as 90). Yet, the rigidity of such rule-based policies is not ideal for sharing every dataset, such as studies with the elderly (e.g., dementia patients). Thus, the law enables publishers to use an alternative, which permits data to be shared in any format, provided the risk of reidentification is appropriately measured and mitigated. In order to appropriately balance the competing needs of minimizing risk (R) and maximizing utility (U), the majority of previous works have focused on optimizing in the context of an anonymization model(e.g., k-anonymity [38]), but this is a more rigid formalism than de-identification [127], thus the solutions may not be optimal in turns of the RU tradeoff. This work develops an algorithm to efficiently discover a set of policies that form an R-U frontier, which offers a collection of mutually exclusive deidentification policies.

As a concrete example, Figure 5.1 depicts how a record from a dataset investigated in our experimental analysis is transformed by one of the frontier policies in comparison with its Safe Harbor and 10-anonymous (i.e., the record is part of a group of no less than 10 records with the same values) versions. In this example, R is defined as inversely propor-

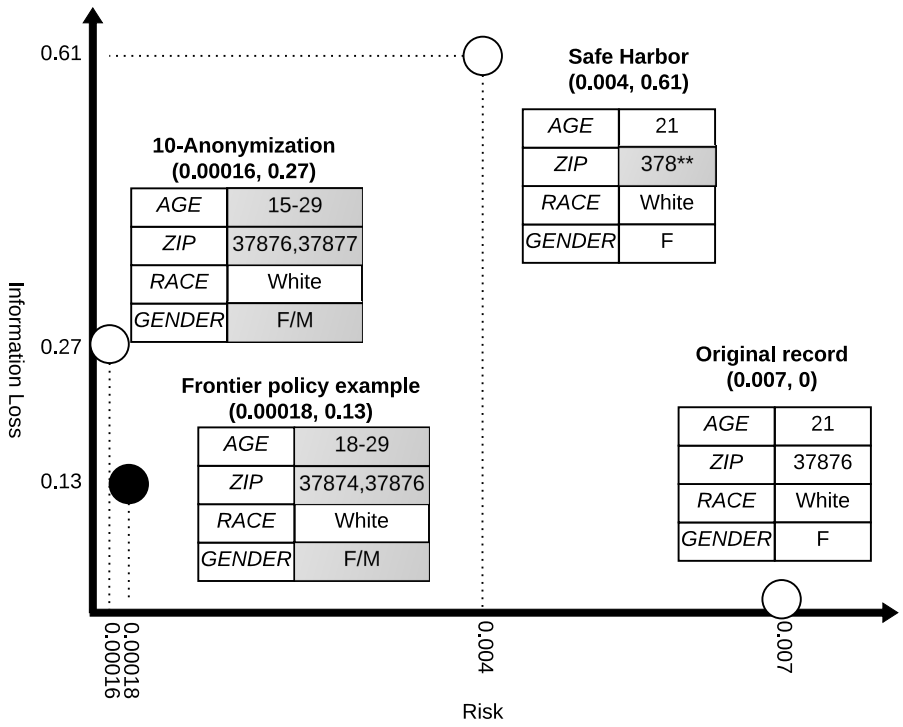


Figure 5.1: An illustrative example of demographic de-identification policies in the risk-utility space (where utility is defined as similarity between the original record and the protected record.

tional to the size of this demographic group defined by the record in a population set, while U is in terms of an information loss metric which represents the discrepancy between the probability density of the record in the original dataset and the transformed dataset. Safe Harbor transforms this record into a group with a large set of ZIP code areas, while 10-anonymization and a policy on the R-U frontier transform it into two different age groups and small ZIP code groups. Based on the population size in these different groups, the record transformed via the frontier policy has slightly higher risk than its 10-anonymous counterpart, while Safe Harbor has the highest risk. On the other hand, the record in the dataset transformed via a frontier policy has lower information loss than its counterparts of both Safe Harbor and 10-anonymous.

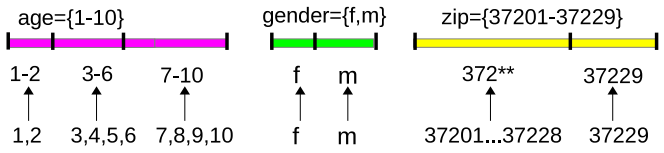


Figure 5.2: An example of a de-identification policy defined over three quasi-identifying attributes, $\{Age, Gender, ZIP\}$.

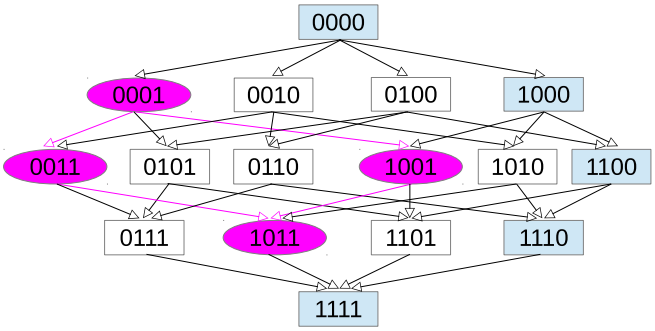


Figure 5.3: An example of a de-identification policy lattice with five quasi-identifying values. Rectangular nodes depict a maximal chain, while oval nodes represent a sublattice.

5.2 The Policy Space

Our policy space is an extremely large set of de-identification strategies. The solution space is limited to the different ways to de-identify the dataset that are compatible with an existing de-identification policy, such as HIPAA Safe Harbor. Each de-identification policy defines a generalization schema for each quasi-identifying attribute in the dataset. In particular, we require a total order in the domain of each quasi-identifying attribute and apply a full-subtree-generalization model [63], which means that the values in a domain are mapped to a set of non-overlapping intervals. As such, a mapping function can be defined by the corresponding partition on the domain of a quasi-identifying attribute.

Figure 5.2 provides an example of a de-identification policy. The set of QI attributes is $\{Age, Gender, ZIP\}$ and the domains are $\{1, \ldots, 10\}$, $\{male, female\}$, and $\{37201, \ldots, 37229\}$, respectively. In this policy *Age*, *Gender*, and *ZIP* are mapped to the aggregated groups: [1-2], [3-6] and [7-10]; [*female*] and [*male*]; and [37201, ..., 37228] and [37229], respectively. This policy is valid because the aggregated groups of values for each QI

compose a partition of the corresponding domain. By contrast, a mapping of ages to $[1-5]$ and $[3-10]$ does not constitute a valid policy because the intervals overlap (i.e., 3, 4, or 5 could be in either interval).

We use the full subtree generalization model because it offers several advantages over alternative models. First, as shown in [128], it enables representation of fine-grained policies, as well as common policies encountered in the real world (e.g., HIPAA Safe Harbor). Second, the set of policies can be structured as a lattice of generalizations which can be systematically searched. Third, it is straightforward to interpret how a policy changes the syntax of the data.

We represent policies as bit-strings. To characterize the translation, let n be the number of values in the domain of a quasi-identifying attribute. After enforcing a total order on the values, they are mapped to a bit-string of size $n-1$. The original domain is represented by a bit-string of 1's, whereas a bit of 0 indicates a demarcation in the partition has been removed to widen an interval[1] (i.e., values have been generalized). For example, the bit-string for *Age* in Figure 5.2 is $[0,1,0,0,0,1,0,0,0]$.

Our algorithm first maps each policy into a quantitative risk-utility (or R-U) space. And, in such a space, policies can be partially ordered and structured on a lattice, through which a systematic search can be conducted. An example of such as lattice is depicted in Figure 5.3.

In preparation for our policy search algorithms, we define two types of subgraphs over the lattice: 1) chain and 2) sublattice. A chain is a totally ordered subset in the lattice. A maximal chain is one that is not a proper subset of any other chain. In Figure 5.3, the rectangular nodes (i.e., [0, 0, 0, 0], [1, 0, 0, 0], [1, 1, 0, 0], [1, 1, 1, 0], [1, 1, 1, 1]) constitute a maximal chain. A sublattice is a subgraph *i*) bounded by an upper node and lower node on some chain and *ii*) contains every node in the set of all chains between them. Any two policies with a chain between them can define a sublattice. An example of a sublattice is

[1]The final value in the domain is implicit in the partition.

shown in the oval nodes of Figure 5.3. Here *sublattice*([0, 0, 0, 1], [1, 0, 1, 1]) defines the set {[0, 0, 0, 1], [0, 0, 1, 1], [1, 0, 0, 1], [1, 0, 1, 1]}.

5.3 Search Algorithms

The size of a typical policy lattice is too large for an exhaustive, systematic search. Thus, we developed two heuristic approaches: 1) Random Chain Search and 2) Heuristic Sublattice Search.

5.3.1 Random Chain

The first strategy is called the Random Chain Search (RCS) and is shown in Algorithm 1.

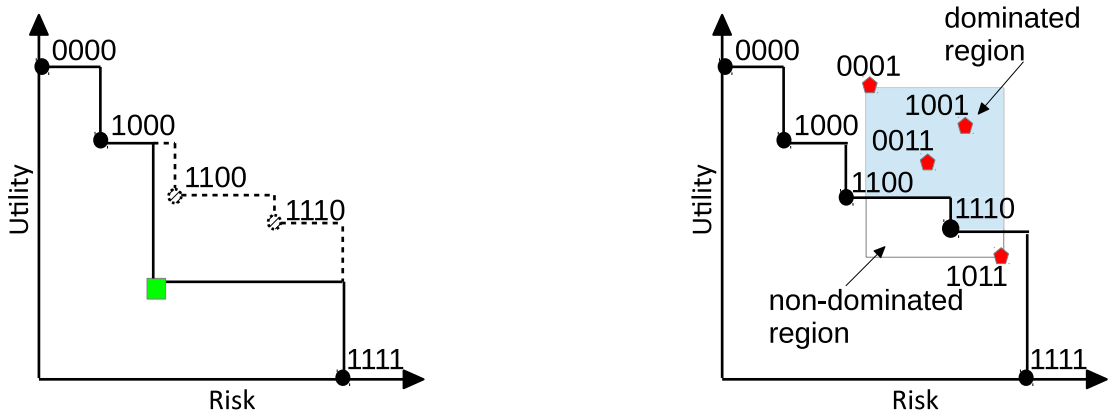
Algorithm 1 Random Chain Search (*RCS*)

Input: n, the maximum number of policies to estimate; L, the length of a policy; T, a dataset
Output: f, the frontier policies

1: $i \leftarrow 0$
2: $f \leftarrow$ InitializeFrontier() {This function returns a non-dominated set of policies, including top and $bottom$.}
3: **while** i ¡ n **do**
4: $c \leftarrow$ selectRandomChain() {This function begins at $bottom$. It iteratively selects a policy at random from the GLB, to which it proceeds until it reaches top. It returns all the policies selected.}
5: **for all** α in c **do**
6: $f \leftarrow$ updateFrontier(f,α)
7: **end for**
8: $i \leftarrow i + L$ {L is the number of policies on the chain}
9: **end while**
10: **return** f

The process begins by assigning an arbitrary non-dominated set of policies in the lattice to the frontier, which is accomplished through *InitializeFrontier*(). Next, we iteratively select maximal random chains, via the *selectRandomChain*(), and update the frontier with policies on the chains. This process iterates until n policies have been searched. Updating

the frontier is accomplished through the function $updateFrontier(f,\alpha)$, which attempts to revise the frontier f with each policy α in the chain. If the frontier does not contain policies that dominate α, it is inserted into the frontier. The frontier then drops all policies dominated by α. As an example of this function, consider Figure 5.4(a).



(a) Frontier update

(b) Sublattice Overlay

Figure 5.4: An example of updating the frontier in the R-U space using policies from Figure 5.3. The current frontier is composed of policies mapped to the stair-step curve. In 5.4(a), the policy mapped to the square will be added to the frontier because it dominates policies currently on the frontier (i.e., [1, 1, 0, 0] and [1, 1, 1, 0]), which will be removed. In 5.4(b), the rectangle represents the bounding region of the R-U mapping of policies in sublattice ([0, 0, 0, 1], [1, 0, 1, 1])

5.3.2 Sublattice Heuristic Search

The RCS algorithm is naïve in that it assumes all regions of the lattice are equally likely to update the frontier. However, this is not the case, and we suspect sublattices can be compared to the frontier to search more efficiently. Consider, given a frontier f, we can draw a stair-step curve in the R-U space that connects all policies on the frontier. An example of such a curve is depicted in Figures 5.4(a) and 5.4(b). It is clear that any policy mapped to the region above the curve will be dominated by at least one policy on the frontier. Additionally, any policy mapped to the region below the frontier will always update the frontier. Thus, this curve divides the R-U space into two regions: 1) dominated

and 2) non-dominated.

Given *sublattice*(α, β), it can be proven that the risk and utility values of policies in the sublattice are bounded in a rectangle defined by the risk and utility values of policies α and β, which we call the *bounding region*. In other words, all policies in the sublattice will have risk in the range [$R_T(\alpha)$, $R_T(\beta)$] and utility in the range [$U_T(\beta)$, $U_T(\alpha)$].[2] For example, in Figure 5.3, all policies in *sublattice*([0, 0, 0, 1], [1, 0, 1, 1]) are mapped to the rectangular area bounded by the R-U mapping of the *top* and *bottom* policies of the frontier in Figure 5.4(b).

To leverage this fact from a probabilistic perspective, we assume that the policies in a sublattice are uniformly distributed in the bounding rectangle. This implies the probability that a policy in a sublattice updates the frontier is the proportion of the lattice's bounding rectangle which falls below the curve of the frontier. Formally, imagine policies on the frontier f are mapped to a set of R-U points that are ordered increasingly by risk $\{(r_0, u_0)$, $\ldots$, $(r_h, u_h)\}$, where h is the number of polices on the frontier. Now, given *sublattice*(α, β), let us assume the policies α and β are mapped to points (r_α, u_α) and (r_β, u_β), respectively.

We compute the area of the bounding region as $(r_\beta - r_\alpha) \times (u_\alpha - u_\beta)$. If we draw a line parallel to the y-axis at each point of the frontier in the R-U space, then the area of the non-dominated region is composed of the resulting rectangles. More specifically, if $r_i < r_\alpha < r_{i+1}$ and $r_j < r_\beta < r_{j+1}$, then the area of the non-dominated region is

$$ND(s, f) = \sum_{k=i}^{j} \max(0, \quad (u_k - u_\beta) \times (\min(r_\beta, r_{k+1}) - \max(r_k, r_\alpha)))$$

Finally, the probability that a policy in the sublattice can update the frontier is computed

[2] A proof sketch for this claim is as follows. Any policy γ in sublattice(α, β) satisfies $\alpha \prec \gamma \prec \beta$ and $R_T(\alpha)$ and $U_T(\alpha)$ satisfies the order homomorphisms. Thus, $R_T(\alpha) \leqslant R_T(\gamma) \leqslant R_T(\beta)$ and $U_T(\alpha) \geqslant U_T(\gamma) \geqslant U_T(\beta)$.

as:

$$H(s,f) = \frac{ND(s,f)}{(r_\beta - r_\alpha) \times (u_t - u_\beta)} \tag{5.1}$$

For example, in Figure 5.4(b), the probability that any policy in *sublattice*([0, 0, 0, 1], [1, 0, 1, 1]) can update the frontier is the ratio between the area below the step curve in the rectangle (*non-dominated region*) and the entire rectangle (*bounding region*). Based on this observation, we introduce a second search algorithm called the Sublattice Heuristic Search (SHS). The steps of the process are shown in Algorithm 2, which we describe here.

Algorithm 2 Sublattice Heuristic Search (*SHS*)

Input: n, the maximal number of policies to assess; TH, the threshold for searching a sublattice; L, the length of a policy; T, a dataset
Output: f, list of frontier policies of the searched policies

```
1:  i ← 0
2:  f ← initializeFrontier()
3:  prunedlist ← ∅
4:  while i ¡ n do
5:      sublattice ← generateRandomSublattice(prunedlist)
6:      p ← H(sublattice,f){Equation 5.1}
7:      if p ≥ TH then
8:          c ← selectRandomChain(sublattice)
9:          for all α in c do
10:             f ← updateFrontier(f,α)
11:         end for
12:         i ← i + length(c)
13:     else
14:         if p = 0 then
15:             prunedlist.append(sublattice)
16:         end if
17:         i ← i+2
18:     end if
19: end while
20: return f
```

As in RCS, this algorithm begins with a call to *intializeFrontier*(), which instantiates the frontier f as a non-dominated policy set. Next, the algorithm instantiates a list to maintain memory of which policies (or sections of the lattice) have been pruned due to dominance by the frontier. At this point, the algorithm iteratively selects a sublattice and tailors its

process depending on the following conditions:

- **Condition 1:** If the entire bounding region of a sublattice is in the dominated region of the frontier, the sublattice is pruned.
- **Condition 2:** If the entire bounding region of a sublattice is in the non-dominated region of the frontier, we search a random maximal chain of the sublattice. Though any of the policies in the sublattice can improve the current frontier, they may dominate one another. Moreover, the entire sublattice can contain a substantial number of policies, which would make a complete search infeasible. By contrast, a maximal chain is the maximal set of policies in the sublattice that can be guaranteed to be on the new frontier.
- **Condition 3:** If neither of the previous conditions are satisfied, we use the update probability to determine if the sublattice is worth further searching. Specifically, if the update probability is greater than a threshold, we search a maximal chain of the sublattice, selected at random. Otherwise, no search is initiated.

5.4 Experiments Setup

5.4.1 Real World Policy: HIPAA Safe Harbor

To perform a comparison with an existing rules-based de-identification policy, we compare our frontier to the Safe Harbor policy of the HIPAA Privacy Rule. This policy enumerates eighteen specific attributes that must be generalized or suppressed from a dataset before it is considered de-identified. Of importance to this study, we focus on Safe Harbor's perspective of demographics. For such features, it states that 1) all ZIP codes must be rolled back to their initial three characters and that codes with populations of less than 20,000 individuals must be grouped into a single code and 2) all ages over 90 must be recoded as a single group of 90+. Safe Harbor does not prevent the dissemination of gender

or ethnicity, but we include these features because they are common demographics, which could be generalized in favor of age and geocodes [129].

5.4.2 Evaluation Dataset

For evaluation, we use two publicly available datasets. The first is the Adult dataset [130], which consists of 32,561 records without missing values. For comparison with Safe Harbor, we restrict the quasi-identifiers to the demographics of Age, Gender, Race. To enable a comparison with respect to geography, we combine the available demographics data from Adult with state-level demographic information obtained from the US Census Bureaus 2000, 2010 Census Tables PCT12A-G [131] to provide each tuple with a 5-digit ZIP code. To mitigate the bias that can be introduced through analysis over a single population, we simulated the Adult dataset for 10 US states: Illinois (IL), Hawaii (HI), Massachusetts (MA), Minnesota (MN), New York (NY), Ohio (OH), Pennsylvania (PA), Tennessee (TN), Washington (WA), and Wisconsin (WI). The Census data of the corresponding states are used as the population statistics to compute the re-identification risk of these synthesized datasets. All of these states, with the exception of HI, correspond to regions that contain academic medical centers participating in the Electronic Medical Records and Genomics (eMERGE) network [132]. These centers are collecting and sharing de-identified data on patients to the public and are actively using the Safe Harbor de-identification policy, but are open to alternatives [133]. HI is selected as an additional state because of its unique demographic distribution (e.g., it has the highest percentage of Asians and the lowest percentage of whites in the country).

To provide analysis on non-synthetic data, we also conducted experiments on the North Carolina voter registration (NCVR) database [120], which contains 6,150,562 records without missing values, each record consists of 18 fields. This is the same dataset as the one used in Chapter 3. For this study, we restricted the dataset to a set of four quasi-identifying attributes, Age, Race, Gender, 5-Digit ZIP Code. We use the entire dataset as the population

and randomly sample datasets to publish. The policy lattice, based on the selected quasi-identifier, contains on the order of 2700 policies, which would take a significant length of time to exhaustively search.

5.4.3 Risk Computation

To compute risk, we adopt the disclosure measure in [128], which is based on the distinguishability metric proposed by [134]. This measure assumes a tuple in the generalized dataset contributes an amount of risk inversely proportional to the size of the population group that matches its QI values. Again, the population information is based on the PCT12A-G Census tables.

For example, imagine a record in the Adult dataset is [39, male, white, 37203]. This record is unique in the dataset, but the census tables show there are 5 people in the region with the same demographics. As a result, this record contributes a risk of 0.2. Further details on this risk computation can be found in [128].

The disclosure risk of the entire generalized dataset corresponds to the sum of the risk of each record. To ensure the risk score for a dataset is normalized between $[0,1]$, we divide this sum by the risk value for the original dataset. This dataset has no generalization and constitutes the maximum risk for all policies in the lattice. Given a dataset D, a population P, the formal definition of the risk for generalized dataset D' is:

$$risk(D',P) = \frac{\sum_{d' \in D'} \left(\frac{1}{g(d')}\right)}{max(risk)} \tag{5.2}$$

$$max(risk) = \sum_{d \in D} \left(\frac{1}{g(d)}\right), \tag{5.3}$$

where $g(d')$ is the size of the population group in P with the set of quasi-identifiers of record $d \in D'$. [128] demonstrated this risk measurement satisfies the order homomorphism.

5.4.4 Utility Computation

To compute utility we use an information loss measure. In particular, we use KL-divergence to measure the loss incurred by a generalized dataset with respect to its original form. This measure satisfies the order homomorphism constraint. Formally, the KL-divergence is computed as:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \tag{5.4}$$

where $P(i)$ and $Q(i)$ are the probability distributions of the quasi-identifying values in the original and de-identified datasets, respectively.

While $P(i)$ is computed from the frequency of quasi-identifying values in the original dataset, $Q(i)$ is an approximation. Specifically, $Q(i)$ is based on the assumption that if several values are generalized to a single group, then the corresponding records are uniformly distributed across the group. For example, imagine the quasi-identifier set is $\{Age, Gender\}$ and there is a record in the generalized dataset [*Age* = [1-2], *Gender* = [male, female]] with a frequency of m. Then, each possible value (i.e., [1, male], [1, female], [2, male], and [2, female]) is assigned a frequency of $m/4$. Following [135], we use the standard convention that $\ln 0 = 0$. Based on this definition, the information loss measure is in the range of $[0, 1]$ and there is no need for normalization.

5.5 Performance Evaluation Results

To conduct experiments on efficiency, we provide a search budget of 14,780 total policies to search. This value represents 20 maximal chain searches (i.e., the policy lattice is composed of 739 levels).

First, we evaluated the efficiency of the search algorithms. We assessed the progress of the algorithms over 20 complete runs. There is minimal variance in actual time to completion between the algorithms, but a significant difference in *how quickly* the algorithms

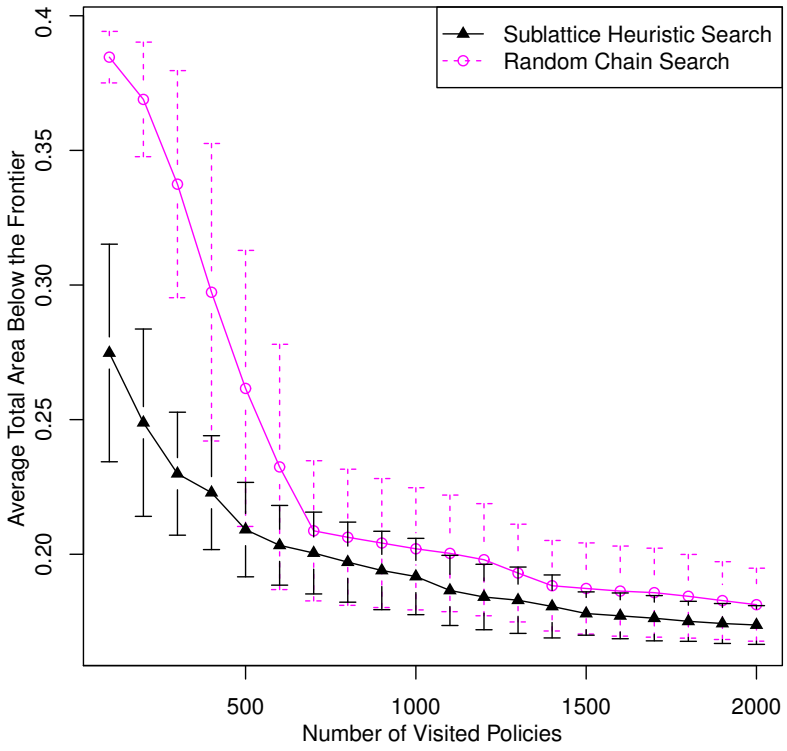


Figure 5.5: The efficiency of search strategies on the Adult dataset as a function of number of policies searched.

converge a high-quality frontiers.

To illustrate this finding, we established checkpoints during the runtime of the experiments. At each checkpoint (e.g., every 100 policies examined), the current average area under the frontier was determined (i.e., smaller areas illustrate better frontiers). The result (mean and the standard deviation) of this evaluation is depicted in Figure 5.5. The result shows that the SHS method dominates the RCS. In particular, after computing 100 policies, the average result of the sublattice search is 28% better than the average result of the random chain search. This result indicates that the SHS method is particularly efficient when a quick solution is needed.

Next, we evaluated the effect of the sublattice heuristic (i.e., area under the frontier) upon searching. In this experiment, we initialized the frontier to a random maximal chain and subsequently generated 24240 random sublattices. For each sublattice, we applied $H()$ to predict the probability that searching a random chain through the sublattice yields

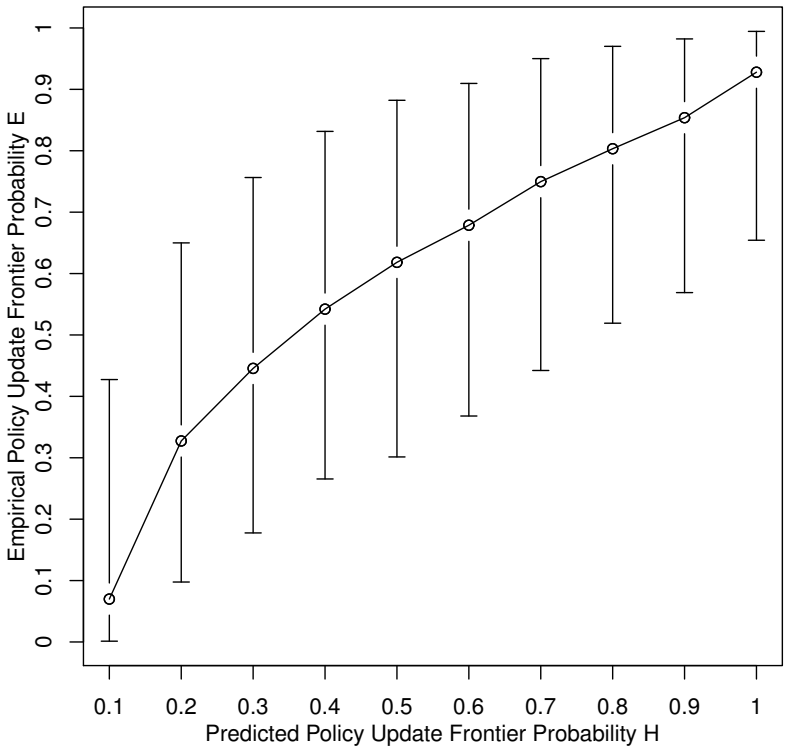


Figure 5.6: An empirical evaluation of the sublattice heuristic H().

frontier updates. We then picked a random chain and computed the empirical probability $E()$ of a policy in a random chain in the sublattice updating the frontier in terms of the ratio of the number of policies that actually updated the frontier and the total number of policies in the searched chain. Finally, we analyze the correlation between the predicted probability $H()$ and the empirical probability $E()$.

We first run a linear regression over the aggregated set of $H()$ and $E()$ values. In particular, we partition the sublattices into 10 groups based on the value of $H()$ (e.g., lattices with $H() \in [0, 0.1]$ are placed in the first group). The representative $H()$ value of each group is assigned to the upper bound of the interval. For the set of lattices in each group, the average and confidence interval of the ratio of number of policies that update the frontier to the total number of policies are computed. The results are depicted in Figure 5.6, where the mean of the actual update ratio clearly increases with the predicted frontier probability. This result suggests that the driving intuition behind the SHS heuristic was reasonable with

respect to the *Adult* dataset.

To further demonstrate the relationship between the probabilities $H()$ and $E()$, we ran a linear regression and a correlation test on the set of values. The Pearson's product-moment correlation coefficient of $H()$ and $E()$ is 0.8643536, with a p-value of 2.2×10^{-16}. This result indicates that the actual probability a policy in the path of a sublattice will improve the current frontier is positively correlated with the estimated probability based on our heuristic.

5.6 Empirical Analysis Results

5.6.1 Frontier Case Studies

The frontier for SHS and 10-anonymization for Adult-TN is depicted in Figure 5.7, while results for the other states are in Figure 5.8. Notably, the results indicate that a region of the frontier discovered by SHS dominates the Safe Harbor policy in all states. Moreover, the frontier region that dominates Safe Harbor results in both greater utility and risk than the results of 10-anonymization.

For illustration, two policies that dominate Safe Harbor (i.e., less risk and better utility) are highlighted in Figure 5.7. The discovered policies exhibit notable differences from Safe Harbor. For instance, both policies generalize race and ages below 90 to larger groups than Safe Harbor (as illustrated in Figure 5.7(c)), but retain more specific geographic information (as illustrated in Figure 5.7(d)). Additionally, the second policy generalizes gender to [Male or Female].

5.6.2 Policies on the Frontier

Table 5.1 reports the number of policies on each frontier. The SHS frontier contains an average of 4,700 policies while the k-anonymity frontier contains an average of 33 and 26 policies when k equals 5 and 10, respectively. This is because SHS can search a signifi-

cantly larger space than the Incognito k-anonymization algorithm, due to the construction of their respective lattices. Even though the number of policies in the SHS frontier is large, these policies are ordered by their R-U values, so a data publisher can quickly locate the policies they are interested in.

5.6.3 Policies Dominating Safe Harbor

The ratio of policies on the SHS and k-anonymization frontiers that dominate Safe Harbor is summarized in Table 5.2. Notice that the SHS frontier contains policies that dominate Safe Harbor in all states. By contrast, 5-anonymization leads to solutions that dominate Safe Harbor for only HI, TN, MN, WA, and WI, while 10-anonymization can only find dominant solution for HI. This is because k-anonymity datasets tend to have more utility loss than does a dataset de-identified through Safe Harbor.

5.6.4 Frontier Ranges

We summarize the result of the comparison of ranges of the k-anonymity frontier and the SHS frontier in Table 5.3. The results indicate k-anonymization solutions are constrained in a very small sub-interval of the SHS frontier. This interval tends to have very small risk and large utility loss. Thus, SHS may be particularly useful when the data publisher is interested in solutions with better utility at the cost of an acceptable increase in risk. For instance, in the state of NY, the maximum risk of the 5-anonymization frontier is only 0.003 that of the SHS frontier. On the other hand, the minimum utility loss of the 5-anonymization frontier is between 0.15 and 0.52, while SHS is always at 0. This phenomenon is visualized is in Figure 5.7 (a), where the 10-anonymization has a much smaller range than that of the SHS frontier. This finding indicates that the SHS frontier can provide solutions in a broader range than the k-anonymity frontier.

Table 5.1: Number of policies on the frontier for the Adult dataset with ZIP codes simulated based on U.S. Census data.

STATE	Number of Policies on Frontier		
	SHS	$k = 5$	$k = 10$
HI	4545	28	25
IL	3999	28	21
MA	3510	29	23
MN	5655	34	25
NY	4374	27	20
OH	3257	39	27
PA	4161	29	23
TN	7766	39	27
WA	5296	33	35
WI	5147	42	34
Average	4771	33	26
St. Dev.	1234	5.5	5.03

5.6.5 Improvement of the Frontier R-U Tradeoff

The frontier R-U tradeoff improvement made by SHS over k-anonymization is reported in Table 5.4. We use the relative difference of AU of the k-anonymization frontier Fk and the SHS frontier Fs to represents the R-U tradeoff improvement rate of the SHS frontier over the k-anonymization frontier: $IR = (AU(F_k) - AU(F_s))/(AU(F_s))$. A positive value indicates Fs improves upon Fk. We truncate the SHS frontier to be in the same range of the corresponding k-anonymization frontier for a fair comparison.

In 9 out of 10 states (OH being the exception), the SHS frontier dominates the k-anonymization frontier. Recall that a positive value in this table indicates the corresponding k-anonymization frontier is dominated by the SHS frontier.

Table 5.2: Proportion of policies that dominate Safe Harbor for the Adult dataset with ZIP codes simulated based on 2010 U.S. census data.

STATE	Proportion of frontier dominating Safe Harbor		
	SHS	$k = 5$	$k = 10$
HI	0.01	0.36	0.28
IL	0.04	0	0
MA	0.01	0	0
MN	0.03	0.09	0
NY	0.001	0	0
OH	0.06	0	0
PA	0.003	0	0
TN	0.01	0.13	0
WA	0.01	0.12	0
WI	0.01	0.10	0
Average	0.018	0.08	0.028
St. Dev.	0.018	0.11	0.084

Table 5.3: Maximum risk values (MAX Risk) and minimum utility loss (MIN Utility Loss) of the frontiers for the Adult dataset with ZIP codes simulated from U.S. census data.

STATE	Max. risk			Min. utility loss		
	SHS	$k = 5$	$k = 10$	SHS	$k = 5$	$k = 10$
HI	0.057	2.6×10^{-3}	1.8×10^{-3}	0	0.15	0.18
IL	0.031	2.4×10^{-4}	4.0×10^{-5}	0	0.43	0.47
MA	0.032	3.8×10^{-4}	9.0×10^{-5}	0	0.36	0.42
MN	0.045	5.2×10^{-4}	2.1×10^{-4}	0	0.36	0.42
NY	0.027	9.0×10^{-5}	4.0×10^{-5}	0	0.48	0.50
OH	0.037	2.9×10^{-4}	7.0×10^{-5}	0	0.45	0.51
PA	0.041	2.9×10^{-4}	7.0×10^{-5}	0	0.49	0.52
TN	0.037	4.3×10^{-4}	1.9×10^{-4}	0	0.36	0.43
WA	0.033	4.1×10^{-4}	2.6×10^{-4}	0	0.35	0.41
WI	0.039	4.6×10^{-4}	2.0×10^{-4}	0	0.36	0.43
Average	0.038	5.7×10^{-4}	3.0×10^{-4}	0	0.38	0.43
St. Dev.	0.009	7.1×10^{-4}	5.5×10^{-4}	0	0.10	0.10

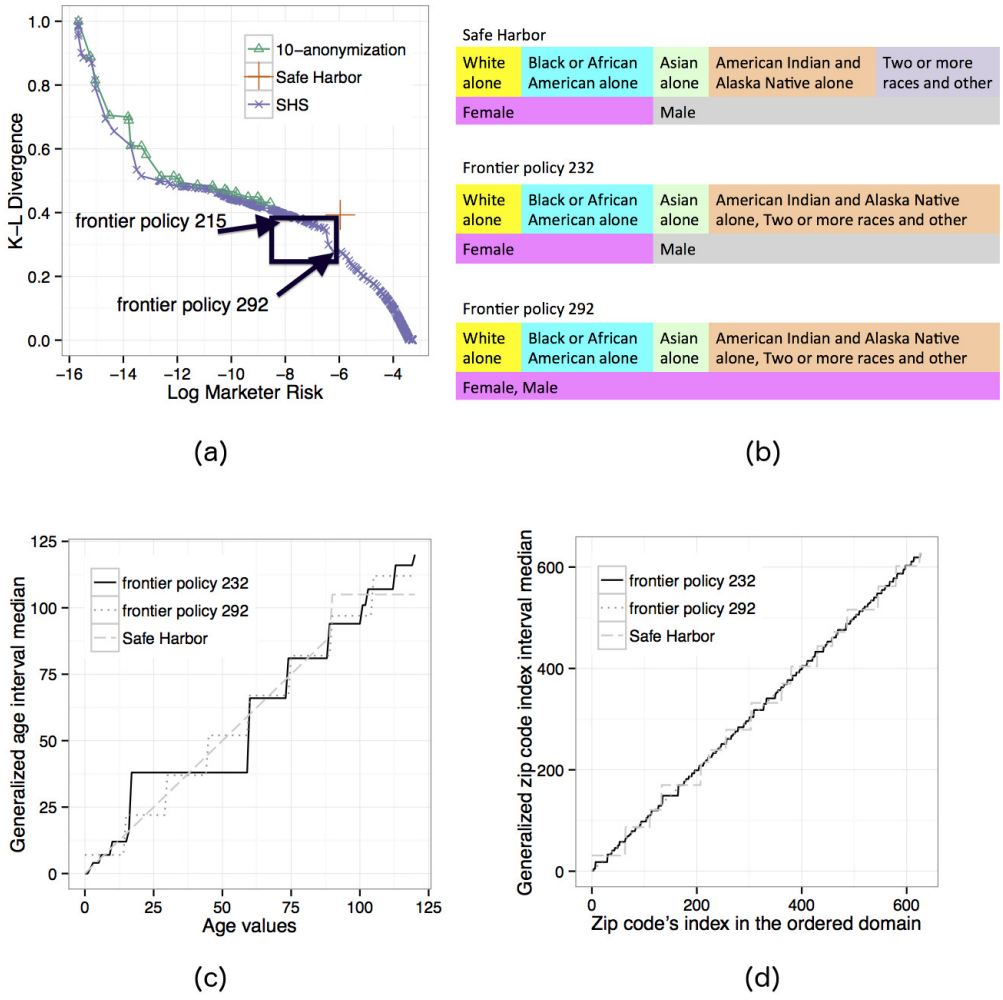


Figure 5.7: Results from the case study for the Adult-TN dataset. (a) A comparison of the 10-anonymization frontier, Safe Harbor policy, and SHS frontier in the R-U space. The policies between the 215^{th} and the 292^{th} on the SHS frontier (in the rectangle) dominate Safe Harbor. (b)-(d) provide a comparison of Safe Harbor and two dominating policies - 232 and 292. (b) A comparison of the generalization rules for race and gender attributes. (c) A comparison of the age generalization rule. The x-axis corresponds to the original age, while the y-axis corresponds to the median of the generalized age interval. (d) A comparison of the ZIP generalization rule. The x-axis corresponds to the original ZIP, while the y-axis corresponds to the median of the ZIP interval. The ZIP codes are represented as an ordinal index, the translation for which can be found in Appendix J of [1].

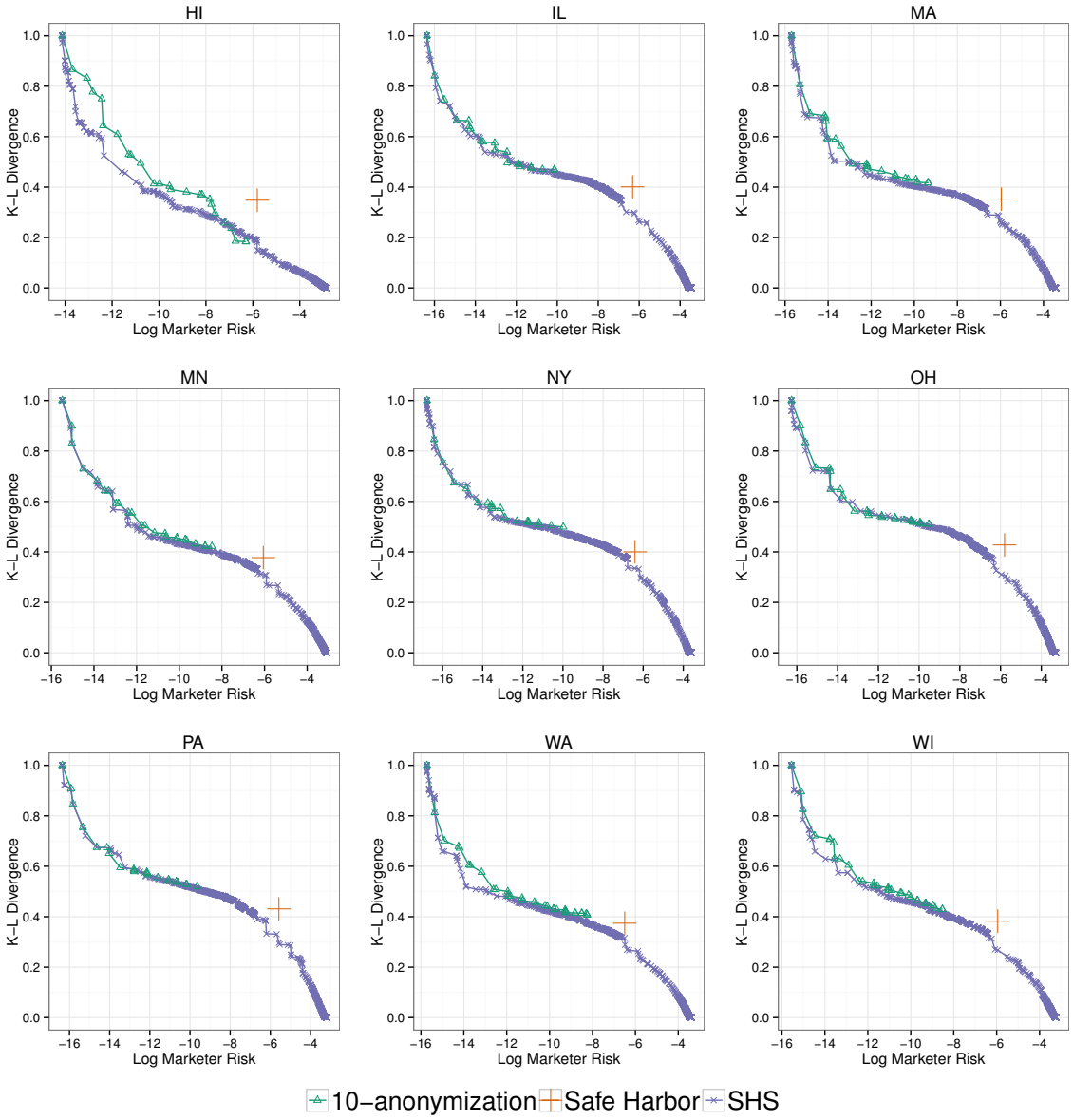


Figure 5.8: A comparison of the 10-anonymization frontier, the Safe Harbor policy, and the SHS frontier in the R-U space for the Adult dataset simulated over nine U.S. states.

Table 5.4: Frontier R-U tradeoff improvement rate of the SHS over k-anonymization (IR) for the Adult dataset with ZIP codes simulated based on U.S. census data

STATE	IR	
	$k = 5$	$k = 10$
HI	0.054	0.050
IL	0.035	0.018
MA	0.058	0.062
MN	0.020	0.049
NY	0.034	0.034
OH	-0.003	0.005
PA	0.018	0.019
TN	0.018	0.050
WA	0.040	0.060
WI	0.026	0.050
Average	0.030	0.040
St. Dev.	0.017	0.019

5.7 Discussion and Conclusions

Organizations that must publish person-specific data for secondary use applications need to make a tradeoff between privacy risks (R) and utility (U). To provide a guideline for data publishers to make this tradeoff, we 1) added a semantic utility metric to an alternative de-identification policy discovery framework, 2) mapped each policy to a two-dimensional R-U space, and 3) formalized the frontier search problem. To solve the problem, we build a set of policies that define a frontier in the R-U space through a heuristic search with a probabilistic basis. We demonstrated that our approach dominates a baseline approach in terms of the quality of the frontier obtained within a fixed number of searched policies.

The empirical analysis results illustrate that the SHS framework is, under many conditions, superior to existing one-size-fits-all policies often invoked in practice (e.g., HIPAA Safe Harbor). We wish to highlight that our empirical analysis was performed over a range of diverse population distributions from 10 U.S. states to mitigate biases in the results. We believe that the SHS strategy has the potential to be a method that overcomes the limitations of a single fixed rule-based policy while being interpretable to health data managers. A healthcare organization, for instance, could present the policy frontier as a documented

method [55] to an Institutional Review Board or legal counsel to justify its selection of a certain degree of protection when sharing data in a de-identified manner.

At the same time, there are several limitations of this work. First, the policy lattice is constructed under the assumption that the set of quasi-identifying attributes is known to the data publisher as a priori. We believe, however, there are several possible ways by which our method could be extended to address this problem. One potential strategy is to construct a policy lattice of the superset of all the possible quasi-identifier sets of attributes and measure re-identification risk as a weighted sum of the risk associated with each potential quasi-identifier. The weight of each quasi-identifier could be dependent on the availability of the corresponding external data resources. An alternative strategy is to construct a policy lattice for every subset with a size no greater than a threshold of the set of all the possible quasi-identifying attributes and search for a policy frontier in each space. Applying the latter would require a strategy to reconcile the de-identification policy associated with attributes that are in the overlapping part of multiple policy spaces (e.g., age, if [Age, Zip, Gender] and [Age, Gender, Race] are both possible quasi-identifiers).

Second, our search strategy does not cover the entire policy space. As such, the frontier is not guaranteed to be optimal. SHS is based on several heuristics and it is possible that more effective approaches could be developed. It may also be possible to develop methods to more systematically and efficiently navigate the space of policies using advanced pruning strategies, such as cost bounding. Moreover, the lattice search process should be amenable to parallel computing techniques as has recently been achieved for k-anonymization [136] provided an appropriate master program that minimizes reassessment of sections of the lattice can be designed.

Third, our investigation is based upon specific measures of risk and utility. In particular, we rely on the marketer risk model, which amortizes the risk over all records in a published dataset. Yet, this is only one way to define risk. The amortization model itself, for instance, can be refined to allow for a discounting function that applies greater weight to

individuals in smaller groups. Beyond the risk model, one could also consider worst-case re-identification scenarios, such as prosecutor or journalist attacks (which state that the risk of a dataset is equal to that of the most risky record) [137]. From the perspective of utility, it is important to recognize that we adopted a generic information loss measure, which was based on the assumption that the specific usage of the dataset is unknown a priori. The data utility function is not necessarily consistent with the usage of the dataset in certain clinical data mining or statistical analysis applications. Nonetheless, if it is known that the dataset will be used in a certain study, then the frontier policy search framework can be customized with an alternative utility function defined by domain experts, provided that the function satisfies the monotonicity requirements of our framework.

Finally, while SHS builds a better frontier than other methods, it can yield a very large number of policies. A data manager would still need to determine which policy is best and it is clear that they could not review every policy on the frontier. As such, a strategy to present the most interesting policy options should be devised.

Chapter 6

Conclusion

This dissertation investigated privacy risk assessment and management theories and technologies for publishing de-identified person-specific information in a manner that balances the tradeoff between data utility and data privacy.

This dissertation developed a principled approach to model the re-identification risk for records in a de-identified dataset of personal information. Most significantly, it developed a generic re-identification process risk model that captures an adversary's behavioral pattern in an environment with limited available resources where we use Markov decision process to represent the adversary's decision making process. The model obtains the adversary's action by solving the optimal decision at each state of the Markov decision process, as apposed to existing methods which make arbitrary assumptions about the adversary's action. This dissertation also developed efficient algorithms to find a solution in a timely fashion.

To operationalize such a re-identification process model to evaluate the re-identification risk in practice, we tackled the problem of assessing the penalty that an adversary would anticipate to be imposed if the privacy attack is caught by the authorities. A novel study of the temporal penalty adopted by a popular data sharing platform was reported in this dissertation. The temporal penalty suspends an user from accessing the dbGaP data for a period of time under the assumption that time diminishes the value of the data. This dissertation investigated this assumption of data value using the impact factors of publications associated with genomic datasets over time, the results suggest there is no evidence of such a dependency.

Finally, provided there is a way to evaluate privacy disclosure risk and an utility function, we defined the de-identification policy frontier discovery (DPFD) problem and developed several heuristic based algorithms to efficiently navigate the solution space to con-

struct a set of solutions with desirable R-U tradeoff rapidly. This part of the dissertation added to the technologies that assist automatically making optimal tradeoff between risk and utility in sharing de-identified personal data.

At this point, we take a moment to highlight several directions for future work.

First, as mentioned in Chapter 3, the re-identification process model is a highly abstract and simplified representation of the actual process. Among the limitations of this process model, we would like to emphasize that it only accounts for one external resource to mount an attack, while in practice several resources are usually leveraged by an adversary to make an intelligent guess and further confirm that guess with a certain level of confidence. A future direction would be to extend the re-identification process model to allow users to explore different external resources. The cost, population covered, attributes covered of these resources may affect the adversary's actions on the orders in which these resources will be exploited. Moreover, the way to access these external resources might also change the attacker's decision making process. For instance, imagine that the entire external dataset is available for downloading after the adversary purchased it. Then the adversary would need to make the decision on whether or not to purchase the dataset based on its meta-information. On the other hand, if the external resource only allows for queries and has different pricing strategies (e.g., unlimited access account or pay for each record), the adversaries will need to decide on the way in which they want to pay for the resource and the queries they want to issue. Accounting for these considerations will bring the process model a step closer to what is happening in practice, even though the cost is that it may cause a state space explosion for the Markov decision process model. More intelligent solvers will need to be developed to efficiently solve the model in a reasonable amount of time.

Second, beyond the limitations of the regression analysis on the temporal penalty of the NCBI dbGaP and the future studies on how to overcome these limitations as presented in Chapter 4, we recognize that being suspended from accessing a database resource has

a different level of impact on different data users (i.e., the potential adversaries) under different circumstances. In other words, the value the data users can obtain from not being suspended from accessing a database resource can be determined by a set of factors, such as the type of study that will be conducted on the data, the investigators' former experience and expertise. This observation inspires a future study on the factors that influence the value of a dataset for secondary use. We anticipate the existing dataset will need to be enriched with new features to conduct this study. The regression analysis based approach can still play an important role given that generalized linear regression models, such as hierarchical linear regression are used. Another phenomenon that caught our attention in particular when conducting the analysis on the temporal penalty of dbGaP is that over half of the publications aggregate multiple datasets to form a large cohort to conduct study. As we mentioned Chapter 4, the approach described in this work is insufficient to model the value of each dataset when it is a subset of the big cohort, from which the publication is generated. Thus, new models need to be defined and developed to answer questions related to the value of a dataset when aggregated with other datasets for analysis.

Third, Risk-Utility solution space for data publishing considered in this dissertation is limited to searching for frontier de-identification policies in the lattice space as described in Chapter 5. The de-identification policy only uses data perturbation to minimize the re-identification risk. However, as our privacy risk process model demonstrated, the data publisher can also force the data users to sign a data use certificate and penalize users, who violate the terms in the data use certificate. Thus, the data publisher's solution space needs to account for additional options, such as a data use certificate, penalization strategies and different levels of penalty in addition to different ways to manipulate the personal data. These extended solutions might not be representable using the lattice structure. New structures can be developed to represent the extended space, but new optimization algorithms will need to be developed to accommodate the extended solution space.

BIBLIOGRAPHY

[1] Weiyi Xia, Raymond Heatherly, Xiaofeng Ding, Jiuyong Li, and Bradley A Malin. R-U policy frontiers for health data de-identification. *Journal of the American Medical Informatics Association*, 22(5):1029–1041, 2015.

[2] Anandhi Bharadwaj, Omar A. El Sawy, Paul A. Pavlou, and N. Venkatraman. Digital business strategy: Toward a next generation of insights. *MIS Quarterly*, 37(2):471–482, June 2013.

[3] Ohbyung Kwon, Namyeon Lee, and Bongsik Shin. Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, 34(3):387–394, 2014.

[4] Leanne Roderick. Discipline and power in the digital age: the case of the US consumer data broker. *Critical Sociology*, 40(5):729–746, 2014.

[5] Kenney Ng, Amol Ghoting, Steven R Steinhubl, Walter F Stewart, Bradley Malin, and Jimeng Sun. PARAMO: A parallel predictive modeling platform for healthcare analytic research using electronic health records. *Journal of biomedical informatics*, 48:160–170, 04 2014.

[6] Andrew R. Post, Tahsin Kurc, Sharath Cholleti, Jingjing Gao, Xia Lin, William Bornstein, Dedra Cantrell, David Levine, Sam Hohmann, and Joel H. Saltz. The analytic information warehouse (aiw): A platform for analytics using electronic health record data. *Journal of Biomedical Informatics*, 46(3):410–424, 2016/02/11.

[7] Elizabeth A McGlynn, Tracy A Lieu, Mary L Durham, Alan Bauck, Reesa Laws, Alan S Go, Jersey Chen, Heather Spencer Feigelson, Douglas A Corley, Deborah Rohm Young, Andrew F Nelson, Arthur J Davidson, Leo S Morales, and

Michael G Kahn. Developing a data infrastructure for a learning health system: the portal network. *Journal of the American Medical Informatics Association*, 21(4):596–601, 07 2014.

[8] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, Melissa A Basford, David S Carrell, Peggy L Peissig, Abel N Kho, Jennifer A Pacheco, Luke V Rasmussen, David R Crosslin, Paul K Crane, Jyotishman Pathak, Suzette J Bielinski, Sarah A Pendergrass, Hua Xu, Lucia A Hindorff, Rongling Li, Teri A Manolio, Christopher G Chute, Rex L Chisholm, Eric B Larson, Gail P Jarvik, Murray H Brilliant, Catherine A McCarty, Iftikhar J Kullo, Jonathan L Haines, Dana C Crawford, Daniel R Masys, and Dan M Roden. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, 31(12):1102–1110, 12 2013.

[9] Katherine M Newton, Peggy L Peissig, Abel Ngo Kho, Suzette J Bielinski, Richard L Berg, Vidhu Choudhary, Melissa Basford, Christopher G Chute, Iftikhar J Kullo, Rongling Li, Jennifer A Pacheco, Luke V Rasmussen, Leslie Spangler, and Joshua C Denny. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*, 20(e1):e147–e154, 06 2013.

[10] Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 20(e2):e206–e211, 12 2013.

[11] Nitesh V Chawla and Darcy A Davis. Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(Suppl 3):660–665, 09 2013.

[12] Jonathan S Schildcrout, Joshua C Denny, Erica Bowton, William Gregg, Jill M Pulley, Melissa A Basford, James D Cowan, Hua Xu, Andrea H Ramirez, Dana C Crawford, Marylyn D Ritchie, Josh F Peterson, Daniel R Masys, Russell A Wilke, and Dan M Roden. Optimizing drug outcomes through pharmacogenetics: A case for preemptive genotyping. *Clinical pharmacology and therapeutics*, 92(2):235–242, 08 2012.

[13] National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies. NOT-OD-07-088, Aug 2002.

[14] Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A. Oniki, Les Westberg, Calvin E. Beebe, Cui Tao, Craig G. Parker, Peter J. Haug, Stanley M. Huff, and Christopher G. Chute. Building a robust, scalable and standards-driven infrastructure for secondary use of ehr data: The sharpn project. *Journal of Biomedical Informatics*, 45(4):763–771, 2016/02/11.

[15] Peter Arzberger, Peter Schroeder, Anne Beaulieu, Geof Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir, and Paul Wouters. An international framework to promote access to data. *Science*, 303(5665):1777–1778, 2004.

[16] Iain Chalmers, Douglas G Altman, Hazel McHaffie, Nancy Owens, and Richard WI Cooke. Data sharing among data monitoring committees and responsibilities to patients and science. *Trials*, 14:102, 2013.

[17] Paul P. Tallon. Corporate governance of big data: perspectives on value, risk, and cost. *IEEE Computer*, 46(6):32–38, 2013.

[18] DANIEL J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–560, 2006.

[19] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine and Ethics*, 25:98–110, 1997.

[20] Michael Barbaro and Tom Zeller. A face is exposed for AOL searcher no. 4417749. *New York Times*, Aug 9, 2006.

[21] Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[22] Latanya Sweeney. Uniqueness of simple demographics in the U.S. population. *Technical report, Carnegie Mellon University*, 2000.

[23] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II*, pages 1–12. Springer, 2006.

[24] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[25] Fida Kamal Dankar and Khaled El Emam. The application of differential privacy to health data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 158–166, New York, NY, USA, 2012. ACM.

[26] US Department of Health and Human Services Office for Civil Rights. Standards for privacy and individually identifiable health information. final rule. *Federal Register*, 67(157):53181, Aug 2002.

[27] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Wright Clayton, Murat Kantarcioglu, Ranjit Ganta, Raymond Heatherly, and Bradley A Malin. A game theoretic framework for analyzing re-identification risk. *PLoS ONE*, 10(3):e0120592, 2015.

[28] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PLoS ONE*, 6(12):e28071, 2011.

[29] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[30] 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), 2010.

[31] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. *l*-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2007.

[32] Dario Freni, Carmen Ruiz Vicente, Sergio Mascetti, et al. Preserving location and absence privacy in geo-social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 309–318, 2010.

[33] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 665–676, 2007.

[34] Luca Bonomi and Li Xiong. A two-phase algorithm for mining sequential patterns with differential privacy. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, pages 269–278, 2013.

[35] Cynthia Dwork. The promise of differential privacy: A tutorial on algorithmic techniques. In *Proceedings of the 52nd IEEE Annual Symposium on Foundations of Computer Science*, pages 1–12, 2011.

[36] Liyue Fan and Li Xiong. Real-time aggregate monitoring with differential privacy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2169–2173, 2012.

[37] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995.

[38] Latanya Sweeney. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[39] Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy. *Natural Reviews Genetics*, 15(6):409–421, 06 2014.

[40] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[41] Stefan Bender, Ruth Brand, and Johann Bacher. Re-identifying register data by survey data: An empirical study. *Statistical Journal of the United Nations ECE*, 18:373–381, 2001.

[42] Mark Elliot and Kingsley Purdam. CASC project deliverable 5D3: TM1 - the evaluation of risk from identification attempts. *CASC Project Computational Aspects of Statistical Confidentiality*, 2003.

[43] Salvador Ochoa, Jamie Rasmussen, Christine Robson, and Michael Salib. Reidentification of individuals in Chicagos homicide database: A technical and legal study. *Massachusetts Institute of Technology*, 2001.

[44] Peter K. Kwok, Michael Davern, Elizabeth C. Hair, and Deborah Lafky. Harder

than you think: a case study of re-identification risk of hipaa-compliant records. *2011 Joint Statistical Meetings*, 2011.

[45] Mandl KD. Brownstein JS, Cassa CA. No place to hidereverse identification of patients from published maps. *The New England Journal of Medicine*, 355:17411742, 2006.

[46] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125, Washington, DC, USA, 2008. IEEE Computer Society.

[47] Khaled El Emam and Patricia Kosseim. Privacy interests in prescription data, part 2: Patient privacy. *IEEE Security Privacy*, 7(2):75–78, March 2009.

[48] Dan Frankowski, Dan Cosley, Shilad Sen, Loren Terveen, and John Riedl. You are what you say: Privacy risks of public mentions. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 565–572, New York, NY, USA, 2006. ACM.

[49] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web*, pages 181–190, New York, NY, USA, 2007. ACM.

[50] The Supreme Court of the State of Illinois. Southern illinoisan vs. the illinois department of public health. 2006.

[51] Federal Court: Canada. Mike gordon v. the minister of health and the privacy commissioner of canada: Memorandum of fact and law of the privacy commissioner of canada. *Federal Court*, 2007.

[52] Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, Natalia Popova, Stephanie Pretel, Lora Ziyabari, Moira Lee, Yu Shao, Zhen Y Wang, Karl Sirotkin, Minghong Ward, Michael Kholodov, Kerry Zbicz, Jeffrey Beck, Michael Kimelman, Sergey Shevelev, Don Preuss, Eugene Yaschenko, Alan Graeff, James Ostell, and Stephen T Sherry. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10):1181–1186, 10 2007.

[53] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 06 2007.

[54] Deven McGraw. Building public trust in uses of health insurance portability and accountability act de-identified data. *Journal of the American Medical Informatics Association*, pages 29–34, 2013.

[55] U.S. Dept.of Health and Human Services. Standards for privacy of individually identifiable health information, final rule. pages pt 160–164, 2002.

[56] Kathleen Benitez and Bradley Malin. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*, 17:169–177, 2010.

[57] Latanya Sweeney. *Computational Disclosure Control A Primer on Data Privacy Protection*. PhD thesis, Massachussetts Institute of Technology, May 2001.

[58] Tore Dalenius. Finding a needle in a haystack or identifying anonymous census records. *Journal of Official Statistics*, 2:329–336, 1986.

[59] Latanya Sweeney. Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10:571–588, 2002.

[60] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 279–288, New York, NY, USA, 2002. ACM.

[61] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Top-down specialization for information and privacy preservation. In *Proceedings of the 21st International Conference on Data Engineering*, pages 205–216, 2005.

[62] Tiancheng Li and Ninghui Li. Towards optimal k-anonymization. *Data and Knowledge Engineering*, 65(1):22–39, 2008.

[63] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional *k*-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Eng*, pages 25–25, 2006.

[64] Gagan Aggarwal, Tomás Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Anonymizing tables. In *Proceedings of the 10th International Conference on Database Theory*, pages 246–258, 2005.

[65] Adam Meyerson and Ryan Williams. On the complexity of optimal *k*-anonymity. In *Proceedings of the 23rd ACM Symp. on Principles of Database Systems*, pages 223–228, 2004.

[66] Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. *The Anonymisation Decision Making Framework*. UK Anonymisation Network (UKAN), United Kingdom, 7 2016.

[67] Mark Elliot and Angela Dale. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, 14:6-10, 1999.

[68] Diane Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9:313–331, 1993.

[69] Elaine Mackey and Mark Elliot. An application of game theory to understanding statistical disclosure events. *UNECE/Eurostat Work session on data confidentiality*, 2009.

[70] Muqun Li, David Carrell, John Aberdeen, Lynette Hirschman, and Bradley A. Malin. De-identification of clinical narratives through writing complexity measures. *International Journal of Medical Informatics*, 83(10):750–767, 2016/12/04.

[71] Richard Bellman. A markovian decision process. *Indiana University Mathematics Journal*, 6(4):679–684, 1957.

[72] Ronald A. Howard. *Dynamic Programming and Markov Processes*. Technology Press of Massachusetts Institute of Technology, Cambridge, MA, USA, 1960.

[73] Joshua Letchford and Yevgeniy Vorobeychik. Optimal interdiction of attack plans. In *Proceedings of the International Conference on Autonomous Agents and Multi-agent Systems*, pages 199–206, 2013.

[74] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[75] Geoffrey J. Gordon. Stable function approximation in dynamic programming. Technical report, Pittsburgh, PA, USA, 1995.

[76] John N. Tsitsiklis and Benjamin van Roy. Feature-based methods for large scale dynamic programming. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 1, pages 565–567 vol.1, Dec 1995.

[77] Craig Boutilier and Richard Dearden. Approximating value trees in structured dynamic programming. In *Proceedings of the 13th International Conference on Machine Learning*, pages 54–62, 1996.

[78] Robert St-aubin, Jesse Hoey, and Craig Boutilier. Apricodd: Approximate policy construction using decision diagrams. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 1089–1095, 2000.

[79] Dimitri P. Bertsekas and John Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, USA, 1996.

[80] Paul J Schweitzer and Abraham Seidmann. Generalized polynomial approximations in markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110(2):568–582, 1985.

[81] Daniela Pucci de Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

[82] Daniela Pucci de Farias and Benjamin Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics OF Operations Research*, 29(3):462–478, 2004.

[83] Dale Schuurmans and Relu Patrascu. Direct value-approximation for factored mdps. In *Proceedings of the 2001 Conference on Advances in Neural Information Processing Systems 14*, pages 1579–1586. MIT Press, 2001.

[84] Craig Boutilier, Richard Dearden, and Moiss Goldszmidt. Exploiting structure in policy construction. In *Proceedings of the 1995 International Joint Conference on AI*, pages 1104–1111.

[85] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Computing Research Repository (CoRR)*, abs/1106.1822, 2011.

[86] Jesse Hoey, Robert St-aubin, Alan Hu, and Craig Boutilier. Spudd: Stochastic plan-

ning using decision diagrams. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 279–288. Morgan Kaufmann, 1999.

[87] Kee-Eung Kim and Thomas Dean. Solving factored mdps using non-homogeneous partitions, 2003.

[88] Thomas Dean and Robert Givan. Model minimization in markov decision processes. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pages 106–111, 1997.

[89] Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas Dean, and Craig Boutilier. Solving very large weakly coupled markov decision processes. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 165–172, 1998.

[90] Satinder Singh and David Cohn. How to dynamically merge markov decision processes. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, pages 1057–1063, Cambridge, MA, USA, 1997. MIT Press.

[91] Reza Shokri. Optimal user-centric data obfuscation. *Technical report, ETH Zurich*, 2014.

[92] Donna E. Shalala. Address at the national press club, Washington, D.C. July 31, 1997.

[93] Edward E. Bartlett. RMs need to safeguard computerized patient records to protect hospitals. *Hospital Risk Management*, (9):129–140, September, 1993.

[94] David F. Linowes and Ray C. Spencer. Privacy: The workplace issue of the '90s. *John Marshall Law Review*, (23):591–620, 1990.

[95] Robert Neil Butler. Who's reading your medical records. *Geriatrics*, 52:7–8, 1997.

[96] L. Jean Camp and M. Eric Johnson. *The Economics of Financial and Medical Identity Theft*. Springer-Verlag New York, 2012.

[97] Alessandro Acquisti, Leslie John, and George Loewenstein. What is privacy worth. In *In Workshop on Information Systems and Economics*, 2009.

[98] Christina Aperjis and Bernardo A. Huberman. A market for unbiased private data: Paying individuals according to their privacy attitudes. *Computing Research Repository (CoRR)*, abs/1205.0030, 2012.

[99] Yiling Chen, Stephen Chong, Ian A. Kash, Tal Moran, and Salil Vadhan. Truthful mechanisms for agents that value privacy. In *Proceedings of the 14th ACM Conference on Electronic Commerce*, pages 215–232, New York, NY, USA, 2013. ACM.

[100] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, pages 199–208, New York, NY, USA, 2011. ACM.

[101] Kenneth C. Laudon. Markets and privacy. *Communications of ACM*, 39(9):92–104, 1996.

[102] Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 774–789, New York, NY, USA, 2012. ACM.

[103] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–92, June 2016.

[104] Richard A. Posner. The economics of privacy. *The American Economic Review*, 71(2):405–409, 1981.

[105] Christopher Riederer, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, and Pablo Rodriguez. For sale : Your data: By : You. In *Proceedings of the*

10th ACM Workshop on Hot Topics in Networks, HotNets-X, pages 13:1–13:6, New York, NY, USA, 2011. ACM.

[106] Aaron Roth. Buying private data at auction: The sensitive surveyor's problem. *ACM SIGecom Exchanges*, 11(1):1–8, 2012.

[107] Aaron Roth and Grant Schoenebeck. Conducting truthful surveys, cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 826–843, New York, NY, USA, 2012. ACM.

[108] George J. Stigler. An introduction to privacy in economics and politics. *The Journal of Legal Studies*, 9(4):623–644, 1980.

[109] Curtis R. Taylor. Consumer privacy and the market for customer information. *The RAND Journal of Economics*, 35(4):631–650, 2004.

[110] Alessandro Acquisti and Hal R. Varian. Conditioning prices on purchase history. *SIMS Working Paper*, 2002.

[111] L. Jean Camp and Catherine D. Wolfram. Pricing security: Vulnerabilities as externalities. *Economics of Information Security*, (12), 2004.

[112] Stuart Schechter. Quantitatively differentiating system security. In *Proceedings of the 1st Workshop on Economics and Information Security*, pages 16–17, 2002.

[113] Jelke G. Bethlehem, Wouter J. Keller, and Jeroen Pannekoek. Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409):38–45, 1990.

[114] Roberto Benedetti, A. Capobianchi, and L. Franconi. Individual risk of disclosure using sampling design information. *Contributi Istat 1412003*, 1998.

[115] Chris Skinner and David J. Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, (14):361–372, 1998.

[116] Elsayed A.H. Elamir and Chris Skinner. Record-level measures of disclosure risk for survey microdata. *Journal of Official Statistics*, (22):525–539, 2006.

[117] Yosef Rinott and Natalie Shlomo. A neighborhood regresion model for sample disclosure risk estimation. *Joint UNECE/Eurostat work session on statistical data confidentiality*, 2005.

[118] Yosef Rinott and Natalie Shlomo. A smoothing model for sample disclosure risk estimation. *IMS Lecture Notes Monograph Series*, (54):161–171, 2007.

[119] Adam Tanner. Harvard professor re-identifies anonymous volunteers in DNA study. *Forbes*, 2013.

[120] (Last accessed Jan 27, 2014) North Carolina Voter Registration Database, ftp://www.app.sboe.state.nc.us/data. Last accessed 27 Jan 2014.

[121] National Institutes of Health. Genomic data sharing policy. August 27, 2014.

[122] William W. Lowrance and Francis S. Collins. Identifiability in genomic research. *Science*, 317(5838):600–602, 2007.

[123] Eugene Garfield. The history and meaning of the journal impact factor. *JAMA*, 295(1):90–93, 2006.

[124] Jevin D. West, Theodore C. Bergstrom, and Carl T. Bergstrom. The eigenfactor metricstm: A network approach to assessing scholarly journals. *College and Research Libraries*, 71(3):236–244, 2010.

[125] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[126] The National Institutes of Health Genomic Data Sharing Governance Committees. Data use under the nih gwas data sharing policy and future directions. *Nature Genetics*, 46(9):934–938, 09 2014.

[127] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of "personally identifiable information". *Communications of the ACM*, 53(6):24–26, 2010.

[128] Kathleen Benitez, Grigorios Loukides, and Bradley Malin. Beyond safe harbor: automatic discovery of health information de-identification policy alternatives. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 163–172, 2010.

[129] Khaled El Emam, Luk Arbuckle, Gunes Koru, Benjamin Eze, Lisa Gaudette, Emilio Neri, Sean Rose, Jeremy Howard, and Jonathan Gluck. De-identification methods for open health data: The case of the heritage health prize claims dataset. *Journal of Medical Internet Research*, 14(1):e33, Jan-Feb 2012.

[130] Moshe Lichman. UCI machine learning repository, 2013.

[131] U.S. Census Bureau. American fact finder website: http://www.americanfactfinder.gov. 2012.

[132] Omri Gottesman, Helena Kuivaniemi, Gerard Tromp, W. Andrew Faucett, Rongling Li, Teri A. Manolio, Saskia C. Sanderson, Joseph Kannry, Randi Zinberg, Melissa A. Basford, Murray Brilliant, David J. Carey, Rex L. Chisholm, Christopher G. Chute, John J. Connolly, David Crosslin, Joshua C. Denny, Carlos J. Gallego, Jonathan L. Haines, Hakon Hakonarson, John Harley, Gail P. Jarvik, Isaac Kohane, Iftikhar J. Kullo, Eric B. Larson, Catherine McCarty, Marylyn D. Ritchie, Dan M. Roden, Maureen E. Smith, Erwin P. Bottinger, and Marc S. Williams. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genetics in Medicine*, 15(10):761–771, 10 2013.

[133] Bradley Malin, Kathleen Benitez, and Daniel Masys. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*, 18(1):3–10, 2011.

[134] Traian Marius Truta, Farshad Fotouhi, and Daniel Barth-Jones. Disclosure risk measures for microdata. In *Proceedings of the 15th International Conference on Scientific and Statistical Database Management*, pages 15–22, 2003.

[135] Daniel Kifer and Johannes Gehrke. Injecting utility into anonymized datasets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 217–228, 2006.

[136] Xuyun Zhang, Chi Yang, Surya Nepal, Chang Liu, Wanchun Dou, and Jinjun Chen. A mapreduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud. In *Proceedings of the 2013 International Conference on Cloud and Green Computing*, pages 105–112, Washington, DC, USA, 2013. IEEE Computer Society.

[137] Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15:627–637, 2008.