Privacy-Preserving Sharing of High-Dimensional Data based on Computational Game Theory

By

Zhiyu Wan

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
December 12, 2020
Nashville, Tennessee

Approved:
Bradley A. Malin, Ph.D.
Daniel Fabbri, Ph.D.
Douglas H. Fisher, Ph.D.
Yevgeniy Vorobeychik, Ph.D.
Murat Kantarcioglu, Ph.D.

DEDICATION

To my parents

ACKNOWLEDGMENTS

First and foremost, I am greatly indebted to my advisor Dr. Bradley Malin for teaching me more than I could ever give him credit for here. He has shown me, by his example, what a great scientist (and person) should be. Most importantly, he is always patient and trust me about my research progress. 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 Washington University in St. Louis 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. I am grateful to Dr. Murat Kantarcioglu at the University of Texas at Dallas, who has kindly given me tremendous academical support to the work that built this dissertation with his sound advice and concrete suggestions.

I would also like to acknowledge Dr. Daniel Fabbri and Dr. Douglas Fisher. Despite their busy schedules, they generously carved out time and provided invaluable expertise and insights on this work.

I thank the wonderful fellow HIPLABers and HIPLAB Alumni: Dr. Weiyi Xia, Dr. Zhijun Yin, Chao Yan, Yongtai Liu, Ziqi Zhang, Jia Guo, Dr. James Hazel, Steve Nyemba, Thomas Brown, Dr. You Chen, Dr. Raymond Heatherly, Dr. Wen Zhang, Dr. Rachel Li, Dr. Wei Xie, Dr. Lina Sulieman, Adarsh Subbaswamy, Dr. Ranjit Ganta, and Grayson Ruhl, all of whom are amazing in creating such a lively and creative research environment where good works and breakthroughs keep emerging.

I thank my parents for their endless love and support. I thank my girlfriend for her accompany and help. I thank all my teachers and professors who have taught me priceless knowledge. I thank all my friends who have shared precious time and memorable moments with me along this journey.

My dissertation research was funded, in part, by the grants U01HG008701, RM1HG009034, R01HG006844, R01LM009989, U01HG8666, U01HG006378, U01HG6379, U01HG006385, U01HG8679, U01HG8680, U01HG8684, U01HG8673, U01HG8701, U01HG8676, U01HG8664, and U01HG004438 from the National Institutes of Health and grant CCF-0424422 from the National Science Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Direct-to-consumer genetic testing (DTC-GT)

United States (US)

Deoxyribonucleic acid (DNA)

Artificial intelligence (AI)

Machine learning (ML)

Information technology (IT)

Sharing, Privacy, Attack, Defense, Optimization, and Game (SPADOG)

Data Sharing Analysis Circle (DSAC)

Data Sharing Privacy Pyramid (DSPP)

Health Insurance Portability and Accountability Act (HIPAA)

Electronic medical records (EMR)

Global Positioning System (GPS)

Single nucleotide polymorphisms (SNPs)

Federal Bureau of Investigation (FBI)

Combined DNA Index System (CODIS)

Genome-wide association studies (GWAS)

Direct-to-consumer (DTC)

Human immunodeficiency virus (HIV)

Acquired immunodeficiency syndrome (AIDS)

Homomorphic encryption (HE)

Secure multi-party computation (SMC)

Database of Genotypes and Phenotypes (dbGaP)

National Institutes of Health (NIH)

Short tandem repeats on Y-chromosome (Y-STRs)

Short tandem repeat (STR)

Log-likelihood ratio (LR)

Wellcome Trust Case Control Consortium (WTCCC)

Shringarpure and Bustamante (SB)

Generative adversarial networks (GAN)

Multi-party computation (MPC)

Software Guard Extensions (SGX)

Phenome-wide association studies (PheWAS)

Domain generalization hierarchy (DGH)

Differential privacy (DP)

Media access control (MAC)

Pure-Strategy Nash Equilibrium (PSNE)

Sequence and Phenotype Integration Exchange (SPHINX)

Precision Medicine Initiative (PMI)

Million Veteran Program (MVP)

Global Alliance for Genomics and Health (GA4GH)

Integrating Data for Analysis, Anonymization, and Sharing (iDASH)

Log-likelihood ratio test (LRT)

Alternative allele frequency (AAF)

Single nucleotide variants (SNVs)

Health and Human Services (HHS)

Information loss (IL)

Generalization intensity (GI)

Backward induction search (BIS)

Lattice-based search (LBS)

Electronic Medical Records and Genomics Pharmacogenomics (eMERGE-PGx)

Backward induction algorithm (BIA)

Genetic algorithm (GA)

Minor allele frequency (MAF)

Random-access memory (RAM)

Augmented SB attack (ASBA)

False positive rate (FPR)

Markov decision process (MDP)

Center for Study of Human Polymorphisms (CEPH)

Sorenson Molecular Genealogy Foundation (SMGF)

Multi-Stage Re-Identification Game (MSRIG)

International Classification of Diseases, Tenth Revision (ICD-10)

Chapter 1

INTRODUCTION

To optimize privacy-preserving data sharing, I built game theoretic models, solved them with the help of artificial intelligence search algorithms, and conducted experiments on genomic datasets.

In this chapter, I introduce the background of my dissertation research and review most works that are related to my research.

1.1 Motivation

Our ability to collect and analyze personal data has grown dramatically over the past decade, a trend that shows no sign of slowing. While this information enables a wide range of institutions to perform novel research in biomedicine and the social sciences, big data has become big business. There is a rapidly expanding market for sharing and selling data for secondary analysis. For instance, in the clinical realm, personal information is routinely stored in electronic health records. The biomedical research domain now supports studies that collect data on a diverse array of participants [1]. And, most recently, the commercial setting has led to a number of ventures where data is collected, such as direct-to-consumer genetic testing (DTC-GT) companies that collect data from various consumers and build repositories that now cover over 10% of the United States (US) population [2]. Many believe that sharing such data beyond their initial point of collection is crucial to maximizing their societal value [3]. However, data sharing efforts are often limited by privacy concerns, particularly over the identifiability of the individuals to whom the data corresponds [4]. Genomic data, which is shared in various settings in the US, provides a clear illustration of the threat and concern. While data managers remove explicit identifiers (e.g., personal names and phone numbers) to adhere to de-identification guidance [5, 6], numerous demonstration attacks have shown that data, and particularly genomic records, can be re-identified through a variety of means [7, 8]. Although institutions and individuals are incentivized to share data [9, 10], they usually lack the ability to identify and assess privacy risks precisely in order to make the optimal sharing decisions.

Game theory, as a branch of applied mathematics, studies the strategic interactions among rational decision makers [11]. With its initial focus on economics, it has influenced many other fields, including political science, biology, psychology, linguistics, and computer science [12]. The founding figure of computing, John von Neumann, created the modern game theory [13] and contributed heavily to the discovery of Deoxyribonucleic acid (DNA), quantum mechanics, and atomic bombs. Shoham [14] surveyed and identified the main areas of interaction between computer science and game theory. Game theory is an integral part of artificial intelligence (AI), theory, e-commerce, networking, and other areas of computer science. One reason is that game theory is by far the most developed theory of the interactions span multiple entities, each with its own information and interests. Computer scientists have taught machines to play games with optimal strategies. Computer or AI has won the best human players in numerous games such as Chess [15], Go [16, 17], Poker [18], and StarCraft II [19]. As long as a problem can be modeled as a game, the search for the solution can rely on algorithms designed by computer scientists. Still, not all real-world problems can be modeled as a game. Game theory has been applied to security and privacy research [20]. However, when applying game theory to the specific area of privacy-preserving data sharing, there are several additional challenges: 1) the adversary's behaviors are difficult to model; 2) privacy breach is difficult to be detected. Two questions remain open even if a real-world problem is successfully modeled as a game: 1) How efficient can the solution of the game be found? 2) How effective can the solution be applied in practice?

With these challenges and questions in mind, I started with the idea of utilizing game theoretic models to help build a better world by solving a practical problem. Eventually, I want to find the answer to the question: How to share data optimally? It might be difficult at first to see the natural connection between data sharing and game theory. Could you imagine what a data privacy game would look like? Now, I use a concrete example to introduce the concept of a data sharing game.

1.2 A Toy Example: The Go-Share-Info Game

Imagine we are playing a variation of the board game *Go*, called the *Go-Share-Info* game, as illustrated in Figure 1.1. This new game uses the same 19x19 board and white and black *stones* as *Go* uses, but with additional rules. The roles of the two players are no longer equal. Player *A*, placing white stones, plays the role of a doctor, who shares information with the assistant (the board) by placing a stone with a message on it. Each message has two parts: the disease and the identifier of a patient, which have already been written on two sides of the stone, respectively. There are 180 messages this player can share by placing corresponding stones on the board. The player can choose not to share a message by erasing one part or two parts of it before placing the corresponding stone on the board. Player *B*, placing black stones, on the other hand, plays an adversary who aims to capture as many patient-disease pairs as possible. Each stone is associated with a score in the range from 1 to 10. At the end of a game, player *A* got all the scores associated with the complete messages left on the board (not been captured), and the player *B* got all the scores associated with the captured stones from player *A* and stones left in player *A*'s bowl, multiplied by a factor of *X*. Factor *X* can be set to 0, 0.5, 1, or 2. Because player *A* knows more information than player *B* knows, the factor is recommended to be set as 2 to balance the two players' difficulties to earn scores. What makes the game fairer is that the winner is determined after a pair of games in which player *A* in the first game will be player *B* in the second game. Whoever gets the highest total score in a pair of games will be the winner. The basic rule of the *Go* remains the same, which is that a stone on the board must have at least one open spot directly orthogonally adjacent to it or must be part of a connected group that has at least one such open spot next to it; Otherwise, the stone will be captured by the opponent. The game starts with the player with one more stone and ends when no stones can be placed on the board by one player. Generally, one stone can be placed by each player during each move. The goal for player *A* is still to minimize the number of stones that will be captured or left in the bowl at the end of the game. However, player *A* can decide to erase parts of the message before placing each stone, which makes the game more complex than the *Go* game. By programming machines to play with itself countless times, engineers may find a winning solution or the optimal strategy for this game.

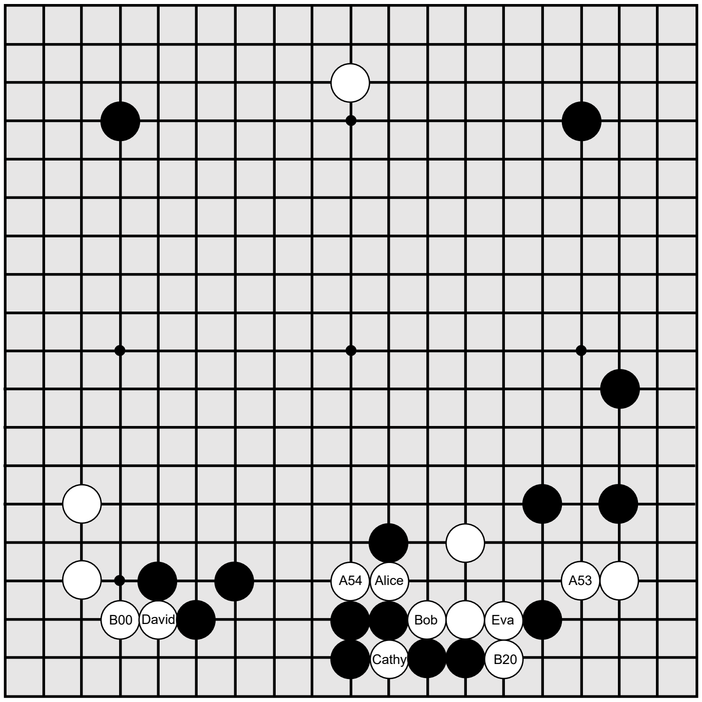


Figure 1.1 An illustration of the Go-Share-Info Game.

A53, A54, B00, B20 are all International Classification of Diseases, Tenth Revision (ICD-10) codes.

1.3 Sharing, Privacy, Attack, Defense, Optimization, and Game (SPADOG) Framework

There have been several works located at the intersection of these two research fields: data privacy and game theory. However, few scholars tried to connect these two fields systematically. Here, I want to pave the avenue from data sharing to game theory and demonstrate that the ultimate answer to the data sharing question will always lead to the game theoretic methodology. To illustrate this idea, I proposed a framework named SPADOG, which represents a combination of data sharing, data privacy, attack, defense, optimization, and game.

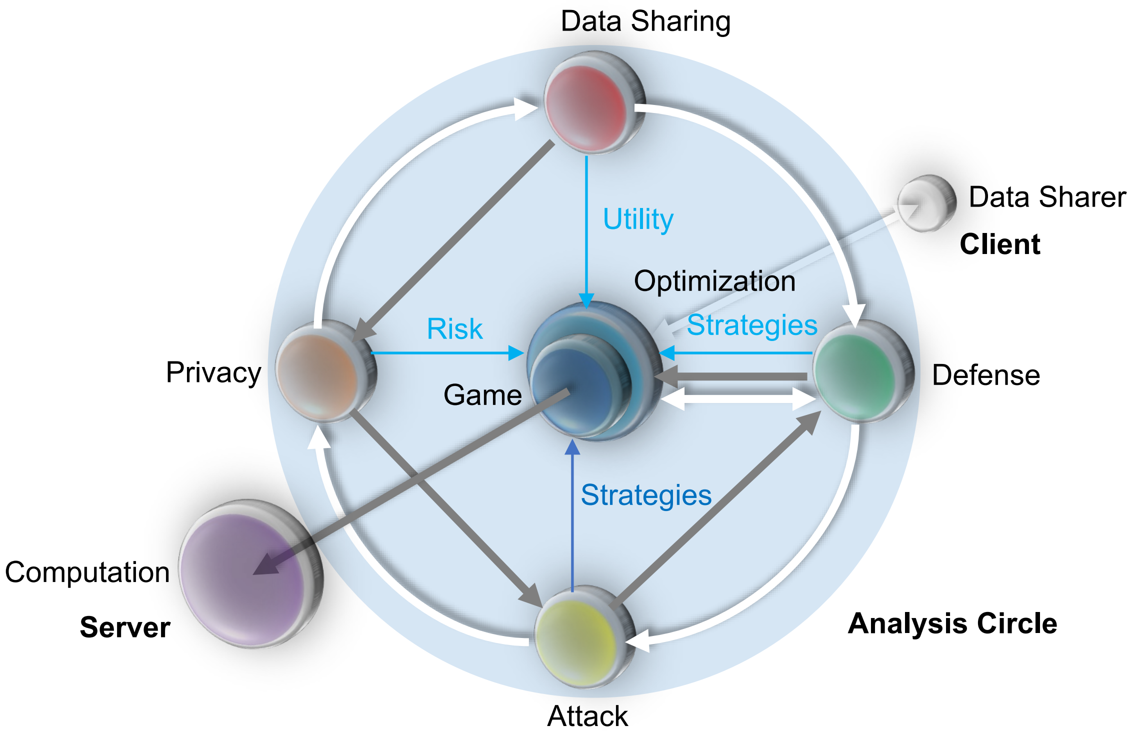


Figure 1.2 Data Sharing Analysis Circle.

Colored plates represent modules in the system. White lines represent data flow. Blue lines represent knowledge flow. Gray lines represent the motivation order.

1.3.1 Data Sharing Analysis Circle (DSAC)

The framework has two representations. The first representation is the Data Sharing Analysis Circle (DSAC), as shown in Figure 1.2. It is a graph model that shows the motivation for each component and information flows in the system.

In Figure 1.2, each plate represents a module of the entire model. There are three types of lines in the figure. The grey lines represent the motivation order of the model. The white lines represent the data flows in the use case. The blue lines represent knowledge flows in the model. The light blue lines are required by the optimization module. At the same time, the light blue lines and dark blue lines are required by the game module.

The DSAC system is represented by a client-server structure, in which the server-side has a backend computing unit and the frontend analysis interface. Each analysis circle is built for a particular problem with a particular type of datasets that need to be shared and a particular attack in consideration. The client needs to choose an analysis circle with a specific utility function, privacy function, attack method, and defense method, and then provide the dataset. After the data has been analyzed by six analysis modules, it will be sent to the backend computing module. At last, the returned result from the server to the client will be the optimal solution.

The analysis circle is constructed as follows:

At first, there is only a data module. When the data module receives the dataset needs to be shared and the given utility function, it returns the utility. Afterward, a privacy module is added. In this case, the privacy module returns the privacy risk based on the dataset and the given privacy function. However, the analysis cycle is built for a particular attack, so an attack module needs to be added to the circle. In this case, after the data module computes the utility, the attack module chooses an attack action according to a set of fixed parameters. Then, the privacy module calculates the privacy loss of the attack. Because the goal of the analysis is to help the data sharer protect the data, a defense module needs to be added to the circle. In this case, before the attack happens, the data is modified according to a protection model with a set of fixed parameters. This time, the returned results include a modified dataset, associated utility, and associated privacy. However, the protection strategy is not guaranteed to be optimal. Then, an optimization module is added, which outputs the optimal protection strategy to the defense module according to the privacy function, the utility function, and the strategy set. However, the attacker's strategy is not guaranteed to be optimal. A game model is then added, which outputs the optimal protection strategy to the defense module by additionally considering the attacker's optimal strategy. However, when the size of the dataset is relatively large, the backward induction algorithm in game theory will be too slow. Then, a computation module is needed, which accelerates the computing. If the global optimum can hardly be found, the best local optimum can be returned. Notice that the order of construction happens to be in the reverse direction of the data flow.

Let us use the toy example to illustrate the construction process of this model. First, player *A* needs to place a stone with information to get scores. A sequence of placed white stones is associated with a fixed sequence of expected utilities. Second, player *A* realizes that white stones being captured would decrease player *A*'s score and increase player *B*'s score. From player *A*'s perspective, a sequence of captured white stones is associated with a fixed sequence of expected privacy losses. Third, player *A* considers the actions player *B* can take to capture stones, including surrounding a group of white stones with black stones. From player *A*'s perspective, a sequence of placed stones with a set of attributes (i.e., location, color, and score) is associated with a sequence of expected utilities (i.e., player *A*'s score) and privacy losses (i.e., player *B*'s score). Fourth, player *A* considers the actions to prevent stones from being captured, including 1) erasing part of the information on a stone before placing them, 2) avoiding any group of white stones to be surrounded by the opponent, and 3) surrounding the opponent's stones. A sequence of placed stones with a set of attributes (i.e., location, color, score, and the number of erased sides) is associated with a fixed sequence of expected utilities and privacy losses. Fifth, player *A* assumes a naïve opponent that adopts a simple strategy. Assuming the location of a black stone depends on the locations and scores of all already placed stones, player *A* can exhaustively search through all the possible sequences of placed stones with a set of attributes and select the sequence with the highest weighted sum of final utility and final privacy. The weighted sum of final utility and final privacy is proportional to the difference between player *A*'s final score and player *B*'s final score at the end of a game. Sixth, player *A* assumes a rational opponent that adopts the optimal strategy. Assuming player *B* will exhaustively search through all possible sequences given placed stones and choose the one with the highest expected difference between player *B*'s final score and player *A*'s final score, player *A* exhaustively searches through all possible sequences and chooses the one with highest difference between player *A*'s final score and player *B*'s final score. This game is a zero-sum game. It can be converted into a minimax optimization problem, and the optimal solution can be found using backward induction with alpha-beta pruning. However, that might not be the best way to find the best solution. Finally, by playing with self almost infinity times, player *A* will find an optimal strategy using a similar approach as used by AlphaGo Zero [SSS+17], which was based on deep reinforcement learning. In situations where there are time and computing power constraints, other heuristics and algorithms can be leveraged to search for a local optimum.

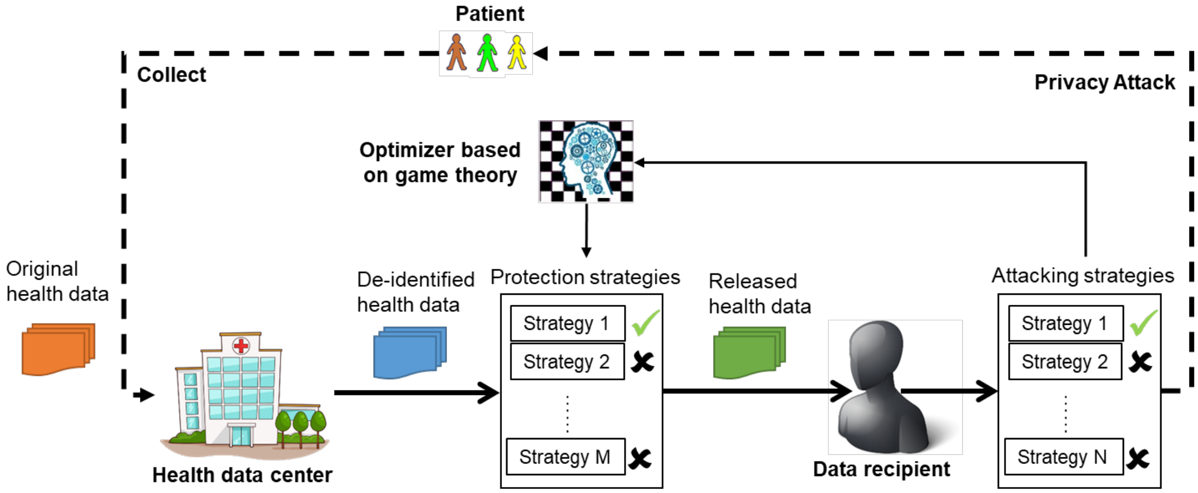


Figure 1.3 Optimal privacy-serving patient data sharing in a healthcare research scenario.

A more realistic example illustrates the working process of the analysis circle is shown in Figure 1.3. In this scenario, a health data center aims to publish patients' data in a privacy-preserving manner to facilitate research. First, patients' data are collected by the health data center. Second, the data center protects patients' data by releasing a de-identified and modified version of it to a group of researchers. One of the recipients is malicious and attacks targeted patients by reidentifying the records and inferring the sensitive information. When there is no optimizer, the patients might detect their privacy is breached and stop sharing data with the health data center. Fortunately, the whole process is simulated by the optimizer based on game theory until an optimal strategy (including not sharing data at all) is found, which prevents the attack while maximizing the shared data's utility to benign recipients.

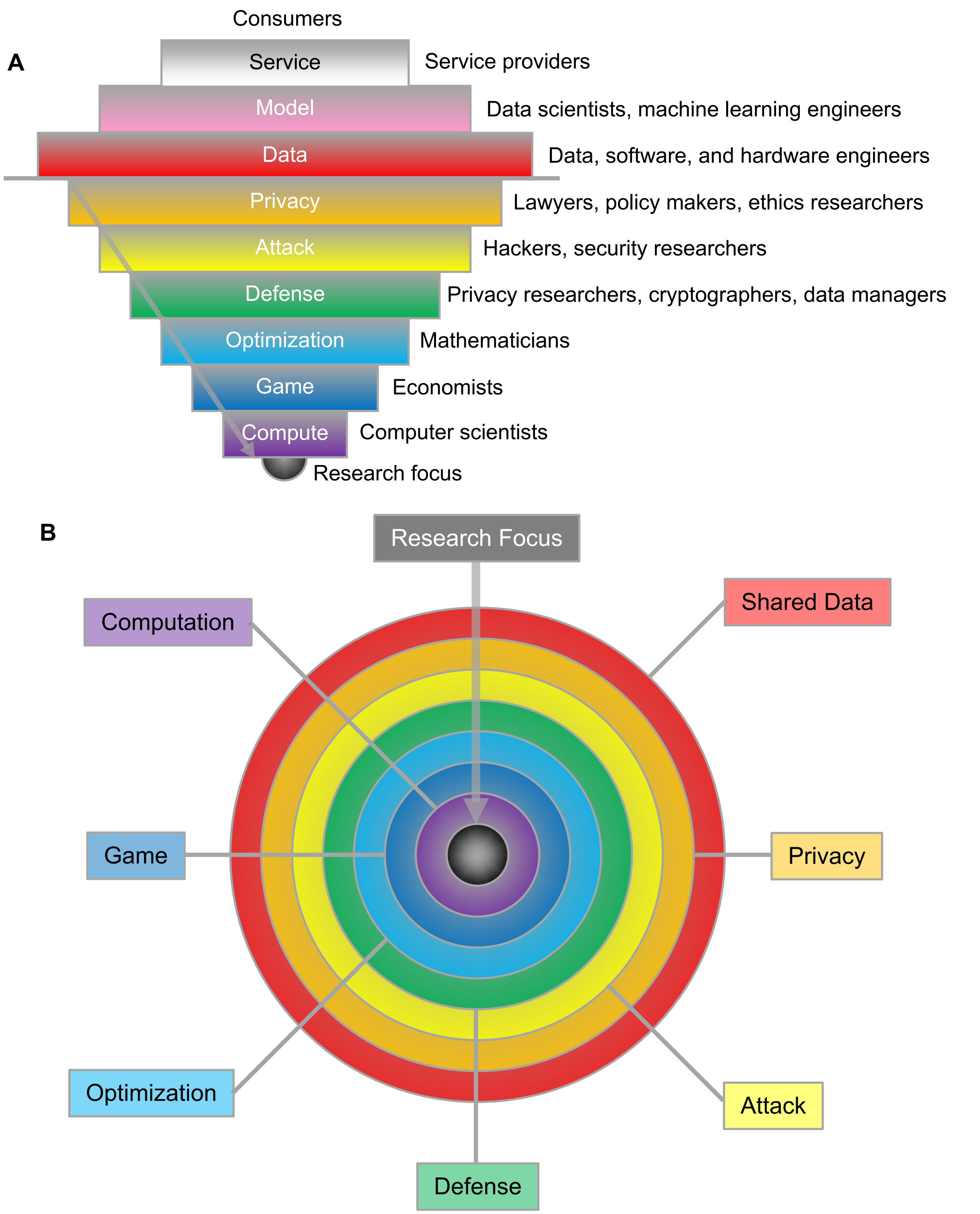


Figure 1.4 Illustration of the Data Sharing Privacy Pyramid (DSPP).

(**A**) Side view of the DSPP. (**B**) Bottom view of the DSPP. Professionals with different knowledge bases or focuses on different levels of DSPP.

1.3.2 Data Sharing Privacy Pyramid (DSPP)

The second representation of the framework is the Data Sharing Privacy Pyramid (DSPP), as shown in Figure 1.4. It is a layered structure that illustrates the knowledge base that supports each research area. The layer is named after a knowledge concept that is represented by researchers and professionals who are working in corresponding research area or whose knowledge can contribute to that layer.

Let us describe the structure from top to bottom. The top layer above the ground line is the Service layer. All service providers are in this layer. They provide services such as news stories, advertisements, and recommended shopping lists through various information channels and media to consumers. With the help of consumers' personal data, personalized services could be provided. Most consumers might not be aware whether, and how, their data are being collected, shared, and analyzed.

The middle layer above the ground line is the Model layer. Data scientists, data analysts, and machine learning (ML) engineers are in this layer. They designed and optimized models and algorithms to learn valuable information, patterns, and knowledge from collected datasets.

The bottom layer above the ground line is the Data layer. Data engineers, software engineers, and hardware engineers are in this layer. They design and improve the systems that collect personal data from the physical space to the cyber (or digital) space and facilitate data sharing. For example, social media platforms help people share information with their friends or to the public. Search engines and virtual assistants retrieve information to users after receiving user' queries. Usually, a service provider would collect certain information from users while providing services to users. Sometimes, the data collected from the users are more valuable than what the users paid for the services. (Notice that most services provided by big information technology (IT) companies are free.) For example, 23andme shared anonymized data with third parties to facilitate healthcare research. Facebook shared user information with Cambridge Analytics to analyze users' political preferences.

Compared to the layers above the ground line, the layers under the ground line are less noticeable by users. However, each of them can affect the adjacent layer, and as a whole serves as the foundation of upper layers.

The first layer under the ground line is the Privacy layer. Lawyers, policymakers, and ethics researchers are in this layer. Warren and Brandeis first raised people's attention to the right to privacy [21]. The right to privacy, like the other rights such as the right to life, is a fundamental human right being protected by laws in the US and other countries. Although privacy behaviors are dependent upon culture and context, even vary dramatically for the same individual [22], the need for privacy is universal across societies over human history, which may root from the deepest instinct of human beings as a well-developed intelligent species. Every health practitioner is required for the oath to protect each patient's privacy. To protect privacy, removing aspects of an individual's identity is codified in laws such as the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [23].

The second layer under the ground line is the Attack layer. Hackers and security researchers are in this layer. At the early age of the development of computer security or network security, the motivation of security attack is purely out of curiosity or to test a bug in a software system. Viruses, spams can hardly directly benefit the attackers. Later, Trojans, Botnets, and ransomware gave the attackers the ability to control and steal information from victims, making the attacks profitable. Security researchers need to test a system's vulnerabilities before it is released and demonstrate a new attack to alert a running system. In a side-channel privacy attack, the only input to the attacker is a target's shared data collected either from the public or from a third party. The output of the attack is the inferred information that can be regarded as a violation of the target's right to privacy.

The third layer under the ground line is the Defense layer. Privacy researchers, cryptographers, system administrators, and data managers are in this layer. They designed or applied models and methods to protect the data from being attacked. For attacks targeting data privacy, there are several families of models that have been developed. Each model has several parameters and different ways to be implemented.

The fourth layer under the ground line is the Optimization layer. Some research groups have shed light on the direction to optimize data protection according to a particular metric. Optimization theory, as a precedent of game theory, studies the way to find the state of a system that lead to a best outcome. One of its successors, control theory, examines how to control a system to reach the desired state, considering real-world constraints in terms of time and environment. For simplicity, researchers usually model an attacker as a part of the environment while optimizing the data sharing.

The fifth layer under the ground line is the Game layer. A few research groups have modeled data privacy problems as games. While most of these privacy games were played between one data sharer and one adversary, others were played among multiple data sharers. How the information and knowledge are distributed in the game shapes the complexity of the models. Game models could be regarded as one kind of optimization problems with multiple decision-makers.

The sixth layer under the ground line is the Computation layer. Few researchers have modeled and tried to solve a practical privacy-preserving data sharing problem using game theory. Instead, options of players in most privacy games are quite limited and thus a brute-force algorithm is sufficient to find the game solution or the Nash equilibrium. However, in real-world scenarios, datasets that need to be shared are usually in a high-dimension or on a large scale. In addition, allowing a data sharer to select fine-grained data sharing policy tends to achieve better balance between data utility and privacy risk. Thus, solving the corresponding game model tends to be computationally expensive and challenging. In these cases, we need computational game theoretic approaches, including heuristics and AI algorithms, to accelerate the search for the game solution.

Note that the professionals working in each layer, usually neglect or try to avoid a deeper layer. For example, people working on data sharing tend to deny the existence of data privacy. People working on data privacy want to rule out the presence of privacy attacks. People working on attacks wish to make protections disappear or ignore the defense mechanism. People working on defense are usually satisfied with good-enough protection solutions instead of chasing the optimal one. People working on optimization try to convert problems with multiple agents into problems with only one decision-maker. At last, people working on game theory try to avoid large-scale strategy space. In contrast, my research recognizes the importance of each layer, blends knowledge concepts from multiple disciplines, opens a door to a brand-new world.

Table 1.1 shows a way to partition each layer in the framework, and it points out the focuses of my dissertation research.

Table 1.1 A taxonomy of knowledge concepts in the Data Sharing Privacy Pyramid

Layer	Way to partition	Category
Data	Type of data attributes	Identity
		Demographic
		Phenotypic
		Image (or video)
		Natural language
		Longitudinal
		Genomic
	The life cycle of data	Collection
		Storage
		Sharing
	Way to share data	Mining
		Modeling
		Publishing
	Application setting	Healthcare
		Research
		Direct-to-consumer
		Forensic
Privacy	Aspect	Confidentiality
		Anonymity
		Solitude
Attack	Attack targeting genomic data	Re-identification
		Attribute inference
		Completion
	Number of stages	One stage
		Multiple stages
Defense	Model	K-anonymity
		Differential privacy
		Access control
		Risk assessment
	Operation	Generalization
		Suppression (or masking)
		Noise addition
		Synthetic data generation
Optimization	Way to combine multiple objectives	Optimize one objective while preserving others
		Pareto Optimization
		Optimize a function that combines all objectives
Game	Number of players	Two players
		More than two players
	Information setting	Complete and perfect information
		Not complete and perfect information
	Number of stages	One-shot game
		Repeated game
	Order of plays	Sequential game
		Simultaneous game
Computation	Algorithm	Linear programming
		Heuristics
		Exhaustive search

1.3.2.1 Data Layer

The data layer can be partitioned in various ways. Here, I only focus on person-specific data that can be collected and shared across entities. One way to partition the data layer is to do so according to the types of data attributes. The first type of data attributes is a direct identifier, such as name, social security number, phone number, and email address. Through each of these attributes, in general, a person can be uniquely identified. The second type of data attributes is a demographic attribute, such as birth date, age, gender, sex, race, ethnicity, education, home address, and zip code. These attributes are prevalent in most datasets. A combination of these attributes may identify a person, so such combination is termed as quasi-identifiers. The third type of data attributes is a phenotypical attribute, such as weight, height, blood pressure, electrocardiogram, symptom, diagnosis, and medication, that may be included in a patient's electronic medical records (EMR). Some of these data are very sensitive, however valuable for medical research. The fourth type of data attributes is a two-dimensional or three-dimensional image (or video) stored as a binary file. It can be a photo posted on a social media platform or a medical image stored in an EMR system. Note that personal photos stored in mobile devices, personal computers, and online clouds are out of the scope of data sharing problem because these data will usually not be shared with others. The fifth type of data attributes is natural language text or voice. The voice commands for virtual assistants are collected by companies to train the assistant. The texts posted on social media can be used to analyze users' behaviors and preferences. It can also be texts in a survey questionnaire or clinical notes in an EMR system. The sixth type of data attributes is longitudinal data. Smart mobile devices such as smartphones, smartwatches, and smart bracelets have the tracking capability to collect and send current Global Positioning System (GPS) location back to service providers with corresponding time stamps. It can also be more valuable and sensitive data attribute such as heart rates and sleeping patterns that are associated with time stamps.

These data attribute types are approximately introduced in the descending order in terms of their time sensitivities. For example, the location of a mobile phone changes every second. A Twitter user can post hundreds of tweets every day. A Facebook user may post photos every other day. A patient might get new symptoms and new weight every other month. A person might change their home addresses every other year. A person seldom changes their name, phone number, and email address.

The last type of data attribute is the genomic attribute. Genomic data have some special characteristics which make it more challenging to be protected, such as high-dimensionality, uniqueness, relatedness, , sensitivity, longevity, and availability [8]. First, its high dimensionality and uniqueness lead to its identifiability. It is almost impossible to find two people sharing the exact same information on genomic markers. Among millions of single nucleotide polymorphisms (SNPs) in a person's genome, only 30

SNPs are required to uniquely identify each person [24, 25]. Just 13 STR markers are sufficient to identify millions of profiles in one of the Federal Bureau of Investigation (FBI)'s forensic databases, Combined DNA Index System (CODIS). Second, strong correlations exist between specific genomic attributes and certain phenotypical attributes, such as certain diseases associated with social stigmas and discriminations. Fourth, before gene-editing technology can be applied to human beings, the same set of DNA markers will almost be unchanged for a person in their lifetime. Furthermore, it is not too difficult to collect discarded bio samples such as saliva, hairs, and skins from an individual and get it sequenced, which is currently not prohibited in the US. As a result, it was debatable if genomic data should be stored in the EMR system in the same way as the other data.

In general, a dataset or a database usually contains more than two types of attributes. In addition, correlations between data attribute intrigue data analysis and incentivize data sharing. For example, sharing datasets with genomic and phenotype attributes is beneficial to the whole society. First, the genetic testing can help doctors and patients with the diagnosis of diseases, especially those that can be treated. The genetic testing was brought into the public spotlight by Angelina Jolie, which was reported by the Time magazine on May 14th, 2013. Second, sharing genomic data can help discover new associations between diseases and genes, especially for rare diseases. Third, to enhance the transparency, the reproducibility, and for the reuse of data for novel investigations, investigators, funded by National Institutes of Health (NIH), are expected to share genomic data from studies to NIH Database of Genotypes and Phenotypes (dbGaP) in a de-identified manner.

The life cycle of data includes data collection, data storage, and data sharing. Here I focus on data sharing. Data sharing can happen in various settings. Let me use genomic data as an example. In healthcare settings, genomic data is collected from patients to healthcare providers to help the diagnosis and prescription. In research settings, genomic data is collected for research purposes. For example, in genome-wide association studies (GWAS), people's genomic data and phenotypes data can be used to determine the correlated genomic marker for certain diseases. In direct-to-consumer (DTC) settings, people send their saliva samples to testing companies to sequence their DNA and receive analysis reports. In forensic settings, law enforcement officers collect DNA samples from suspects and identify the criminals using DNA tracing techniques. In different settings, people sharing the same type of data for different purposes and get a different level of data utilities. For example, in healthcare and forensic settings, the sharing of data is essential and sometimes mandatory. Whereas, in the research settings, the sharing of data is relatively voluntary. And in the DTC settings, the sharing of data can depend upon a consumer's need and interest. In addition, data can be shared on three levels. On the basic level (namely, data publishing), data is shared as it is stored in a data table. On the middle level of data sharing (e.g., federated learning and transfer learning), the models trained on a dataset, instead of the raw dataset, is shared. On the top-level (namely, data mining) the knowledge learned from a dataset is shared.

1.3.2.2 Privacy Layer

Privacy is an umbrella term that covers many definitions. Generally, privacy can be categorized into physical privacy and information (data) privacy. Information privacy is as essential as physical privacy. Computers store information as data. The Internet makes the data remotely accessible. With the development of information technology, personal information is collected by sensors and stored in the database and shared through the networks. Here, I will focus on data privacy, especially genomic data privacy. One way to partition the privacy layer is to categorize it into three aspects: confidentiality, anonymity, and solitude [26]. Confidentiality is defined as keeping the secrets between two parties and preventing unauthorized access. Anonymity is defined as preventing the leakage of identity or membership information [26].

Confidentiality is the primary concern when the data are being used (or shared with authorized secondary parties). Usually, cryptographic tools are employed to fulfill the same purpose: to keep the secrets between two parties (e.g., the individual and the storage facility) will not be accessed by unauthorized parties. A message is encrypted by the sharer and decrypted by the recipient. Data privacy can be breached even if the data sharing process is secured, which indicates that the data needs to be protected after sharing to anticipated (authorized) recipients.

Anonymity is the primary concern when the data are being re-used (or being shared with a third-party or public). It guarantees the information not disclosed by the sharer (i.e., the identity) not be inferred by any recipients. A data record contains three types of attributes: the identifiers, the sensitive attributes, and other attributes. Generally, in a public environment, people will not share sensitive information using their real identities, and they would only share sensitive information anonymously.

1.3.2.3 Attack Layer

Side-channel attacks are those attacks that can infer or uncover information that has not been shared. It can be regarded as a privacy attack if the inferred information is person-specific sensitive information. Here I focus on privacy attacks that target genomic data. They can be categorized into three classes: re-identification (or identity tracing) attacks, attribute inference attacks, and completion attacks [7]. De-identification (i.e., removing identity information) is not enough to protect the privacy because de-identified information can be re-identified (e.g. demographics [27, 28, 29], and genome sequences [30, 31]). In other words, these identity attributes can be recovered by the information remained in other attributes being shared. Two common types of privacy attacks are record linkage attack (i.e., re-identification attack) and attribute linkage attack (i.e., attribute inference attack) [32]. A record linkage attack usually works by linking a de-identified record with an identified dataset using quasi-identifiers. One instance of record linkage attack was shown by Sweeney in her report [33] that a voter registration list can be linked to de-identified hospital records to identify individuals. Even if the sensitive attribute is not shared, the attacker can uncover those sensitive attributes (e.g., demographics [31], genotypes [34], phenotypes [35], memberships [30, 36]) from other attributes given associations among attributes in an attribute inference attack. For example, by comparing records with summary statistics from the case and control datasets, an attacker can detect the memberships of the case dataset, which might be related to a sensitive diagnose such as human immunodeficiency virus (HIV) / acquired immunodeficiency syndrome (AIDS) that can cause social stigma and discrimination. In a completion attack [37, 38], complete genomic data are imputed from partial genomic data or reconstructed from other data using correlations or linkage disequilibrium information. In addition, an attacker can access multiple resources and combine them in a stage-wise manner. In other words, an attack can involve multiple stages of attacks and multiple datasets, which all have records associated with the targeted individual. For instance, a completion attack can be succeeded by an attribute inference attack, which is further succeeded by a record linkage attack for identity tracing in some cases [39].

1.3.2.4 Defense Layer

Defense models and methods are developed to protect privacy. Here I focus on technical safeguards for genomic privacy. To protect confidentiality, main protection models include cryptography (e.g., homomorphic encryption (HE) [40], secure multi-party computation (SMC) [41], secure hardware [42]) and access control (e.g., Beacon services [43], blockchain [44]). To protect anonymity, several anonymization models (e.g., K-anonymity [45], L-diversity [46], t-closeness [47], differential privacy [48]) have been proposed to counteract the linkage attacks. Data anonymization operations include generalization, masking, suppression, obfuscation (i.e., noise addition), and synthetic data generation. Risk assessment can be introduced to evaluate the performances of these methods. Because most models rely on pre-determined parameters, it is hard to choose a parameter suitable for all different environments. And it is harder to guarantee the choice is optimal.

1.3.2.5 Optimization Layer

Notice there is always a tradeoff between data utility and privacy protection. Finding solutions balances perfectly between privacy and utility is a two-objective optimization problem. There are three ways to solve a two-objective problem. The first one is optimizing one objective while keeping the other one above a certain level. The second one is finding a set of Pareto optimal solutions (i.e., the Risk-Utility frontier [49, 50] for this specific problem). The third one is optimizing a payoff function combining both objectives (i.e., utility and privacy for this specific problem), especially as a weighted sum of both. Notice that the attacker's behavior is assumed to be a part of the environment which is either pre-fixed or randomly determined. If the payoff function is linear in terms of controlled variables, the solution can be found using the linear programming. Otherwise, the non-linear programming could be used. When the control variables are integers instead of continuous variables, the programming problem becomes an integer programming problem which is more difficult to be solved. The difficulty increases further when the control variables are binary variables and thus the programing problem becomes a zero-one integer programming problem. In these cases, search algorithms need to be leveraged to find the optimum.

1.3.2.6 Game Layer

While previous works have tried to optimize the two objectives, none of these optimization models consider the interactions between the data protector and the adversary. Most of previous risk analyses assumed a worst-case scenario where the adversary has unlimited means and resources. The decisions people (e.g., most cryptographers) were making were based on what is possible, not what is probable.

In my adversarial model, the attacker is driven by economic incentives. He or she only attacks when it is profitable. Game models take the adversary's responses into consideration, which makes the model more accurate. With an accurate risk assessment, a portion of the data can be shared in an acceptable risk level. The rationality assumption regarding the adversarial models is reasonable, especially when machine learning techniques are being employed to breach people's privacy.

To find the perfect balance between utility and privacy, I introduced game theory to several data privacy problems, which created opportunities to improve the utility and decrease the risk simultaneously. To solve these problems, I further assumed the data sharer is also driven by economic incentives.

According to the problem and assumptions, different settings in the game model need to be determined. For example, for the healthcare scenario shown in Figure 1.3, we can use a two-players one-shot Stackelberg game with complete and perfect information. The entire data sharing process based on game theory works in three steps: 1) the data sharer collects identified health data from data subjects; 2) the sharer selects an optimal data sharing (or protection) strategy; and 3) after receiving the data, the adversary selects an optimal attacking strategy.

For the toy example, we can use a two-player sequential game with complete but imperfect information. It is a game with complete and imperfect information because player B does not know all the actions taken by player A, but both players know both players' utility functions and strategy sets. For a rational player with imperfect or incomplete information, they will try to infer the information and choose the strategy that maximizes the expected payoff.

1.3.2.7 Computation Layer

Typical game theoretic models have relatively small strategy space. For example, in the Prisoner's dilemma model, each of the two players has only two actions. Thus, the Nash equilibrium can be found easily, no matter the players use fixed or randomized strategies. However, for a data-sharing problem, even if we just consider the simplest data anonymization operation – data masking, the strategy space for the data sharer increases exponentially with the size of the dataset. For a dataset with hundreds of data subjects and ten attributes. The size of the strategy size is already at least 2 to one thousand, which is much larger than the number of stars in the universe (2 to 70). The attacker's strategy space can also be huge. The brute force algorithm will not work in these cases. Heuristics such as genetic algorithm and pruning techniques can be used to accelerate the computation.

1.4 Literature Review

In this section, I describe the background and preliminary literature in game theory and genomic privacy and review literatures that are closely related to my dissertation research by pointing out their contributions, their limitations, and their connections to my dissertation research. Note that they are categorized and ordered according to the SPADOG framework.

1.4.1 Data Sharing

Data are premium oils in the big data era. Powerful machine learning and artificial intelligence engines are driven by high volumes of data, including personal data. Person-specific data can be used to make customized recommendations. Online shopping companies and search engines used users' information to provide accurate predictions. Online social networks can use users' shared information to build user profiles and analyze their political preferences. Personal Advertisements could be distributed more accurately based on users' online interactions and search queries. Our voices, faces, typed words, moving trajectories, all information about every one of us has the potential to be collected, shared, and analyzed in this data-driven society.

Genomic data is increasingly collected by a wide array of organizations [51], ranging from direct-to-consumer genomics companies [52] to clinical institutions [53]. This data serves as the basis of discovery-driven research [54] and, more recently, for personalized medicine programs [55, 1]. However, as the quantity and coverage of genomic data grow, so too does the chance for the discovery and reporting of rare alleles [56, 57]. This is challenging for researchers and clinicians who aim to discern if such an allele (or combination of alleles across the genome) is meaningful. To mitigate uncertainty, there is a desire to open data held by one organization to those who may need it elsewhere [58, 59]. There are some initiatives, like the Personal Genome Project [60], that freely and publicly share genomic data linked to phenomic data.

In the DTC settings, more than 25 million people have had their DNAs get genotyped through direct-to-consumer genetic testing (DTC-GT) companies (e.g., 23andMe[1] and AncestryDNA[2]) 23andMe or AncestryDNA. Some of them uploaded their raw data to recreational genomic databases for receiving free services such as ancestral genealogical search (e.g., Ysearch[3]) and relative matching (e.g., GEDmatch[4]), which are not free on 23andMe. Some of them uploaded their genomic data anonymously to open research projects (e.g., Personal Genome Project[5] [60] and OpenSNP[6] [61]) for the social good. There are even online marketplaces (e.g., Nebula Genomics[7] [62] and Luna DNA[8] [44]) mostly based on blockchain technology, where data providers could upload genomic data and get paid by third parties [10].

1.4.2 Data Privacy

Although personal data are inevitably being collected, and data sharing is mandatory in certain circumstances (e.g., national security and public health), a lot of people will not share their personal information publicly and proactively without restrictions, due to the concerns about privacy risks [4],

[1] 23andMe. https://www.23andme.com/.
[2] AncestryDNA. https://www.ancestry.com/dna/.
[3] Ysearch. https://www.ysearch.org/.
[4] GEDmatch. https://www.gedmatch.com/.
[5] Personal Genome Project. https://www.personalgenomes.org/.
[6] OpenSNP. https://opensnp.org/.
[7] Nebula Genomics. https://nebula.org/.
[8] Luna DNA. https://www.lunadna.com/.

especially in the healthcare environment [63, 64]. Considering the sensitivities of most health data, people prefer to share data with trusted health providers. These data holders have responsibilities to keep the data safe from privacy breaches. Historically, this has been achieved by removing aspects of an individual's identity (e.g., suppression of certain demographics), a practice codified in laws around the world, such as the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [5] and the General Data Protection Regulation [65]. Such laws provide specific guidance on how to share de-identified or anonymized information.

The HIPAA Privacy Rule [5] states that only "individually identifiable" information is covered by the regulation. It goes on to state that data is no longer subject to the regulation when it is de-identified. Specifically, this is defined as "information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable." The Privacy Rule then provides alternative implementation specifications, one of which must be followed to ensure that data meets the de-identification definition. The first implementation specification is the Safe Harbor policy, which enumerates 18 attributes that must be generalized or suppressed.

Genomic data are a special type of personal data that is not directly protected by HIPAA. Due to privacy concerns, research projects are not allowed not share individual-level genomic data publicly without the data providers' consents [6, 66]. Corresponding data use agreements (DUAs) explicitly prohibit re-identification [67]. Some DTC-GT companies have privacy policies that promise that they will not share consumer' genomic data with untrusted third parties [2]. Open research projects developed open consent to inform volunteers about the privacy rights they are giving up [68].

To provide intuition into knowledge derived from genome-based investigations, data custodians have turned towards sharing data in statistically aggregated forms (e.g., sharing only allele frequencies) about the pool of individuals who were in a study or were treated in a clinical setting. The practice of aggregated data sharing began on a large scale in the mid-2000s, with programs like the Database of Genotypes and Phenotypes (dbGaP) at the National Institutes of Health (NIH) [69], which aimed to standardize and centralize genomic data, making it easier to access. Summary statistics about the allele rates were made publicly accessible over the Internet because it was assumed that the privacy risks for such data were minimal.

1.4.3 Attacks on Genomic Data Privacy

Erlich and Narayannan analyzed three categories of breaches (i.e., completion, identity tracing, and attribute disclosure attacks) towards genomic data [7]. Naveed et al. [8] reviewed the state-of-the-art privacy attacks on genomic data.

1.4.3.1 Re-identification Attacks Targeting Genomic Data Sharing

In a typical re-identification attack, the adversary re-identifies an anonymous record by linking it to an identified dataset upon a set of common attributes called quasi-identifiers [70, 71]. Malin and Sweeney first demonstrated re-identification attacks on genomic datasets [72, 73]. It is discovered that 30% pedigree structures collected from public obituaries are unique and can be linked upon to re-identify genetic datasets [74]. In one of my adversarial models, the adversary re-identifies genomic records in a multi-stage manner.

Sweeney et al. re-identified participants of the Personal Genome Project [60] by linking these participants' data records to voter registration lists upon three demographic attributes (namely, gender, birth date, and ZIP code), and verified results by leaked names embedded in the filenames [75]. Their attack demonstrates the privacy risk of an open-access database. I considered a similar open-access database that is targeted by the adversary in my adversarial model. However, my re-identification attack model's record linkage stage is based on only two demographic attributes (namely, state of residence and age). They inferred the full name of a record from the filename and used it for verification. In contrast, I inferred the surname from the genomic attributes and used it as a quasi-identifier in the record linkage stage.

Gymrek et al. re-identified 50 participants of the 1000 Genomes Project by first inferring surnames from short tandem repeats on Y-chromosome (Y-STRs) and then linking genomic records to identified public resources upon demographics and those inferred surnames [31]. In response, the National Institutes of Health (NIH) moved the age information of participants in this project into access-controlled databases [76]. I used the same attack in my adversarial model. However, some datasets they used for surname inference (e.g., the Sorenson Molecular Genealogy Foundation) are no longer available. As a result, I simulated datasets on a larger scale in my experiments. Additionally, I developed a risk assessment framework based on which I evaluated several protection mechanisms.

Lippert et al. re-identified records in a whole-genome sequencing dataset by first inferring visual traits from genotypes [35]. Other works re-identified records by first inferring genotypes from phenotypes (e.g., quantitative traits [34], visual traits [77], 3D facial traits [78]) or summary statistics like linkage disequilibrium [79, 80, 81]. My protection model has the potential to deal with these multi-stage re-identification attacks, although the inferred attribute in my experiments is the surname instead of a phenotypic attribute.

1.4.3.2 Membership Inference Attacks Targeting Genomic Data Sharing

Homer et al. [30] demonstrated that an adversary could apply a statistical inference attack to detect the presence of a known individual's DNA sequence in a pool of subjects (e.g., a case-control cohort of individuals positively-diagnosed with a sexually transmitted disease). Specifically, distances between an individual's sequences to the allele frequencies exhibited by the pool versus some reference population were measured. When the target was deemed to be sufficiently biased towards the pool, the adversary could assign the target with the membership. As a result, the NIH, Wellcome Trust, and other genomic data custodians removed all aggregated data of human genomes from public websites [66].

Visscher et al. [82] identified the relations between the detection power, the pool size, and the number of independent single nucleotide polymorphisms (SNPs), pointed out the issue in the selection of reference population, and connected together with the framework based on likelihood ratio and the one based on linear regression. Sankararaman et al. [36] and Jacobs et al. [83] respectively demonstrated the optimal detection power of likelihood ratio statistics both analytically and empirically. Braun et al. [84] criticized the original formulation used by Homer et al. by showing its potential high rate of false positives due to its ignorance of linkage disequilibrium.

Following Homer et al.'s work, Sankararaman et al. [36] improved the statistical detection method with a log-likelihood ratio (LR) test (proved to have the most detection power among all the hypothesis tests). Other than the empirical method, which needs actual data to compute the results, they developed an analytical method that can estimate the number of SNPs as a function of the size of the pool and the two bounds with almost no computation effort. However, such an analytical method is valid only if the pool is large enough (i.e. more than 100 individuals) and SNPs are common and independent from each other, which requires a pre-filtering process that will remove all SNPs with minor allele frequency larger than 0.05 and retain a subset of independent SNPs. Note that, SNPs that are most useful to the data publisher have a great chance to be filtered out during this pre-filtering process. Failing to model the incentives of the data publishers and the risk of the adversaries made this general guidance less practical because a rational adversary may give up the attack and a data publisher may lose the incentive to share data in real scenarios. The analytical method was computationally efficient, highly scalable. However, it is not accurate and needs pre-filtering. On the other hand, the running time of the empirical method increased exponentially with the numbers of SNPs and individuals. For simplicity, they ranked the SNPs according to SNP-disease associations (p-values). They evaluated their methods on real datasets (58C and UKBS control group of the Wellcome Trust Case Control Consortium (WTCCC)) consisted of 33,138 SNPs and 2,927 individuals and implemented an open-sourced tool. Their detection power was underestimated due to the fact that the attacker does not know the ground truth.

Wang et al. [79] extended Homer et al.'s attack for GWAS by using more aggregated data published by the GWAS researchers such as p-values and coefficients of correlations for each SNP pair (r-square) instead of allele frequencies. The first step of their attack is to recover all the single SNP allele frequencies, pair-wise SNP frequencies, and signed allele correlations for each SNP pair by solving equations and constraints. Then they defined a new hypothesis test statistic, which they claimed to be more powerful by making use of more statistic information. A Markov model trained on limited data from HapMap is used to simulate the population (both pool and reference). Based on the simulated dataset, they studied the sensitivities of test statistics to various aspects, such as the size of the pool, the population of reference, the precision of reported r-square, etc. In addition, regression coefficients [85] could be leveraged to improve inference power further.

Membership inference against the Beacon service is demonstrated by Shringarpure and Bustamante (namely, SB attack) as repeatedly submitting queries for minor alleles present in the targeted genome [86]. The robustness of SB attack against SNP hiding strategy could be further improved by considering SNP correlations [87].

Membership could be inferred even if aggregate statistics are released with significant noise [88]. Even identities and traits could be inferred from differentially private GWAS statistics [89].

Determined membership could further reveal the participant’s sensitive phenotypes [90]. Even machine learning models trained on a genetic dataset could reveal the genotypes and/or memberships of the participants [91, 92], a risk that would be exacerbated by the developments of membership inference attacks against machine learning models [93] such as generative adversarial networks (GAN) [94].

1.4.3.3 Imputation Attacks Targeting Genomic Data Sharing

Multiple studies suggest limiting the number of published genetic variants [90], however protecting genetic datasets by data masking/hiding is difficult, considering the correlations between genetic variants and the well-established genotype (or SNP) imputation techniques [95, 96]. It is possible to infer someone’s predisposition for Alzheimer’s disease from a locus that has been masked [97]. 90-98% of forensic STR records can be connected to corresponding SNP records and vice versa, based on correlations (i.e., linkage disequilibrium) between STR and SNP markers [98].

1.4.3.4 Reconstruction Attacks Targeting Genomic Data Sharing

Wang et al. [79] developed another attack that is totally different from the one of Homer et al.'s, which tries to recover all the SNP sequences in the pool based on the reported statistics. According to the evaluation based on simulated datasets drawn from Markov models, Wang et al. concluded that their first attack method has more power (80%) than Homer et al.'s method (9%) with the same fixed false-positive rate of 0.05. Their dataset includes 200 sequences of 174 SNPs in each of the case (i.e., pool) and control (i.e., reference) groups. It took 12 hours to successfully recover the 174 SNPs from all 100 individuals using the second attack. One limitation of their work is that the high computational complexity of their approach makes their attack not scalable.

In a related paper [80], Cai et al. proposed a new recovery attack which is similar to the one of Wang et al.'s but claimed to be more successful. After the successful recovery of all the genetic sequences in the pool, they identified a subset of the reference population that belongs to the pool. In their experiment conducted on 8 WTCCC datasets, where the size of the pool is about 3,500, the size of the reference is about 15,000, the number of re-identified individuals is about 100, and the number of published genotypes is 250. The first step of their attack is recovering the co-occurrence matrix given genotype frequencies, genotype-disease associations, and genotype-genotype correlations, which are all publicly available. The second step is finding presence proofs which are sets of genotypes belonging to at least one individual in the pool. Although the false-positive rate is zero in their experiment, the detection is not powerful enough, being compared to others. Moreover, Cai et al. failed to provide any practical countermeasures for data publishers in the paper.

Due to the similarity between relatives' genetic records, if someone's identified genetic record is shared, it is possible to infer their relatives' genotypes [37] and their predispositions to certain diseases using Bayesian approaches [38, 99]. Specifically, Humbert et al. [38] proposed a reconstruction attack that was based on graphical models and belief propagation. In the attack model, the adversary wants to reconstruct actual genomic sequences of a family from observed genomic sequences of them. In other words, the adversary can infer the target's genome sequence from the target's relatives' sequences which are assumed to be publicly available. The background knowledge of the adversary includes familial relationships, linkage disequilibrium, and minor allele frequencies. More powerful reconstruction attacks are proposed to infer individuals' genotypes from their relatives' genotypes and phenotypes [100, 101, 102]. With known pedigree structures, model-based simulations show that undisclosed genomes get inferred easier when more individuals share their genomes, especially in a homogeneous population [103].

Ney et al. demonstrated that the entire consumer genomic database could be extracted by uploading artificial records [104]. Their attack not only can re-identify anonymized genomic records but also convert a query-based database into an open-access database, demonstrating the vulnerability of a recreational genomic database. In the scenario where users left GEDmatch due to privacy concerns, Ney et al.'s attack showed that a potential adversary might have extracted the entire database beforehand even if the data of users who have left are deleted. Following their attack, my model has the potential to protect open-access research databases (e.g., OpenSNP [61]) and query-based consumer genomic databases (e.g., GEDmatch).

1.4.3.5 Familial Search Attacks Targeting Genomic Data Sharing

In the DTC settings, customers can upload their raw genetic data to online repositories for recreational purposes (e.g., familial search). The high-profile forensic case of the Golden State Killer brought the long-range familial search and its potential privacy risk under the spotlight [105]. Erlich et al. estimated that, in a consumer genomic database (e.g., GEDmatch) of more than one million individuals, 60% of individuals of European descent would be linked to a family member as far as a third cousin [106], similar to the case of the Golden State Killer. Their estimation was based on theoretical models and simulations. Subsequently, Kim et al. demonstrated the possibility of performing these familial searches of short tandem repeat (STR) databases using single-nucleotide polymorphism (SNP) profiles or vice versa [107]. With the help of a genealogical database, the triangulation of targets will be far more powerful [108]. I considered the same type of STR databases. However, the adversary in my model only links targeted individuals to themselves instead of their family members.

1.4.4 Defenses for Genomic Data Privacy

Malin evaluated the technologies for protecting genetic privacy [109]. Erlich and Narayannan analyzed three categories of mitigation techniques (i.e., access control, anonymization, and cryptography) [7]. Naveed et al. [8] presented a framework to systematize the analysis of threats and the design of countermeasures and strategies for mitigating attacks. Tang et al. reviewed technologies for protecting genomic data analytics in the clouds [110]. Wang et al. studied the clinical, technical, and ethical aspects of genetic privacy [111]. Shi et al. categorized protection tools into three groups: controlled access, data perturbation, and cryptography [112]. Aziz et al. categorized genetic privacy into three groups: privacy-preserving data sharing, secure computation, and storage, as well as query privacy [113]. Mittos et al.

systematically reviewed representative papers on privacy-enhancing technologies for genetic privacy and surveyed 21 experts' opinions on identified challenges [114]. Berger and Cho reviewed emerging technologies for privacy-enhancing genomic data sharing such as multi-party computation (MPC), HE, and software guard extensions (SGX) and their challenges [115].

In a survey [32], Fung et al. reviewed the developments in the field of privacy-preserving data publishing, compared the differences between privacy-preserving data publishing and privacy-preserving data mining, and gave a list of desirable properties of a privacy-preserving data publishing method. They also reviewed and compared existing methods in terms of privacy models, anonymization operations, information metrics, and anonymization algorithms. I provide additional reviews regarding privacy risks and data protection models. First, I review the de-identification and anonymization models. In doing so, I justify the generalization strategy invoked in the data protection strategy studied in my dissertation research. Second, I recount the different ways in which re-identification risks are formalized and quantified.

1.4.4.1 Cryptographic Tools

Ayday et al. overview cryptographic works for protecting genetic data [116]. Studies like genome-wide association studies (GWAS) may need millions of records that are usually distributed among multiple repositories. Computations of GWAS statistics can be protected by cryptographic tools. HE is utilized when the computations of statistics are outsourced to external data centers (e.g., counts [40]) or public clouds (e.g., allele frequencies [117], chi-square-statistics [118], regression coefficients [119], counts [120]). SMC enables the computations of statistics over distributed encrypted repositories, without even the local statistics being released [41] and facilitates quality control and population stratification correction in large-scale GWAS [121]. Encrypted hardware is leveraged to perform secure count queries on a genetic dataset in an untrusted cloud [42]. Without the burden of computation on encrypted data, Intel's Software Guard Extensions (SGX) can isolate the computation process in a protected enclave [122]. One study utilizes a combination of HE and SMC to protect aggregated GWAS data [123]. A combination of HE and SGX can securely perform GWAS analysis with high efficiency [124].

1.4.4.2 Anonymization Models

Privacy-preserving data publishing approaches have been introduced to protect the anonymity of genomic data before sharing. The operations that can be applied to anonymize a record can grossly be characterized as: i) generalization, suppression (or masking) [45, 125, 36], ii) randomization (or noise addition) [126, 127], and iii) synthetic data generation [128, 129]. I focused on generalization and masking because they were widely adopted in data protection policies and anonymization approaches. De-identification policies, such as HIPAA's Safe Harbor [23], often used rules in the form of an enumerated list of features that need to be generalized (e.g., 5-digit Zipcode needs to be generalized to first 3-digits, provided there are at least 20,000 people in the region). Similar rule-based policies have been invoked in other countries, such as Canada [130].

1.4.4.2.1 Generalization and suppression (or masking)

Beyond rule-based policies, other anonymization approaches have focused on ensuring the dataset itself satisfies a certain level of protection. For instance, the k-anonymity models [45] ensure each record in a dataset can be linked to at least k records in the dataset with the same quasi-identifiers (i.e., attributes used for linkage). While k-anonymity can be achieved through any of the aforementioned operations [131], the most common approach is generalization or suppression [132]. Subsequent models further improve the protection power [46, 47]. K-anonymity models can be applied to both genetic and non-genetic attributes in a dataset. Some studies generalized nucleotides into broader types to satisfy 2-anonymity [133, 134]. Some studies generalized diagnosis codes in GWAS and Phenome-wide association studies (PheWAS) datasets to satisfy 5-anonymity [125, 135]. I adopt the generalization model without enforcing a specific protection parameter. Rather, I searched for a generalization that maximizes the payoff for the publisher of a record.

Several generalization models have been developed (particularly with applications for k-anonymization), and it is important to clarify which is used in my dissertation works. Specifically, I used a full-domain generalization model [136], which is the cross-classification of the domain generalization hierarchy (DGH) for each attribute. This is the most frequently used generalization model in practice, but my framework can be extended for other generalization models, such as full-subtree generalization [137] and multi-dimensional generalization [138].

In the re-identification game, which I studied in my dissertation research, I assumed the defender's strategy set is derived from de-identification or anonymization models. Various models have been developed in my dissertation research, but most of them aim to transform the attributes that could be used to ascertain an individual's identity to address identity disclosure risk. While there are other privacy concerns (e.g., attribute disclosure [46], membership disclosure [139, 140]), I focused on identity disclosure because of its direct relationship with existing privacy laws and a broad class of data protection methodologies [32]. I believe that my game theoretic framework will generalize to other privacy models, such as differential privacy [48], which applies random noise to shared data. In general, I focused on identity disclosure to illustrate the novel perspective that games can bring to the anonymization problem.

1.4.4.2.2 Noise addition

The differential privacy (DP) model [48] has been used to prevent membership-inference attacks (e.g., Homer et al.'s attack) by adding noise to summary statistics (e.g., allele frequencies, chi-square statistics, and p-values) [141, 142]. However, it is inherently challenging to use DP techniques to protect GWAS data because the number of outputs (e.g., the correlations between SNPs) is much larger than the number of inputs (i.e., the number of participants) [143] and a large amount of noise is required even for releasing statistics of a small number of SNPs [7]. To effectively preserve privacy, the resulting utility of the DP model for personalized Warfarin dosage would be so low that patients would be at risk of death [91]. With a weaker adversarial model, a better utility could be achieved [126]. By adding noise to input instead of output, accuracy could be improved [144]. Using DP, meaningful GWAS results of 2 SNPs are returned based on two state-of-the-art statistics in a heterogeneous population of hundreds of individuals [127]. A combination of HE and DP provides a practical privacy-preserving solution for accessing aggregated genetic data [145, 146]. To avoid high cost of accuracy in DP models, a Markov chain Monte Carlo model based on risk-assessment is utilized to adding noise to aggregate genetic data [147].

1.4.4.2.3 Synthetic data generation

Deep learning models such as the Generative adversarial network [148] could be harnessed to protect genetic privacy by generating synthetic datasets for sharing [149].

1.4.4.3 Access Control for Genomic Data Sharing

Few genetic datasets are open for public access at the individual-level without the data subjects' consent. For GWAS, NIH established a two-tier access-controlled system that hides individual-level genotypes and phenotypes behind the firewall and only publish summary statistics (e.g., minor allele frequencies, chi-square-statistics) that are useful for meta-analysis.

Accesses to databases like dbGAP are restricted [69], which may slow down research advances. An alternative is a semi-trusted registration-based query system [150], with recording and auditing capabilities. The open-accessed Beacon service provided by the Global Alliance for Genomics and Health (GA4GH) let users query on presence information instead of allele frequencies for genetic datasets as a strategy of trading data utility for privacy [43].

Participant-centric access control allows the original participant to control access to their genomic data [7]. Kaye et al. overviewed and analyzed various participant-centric biomedical research initiatives that protect privacy and maintain public trust [151]. Clayton et al. systematically reviewed works that consider genomic data privacy from individuals' perspectives [26]. I designed data sharing strategies from the participants' perspectives and conducted experiments accordingly.

Through a single-blinded randomized trial in a healthcare environment, McGuire et al. found out that 53.1% of participants are willing to publicly release their genomic data with no controls [63]. I would argue that participants in the healthcare environment care more about the utility of health data. Additionally, the number of high-profile privacy attacks targeting genomic data keeps increasing recently [7, 114].

Through a randomized trial in a healthcare environment, Oliver et al. found out that privacy risks were largely outweighed by utility regarding genomic data sharing and urged policymakers to respect participants' diverse assessments on the privacy risks and the utility and participants' privacy-utility determinations [64]. I would argue that participants' intuitive assessments of their privacy risks are not accurate, so I used an analytical model and concrete experiments with various cost-and-benefit combinations to help participants and policymakers make more informed and strategic decisions.

Deuber et al. developed a system based on garbled circuits that enables a data subject to arbitrarily decide which third-party can compute over their encrypted genomic data stored in the cloud [152]. Whereas in my model, the data subject decides to share data with an open-access database, so no encryption is required. Furthermore, my model aims to provide an optimal sharing strategy.

Roberts et al. concluded that paying individuals for genomic data sharing benefits both the company and consumers and asked for the best ways to incentivize individuals to share data [10]. Blockchain, a distributed and secure database model, has been introduced to genomics to facilitate data sharing, to avoid a single point of failure, and to protect privacy [153, 154]. For example, Nebula Genomics incentives data subjects to sell data directly to third parties (e.g., researchers) using its cryptocurrency while protecting the privacy of data subjects with the help of HE and SGX [62]. Other companies, like EncrypGen and Luna DNA, have similar mechanisms [44]. Blockchain frameworks have been proposed for the privacy-preserving sharing of GWAS datasets [155] or machine learning models trained on genetic datasets [156]. I considered a similar economic model in my system where data subjects are incentivized to share data, and my models can determine an optimal sharing strategy.

1.4.4.4 Countermeasures for Genomic Privacy Attacks

Wang et al. [79] mentioned three potential countermeasures for their attacks: downgrading the precision of reported statistics, publishing less data (dropping data with small r-square), and adding noise. However, all these countermeasures will render GWAS papers less informative. However, their experimental results indicated that their attack was robust to the first two countermeasures. Wang et al. further suggested enforcing differential privacy or selectively removing some data to defeat their attacks. Although they proposed two powerful attacks, the major limitation of their work is the missing of concrete study on the countermeasures.

In a follow-up paper from the same group [157], Zhou et al. presented a preliminary framework for a risk-scale system for the data publisher. They assumed that the adversary could not accomplish the task that needs exponential computing power. They provided releasing guidance conditioned on the relationship between the number of sequences and the number of SNPs, making use of the high computational complexity of the two attacks in Wang et al.'s paper [79] (proved to be NP-complete). They considered two scenarios (types of attack): recovery attack given allele frequencies and identification attack, given test statistics (r-square). However, their framework was not practical because the computational cost is only a small portion of a rational adversary's payoff.

Following Homer's work, Sankararaman et al. [36] provided quantitative guidelines for the data publisher to publish a subset of SNPs without compromising the data subjects' privacy. Specifically, they estimate the upper bound of the number of SNPs that can be published safely given the size of the pool, the maximal allowable detection power (true-positive rate), and the maximal allowable false-positive rate.

Ayday et al. [158] defined the genomic privacy of a patient from the medical center's point of view and discuss protection strategies such as policy enforcing and obfuscation.

Strategies have been proposed to address the SB attack, such as adding noise [159, 160], imposing a query budget [159], and providing false answers for most discriminative variants as I proposed.

1.4.5 Optimization for Data Privacy Protection

Rule-based policies were not necessarily optimal, per se, and thus numerous anonymization approaches have been proposed to discover policies that maximize data utility while satisfying a risk threshold [161, 125, 49]. Accurately assessing utility and privacy risk is the first step before optimizing the utility while protecting privacy [50].

1.4.5.1 Re-identification Risk Assessment

No system is impregnable to attack and, thus, re-identification risk assessments must be performed. Elliot and Dale suggested three models of re-identification risks [162]: i) prosecutor, ii) journalist, and iii) marketer. For these risks, it is assumed there is a published dataset, which is based on a sample of a broader population. The prosecutor and journalist risks correspond to the most re-identifiable record in the dataset and population, respectively. The marketer risk, by contrast, corresponds to the average risk of all records in the dataset. When Dankar and El Emam provided mathematical definitions [163] of these scenarios, it was assumed that the attack will always be attempted. However, as I showed in my dissertation works, the cost of a privacy violation (e.g., the expected loss in terms of a fine) greatly influences this decision. There have been investigations into the cost of privacy violations. For instance, Banerjee et al. introduced quantitative methods [164] to define privacy violations and their consequences. They provided definitions of sensitivity and severity (of privacy violations), considering the level of which the data subjects are concerned with regards to their own privacy. Lebanon et al. designed a decision theoretic framework [165] to assess privacy risk that accounts for both the entity identification and the disclosed information sensitivity. However, these models did not consider the decision-making process of the adversary with varying strategies as in my game models.

One of the challenges associated with re-identification is that an adversary must obtain a degree of background knowledge in order to perpetrate their attack. In certain instances, this knowledge may be gained by observation, such as when the adversary sees an ambulance leaving their neighbor's house. Yet such information may be difficult to come by and, thus, it has been suggested that reasonable adversaries are more likely to use resources, such as public records or information brokers, which can be gathered or queried *en masse*. In this regard, there has been some investigation into the credentials and costs associated with gathering such resources. In particular, Benitez and Malin [166] illustrated that voter registration records, which have been used for re-identifications, have a wide range in price (from $0 to $17,000), which is set by the state or municipality making them available, and the amount of information useful for re-identification (e.g., demographics) is not correlated with the price (e.g., the most expensive resource actually had the least amount of information).

1.4.5.2 Risk Assessment for Privacy-Preserving Genomic Data Sharing

Malin et al. reviewed the problem of identifiability in biobanks from a risk assessment perspective [167]. Usually, the privacy risk is quantified with a specific attack into consideration. Humbert et al. quantified the genomic privacy resulting from their reconstruction attack, in which the adversary can infer the target's genome sequence from his or her relatives' sequences [38]. Humbert et al. maximized the utility while protecting genomic privacy in a family by optimizing the way of masking SNPs [168]. Afterward, Kale et al. revisited the problem with an improved approach [169]. Later, Humbert et al. quantified the interdependent privacy risks associated with more powerful reconstruction attacks they proposed to infer individuals' genotypes from relatives' genotypes and phenotypes [100, 101]. My protection model also quantified an individual's utility and privacy and optimized the way to mask genomic attributes. However, I considered STR instead of SNP markers and chose a different way to solve the multi-objective optimization problem. Besides, I did not consider other family members' privacy risks. Because of well-established genotype imputation techniques and the correlations between genomic attributes, masking approaches are more suitable for independent genomic attributes [95, 96]. The STR markers considered in my experiments do not have strong correlations, making the masking approach suitable for my problem.

Craig et al. discussed approaches for assessing and quantifying the privacy risks to participants that result from the sharing of summary-level data with a membership inference attack in consideration [170]. Wagner systematically compared and evaluated 24 genomic privacy metrics in four possible attack scenarios, in all of which the adversary aims to infer a person's genomic markers [171].

Compared to these quantification models, one of our models quantified both utility and privacy and designed a measurement that accelerates the search process for the best protection strategy against the SB attack [86]. One of our models considers a multi-stage re-identification attack instead of a one-stage attack (e.g., reconstruction attack, genotype inference attack, membership inference attack). Moreover, this model not only quantifies both utility and privacy but also optimize them simultaneously. In addition, the adversarial behaviors in most of these adversarial models are pre-fixed instead of strategically determined, which might be an assumption too simplified for a real-world scenario.

1.4.6 Game Theoretic Models

HIPAA acknowledges this fact by stating in the second implementation specification that "information is not individually identifiable health information only if: 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." Here, I highlight the fact that the notion of de-identification is explicitly tied to a risk assessment that accounts for the capabilities of a reasonable adversary. I believe this is a clear justification for the application of the game-based approach to the de-identification problem.

Although optimizations regarding privacy-preserving data sharing can explicitly trade-off utility and privacy, they fail to account for the behavior of would-be attackers who attempt to violate the privacy of individuals in the shared data. Furthermore, the side-effect of this line of research, which focuses on worst-case scenarios, has been the increasing promotion and, at times, adoption of data sharing practices that may unnecessarily impede research [172]. In other words, the research community often focuses on what is possible, as opposed to what is probable. To date, there has been little evidence that such attacks will be realized in practice for any reason other than a demonstration [173]. Rather than viewing de-identification as a dichotomous problem of "broken" or not, it should be considered as a matter of risk over a continuous range.

Recently, risk-based approaches have been proposed for decision support in de-identification [174]. However, it should be recognized that a key source of re-identification risk is a decision on the part of the data recipient to *attempt* re-identification. It is natural that the anticipated data recipient, in most practical data sharing settings, will only attempt re-identification if there is a tangible economic benefit (e.g., monetary gain) to doing so.

Game theory can be invoked to provide a more realistic view of how to measure the risks and recommend protections. A game theoretic approach offers an alternative that trades off utility and privacy in a way that explicitly accounts for adversarial behavior and capabilities and can potentially help the data sharers and the policymakers to assess the privacy risk and find the best protection strategy.

1.4.6.1 Background of Game Theory

Game theory is a branch of applied mathematics that studies the interactions among rational agents and their behaviors [11]. Four key components for a game include the players, their strategies, payoffs, and the information they have [175]. A game G with n players can be defined as a triplet <*P*, *S*, *U*>, where $P = \langle p_1, \dots, p_n \rangle$ is the set of players. $S = S_1 \times \dots \times S_n$ is the set of strategy profiles, where S_i is the set of strategies that player i can take. A strategy is a plan of actions the player can take. U is the set of payoff functions. The payoff of a player can be affected by the actions of all the players. Thus, the set of payoff functions can be repressed as $U(s) = \left(u_1(s), \dots, u_n(s)\right)$ where $u_i: S \to R$.

A strategy profile $s^* \in S$ is a Nash equilibrium if no player can do better by unilaterally changing his or her strategy, that is:

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*), \forall i \qquad (1.1)$$

where s_{-i} is a strategy profile of all players other than player i. When all players adopt a fixed action as their strategy, s^* is a pure-strategy Nash equilibrium (PSNE). When all players randomly choose an action according to a probability distribution, s^* is a mixed-strategy Nash equilibrium.

A complete-information game is one in which the players know others' strategies and payoffs. A perfect-information game is one in which the players know others' previous moves. A zero-sum game is one in which all players' payoffs always summing up to zero. In AI, the most common games are deterministic, turn-taking, two-player, and zero-sum games of perfect information, such as Chess [15] and Go [16, 17]. Two-player zero-sum game models have been recently applied in machine learning and statistics to solve data-driven problems optimally [148, 176].

In a two-player Stackelberg game [177], there are two players: the leader p_1 and the follower p_2. The leader makes the first choice, and its strategy is then observed by the follower.

Let us assume that the follower then plays the optimal strategy in response to the leader (that is, the *best response*), defined as

$$s_2^*(s_1) = \underset{s_2 \in S_2}{\operatorname{argmax}} \, u_2(s_1, s_2) \tag{1.2}$$

Using backward induction, a brute force methodology, the leader's best strategy can then be determined:

$$s_1^* = \underset{s_1 \in S_1}{\operatorname{argmax}} \, u_1\big(s_1, s_2^*(s_1)\big) \tag{1.3}$$

Next, I contextualize my dissertation research with respect to other investigations into game theory for characterizing and addressing privacy and security concerns.

1.4.6.2 Game Theory Applied to Security

Game theoretic frameworks have been introduced to model privacy and security problems. Manshaeiy et al. [20] surveyed applications of game theory in addressing network security and privacy problems in six main categories: security of the physical and media access control (MAC) layers, security of self-organizing networks, intrusion detection systems, anonymity and privacy, the economics of network security, and cryptography.

Security game has become an important area of research in Artificial Intelligence. Tambe et al. applied computational game theory, a new area in the field of game theory, to real-world security problems as security games [178]. Tambe et al. [179] proposed a security game model between the defender (e.g., police officers) and the adversary (e.g., terrorists) to optimize the allocation of limited security resources, which is extended by An et al. [180] by considering the surveillance cost and partial knowledge of the adversary. Tambe et al. [179] reviewed how games are designed and applied for the protection of security. The problems they were dealing with are limited security resources allocation and scheduling problems considering rational adversaries. Based upon Bayesian Stackelberg game models, new algorithms have been developed and deployed in multiple real-world applications: ARMOR has been deployed at Los Angeles International Airport since 2007 [181]; IRIS has been used by US Federal Air Marshals since 2009 [182]; PROTECT has been deployed in the port of Boston for US coast guard patrolling since April 2011 [183, 184]; GUARD is under evaluation by the US Transportation Security Administration [185]; and TRUSTS is being tested by the Los Angeles Sheriffs Department [186].

1.4.6.3 Game Theory Applied to Data Privacy

Game theoretic approaches have been applied to various problems regarding data privacy. The first paper in this area traced back to 2003. Acquisti et al. [187] relied on a repeated simultaneous game model with complex utility functions to study the incentives and behaviors of participants in anonymity networks (like The Tor network). Many assumptions about the players' strategies and behaviors are discussed. However, those utility functions were not specified, and no experiments were conducted. Chen et al. [188, 189] proposed a general game model to analyze the effect of privacy concerns on the behavior of selfish agents and the Nash equilibrium. The game is named Coupon game, which is based on the signaling game, in which player *A* aims to discover player *B*'s secret type by offering a coupon. Player *B* chooses to respond randomly or not. Three different payoff functions are considered in their model.

Rajbhandari and Snekkenes [190] defined a normal form game between a user (i.e., the defender) and a service provider (i.e., the attacker) for assessing privacy risk. In this game, the user chooses whether or not to provide private information, while the service provider chooses whether or not to exploit the user's private information. The payoffs representing saved or lost time, which should be collected by conducting experiments and surveys, are determined arbitrarily in their demonstration. Using their set of payoffs, the game in normal form has no pure-strategy Nash equilibrium.

1.4.6.3.1 The application of the Stackelberg game model to privacy

The Stackelberg game model [177], which I leverage to model the re-identification game, has been used in various contexts. Bruckner and Scheffer [191], for instance, modeled the adversarial prediction problem as a single-shot game, which is one kind of Stackelberg game as I used in my model, and explored the conditions for the existence of unique Nash equilibrium. Bruckner and Scheffer [192] modeled the adversarial prediction problem as a Stackelberg game between a data generator (leader) and a learner (follower). Here, the leader generates data based on the follower's prediction models to create confusion, while the follower adjusts the prediction models to account for the leader's response.

1.4.6.3.2 Privacy games related to machine learning

Some papers applied game theory to privacy and machine learning. In Chapter 3 of his dissertation [193], Loiseau summarized his work relating to game theory and statistical learning for online privacy. He focused on investigating algorithms to learn from personal data and how they affect privacy. First, Ioannidis and Loiseau [194] formulated a non-cooperative game model to account for the public good nature of the learning outcome in the context of linear regression. Later, Chessa et al. [195] formulated a game theoretic model, in which individuals take control over participation in projects and data analysts set requirements for data precision. Meanwhile, Liu and Chawla [196] proposed a game model for adversarial learning based on the assumption that the players have uncertainty with respect to each other's payoff function. More recently, Oh et al. [197] introduced a general game theoretic framework for the user-recognizer dynamics, and they presented a case study that involves state-of-the-art adversarial image perturbations and person recognition techniques.

1.4.6.3.3 Location privacy games

Some papers applied game theory to location privacy. Gianini and Damiani [198] modeled the location privacy protection problem as a two-player zero-sum signaling game. However, the problem of privacy-preserving data sharing on which I investigated can hardly be modeled as zero-sum games because the sum of the payoffs of both players will not be zero. Freudiger et al. [199] utilized a game theoretic model to analyze the behaviors of nodes in mobile networks to protect location privacy. They analyzed both complete information and incomplete information multi-player game models regarding the Nash equilibriums. However, no real-world data were used in the investigation. Liu et al. [200] applied a non-cooperative Bayesian game to distributed dummy user generation in both static and time-aware contexts. They analyzed the Bayesian Nash equilibrium, proposed a strategy selection algorithm, and conducted a simulation on real-world data. Wang and Zhang [201] identified the context privacy problem considering the context dynamics and formulated the interactive competition between a smartphone user and a malicious adversary as a zero-sum stochastic game. In their model, the user's action is to control the released data granularity while the adversary's action is to select the source of sensing data. Olteanu et al. [202] proposed a game theoretic framework for analyzing and predicting the strategic behaviors in terms of (co)-location information sharing of online social network users.

Shokri et al. [203] formulated the location privacy problem as Stackelberg Bayesian games with the consideration of a user's service quality and adversary's cost. Shokri [204] developed a Stackelberg Bayesian game model to design location data obfuscation (e.g., adding noise) mechanisms to balance the utility-privacy tradeoff optimally. Specifically, in his model, the user (i.e., data publisher) wants to minimize the utility cost and the privacy risk. However, the data obfuscation increases the cost and risk at the same time. The rational adversary wants to minimize inference error (i.e., the user's privacy risk). The historical data the user published is used for the prior distribution (which is known to the user and the adversary) of the user's current data point. Shokri considered two types of users: i) the utility-sensitive users that want to minimize utility cost while guaranteeing the desired level of privacy, and ii) the privacy-sensitive users that want to minimize the privacy risk while guaranteeing the desired utility. For each type, Shokri provided three mechanisms: 1) optimizing differential privacy bound, 2) formulating a non-zero-sum Stackelberg game, and 3) combine them together. The information-sharing process is like this: the user has a secret *s* (belongs to *S*), obfuscate it to observable *o* (belongs to *O*), and the adversary infers it from *o*. The utility cost is a distance between *s* and *o*, and the privacy risk is the distance between the inferred secret and *o*. The strategy space of the user is a distribution over *O*, and the strategy space of the adversary is a distribution over *S*. For the first mechanism, the Bayesian approach is highly compatible with the definition of differential privacy. Other than the optimal attacker, Shokri further introduced a Bayesian inference attacker. Shokri concluded that i) the third protection mechanism is best, and ii) the optimal attacker is better. Shokri developed an excellent theoretic model. However, he failed to discuss the sensitivities of the results to the two distance functions. Another limitation is the unwarranted assumption that the prior knowledge of the adversary should be less informative than the user's.

Later, Shokri et al. [205] formalized a Stackelberg Bayesian game model enabling the design of optimal user-centric location obfuscation mechanisms respecting each individual user's service quality requirements (i.e., data utility), while maximizing the expected error that the optimal adversary incurs in reconstructing the user's actual trace.

1.4.6.3.4 Privacy games related to Online Social Networks

Some papers applied game theory to privacy in online social networks. Biczok and Chia [206] tackled the issue of interdependent risks caused by agents with misaligned incentives regarding their privacy in online social networks by using a game theoretic framework. Pu and Grossklags [207] went one step further by studying large groups of users who take others' preferences into account when making their own decisions. Chen et al. [208] modeled and analyzed the privacy settings of social networks from a game theoretic perspective by introducing three types of game models. The first is a two-user game model, which investigates the relationship between two users when they disclose profile attributes. The second is the basic evolutionary game model, which shows the dynamic behavior of multiple users as in large-scale online social networks. The third is a weighted evolutionary game model, which considers the influence of attribute importance and network topology. They concluded that the network topology has a limited effect on the privacy dynamics of the network in the absence of risk.

1.4.6.3.5 Audit privacy games

Some papers applied game theory to privacy in auditing. Blocki et al. [209] introduced the notion of games for auditing the use of medical records in the context of a primary care setting. In their Stackelberg game model, the defender is the data publisher (e.g., the hospital or the auditor), and the adversary is the data recipient (e.g., the employee or the auditee). The defender's action space in the audit game includes two components. The first component is the allocation of its inspection resources to targets. Only one target is audited, and only one target is attacked. A randomized strategy is a probability distribution on each of the targets being audited, which exists in a standard model of security games in [179]. Second, they introduced a continuous punishment rate parameter, which brings the defender a cost. The adversary's utility includes the benefit from committing the violations and the loss from being punished if caught by the defender. Their model is very similar to the Stackelberg game model I used. However, the data-sharing problem I targeted is in a research setting where the data sharer has no auditing capability. The protection in my game model happens before the data sharing and the attack rather than happening after the data sharing and the attack as in their setting.

Later, Blocki et al. [210] updated their audit game model such that the defender has multiple resources using which he or she can audit a set of targets. In addition, each resource may be restricted to a subset of potential violations. Cardenas et al. [211] formulated the problem of privacy-preserving demand response as a control theory problem, and they showed how to select the maximum sampling interval for smart meters in order to protect the privacy of consumers while maintaining the desired load shaping properties of demand-response programs. However, attackers were not modeled as players in their games.

Later, Yan et al. [212] modeled the interaction between a database auditor and potential attackers as a Stackelberg game in which the auditor chooses an auditing policy and attackers choose which records in a database to target. In addition, they performed an extensive evaluation based on two real datasets [213].

1.4.6.3.6 Game theory applied to privacy-preserving data mining

Some papers applied game theory to privacy-preserving data mining. Kargupta et al. [214] applied a game theoretic approach to multi-party privacy-preserving distributed data mining, especially to secure multi-party computation with the example of sum computation. Their contributions include describing a penalty mechanism to shift the equilibrium from collusion to no collusion among players. However, in this game, every party is both a data publisher and an adversary, which is different from my two-player game models. Kantarcioglu et al. [215] made recommendations for firms' decisions and the government interventions considering consumers' decisions based on a Stackelberg game model. In their game, the firm is the leader, and the consumer is the follower. They thoroughly analyzed the properties of different types of joint distributions on random variables in the model. Nix and Kantarcioglu [216] proved that the mechanisms designed, rewarding players based on their contributions, in an information-sharing game for classification analysis are individual-rational and incentive-compatible under the assumption that deviation from true value would decrease the utility (i.e., accuracy) in both non-cooperation and cooperation cases. Three real datasets were used to examine the model's practical capability. Wang et al.

[217] introduced a game theoretic approach to the design of a payment mechanism. Their model made the quality of the collected data controllable through a parameter by making sure that everyone's strategy in a Nash equilibrium is to participate and symmetrically randomize data and by guaranteeing differential privacy.

Some papers specifically applied game theory to secure multi-party computation (SMC). Miyaji and Rahman [218] built a two-party secure set-intersection protocol in the game theoretic setting using cryptographic primitives. The constructed SMC protocol satisfied computational versions of strict Nash equilibrium and stability with respect to trembles. Some papers specifically applied game theory to secure data integration. Mohammed et al. [219] leveraged a two-player two-action information-sharing infinite repeated game to K-anonymize large integrated private data from multiple data providers. The game was based on a designed interactive protocol, and the K-anonymization was based on a top-down generalization hierarchy tree search.

1.4.6.3.7 Game theory applied to privacy-preserving data publishing

Several papers applied game theory to privacy-preserving data publishing. Kumari and Chakravarthy [220] set up a cooperative privacy game in which each tuple in the data table behaves as a player, trying to preserve their privacy and, in turn, helps in preserving other's privacy. Their game model was different from mine in terms of player modeling. Li et al. [221] designed a game theoretic framework for a data publisher to evaluate the tradeoff between re-identification risk and the value of shared unstructured natural language data. Their game model was built upon unstructured data, while my game models were built upon structured high-dimensional datasets. Wu et al. [222] constructed a game model of multiple players. In their model, each player publishes the dataset sanitized by differential privacy. They demonstrated the sufficient conditions of the existence and uniqueness of the pure Nash Equilibrium and referred to the price of anarchy to get the efficiency of the pure Nash Equilibrium. However, the adversary is not modeled as a player in their game model. In addition, they assumed a worst-case scenario for the privacy risk and did not specify utility functions in the game model.

Duong et al. [223] proposed a game theoretic framework to maximize the publisher's utility (payoff) by adopting proper strategy while considering multiple rational adversaries with structured background knowledge and sharing rules. Specifically, the published dataset is a de-identified table with sensitive information (e.g., disease) within it. The goal of the adversary is to re-identify a target in the table. The background knowledge can be classified into three categories of facts. The sharing rules force everyone to share only true information and the same amount of information. The benefit of the adversary is defined as the value of the target record divided by the square of the number of successful adversaries. The strategies of the publisher are different ways to generalize the table into equivalence classes. The utility of the publisher is a normalized weighted sum of information loss and probability of a breach. They simplified this model from a Stackelberg game to a normal-form game. In each simulation, an attacker's background knowledge was sampled from a prior distribution. They made three different assumptions for attackers' rationalities in different scenarios. Attackers can act rationally by playing a Pure-Strategy Nash Equilibrium (PSNE) or act irrationally by sharing no information or sharing all information. The results showed that the publisher would get the worst utility if all attackers share all information. The limitation of this work is that it lacks a strict theoretic framework and enough real data to validate the framework. The weight factor in the publisher's utility function is hard to be decided. The authors failed to consider the cost of attackers, which is reasonable but not always the case.

1.4.6.3.8 Genomic privacy games

There exists only one paper, other than my papers, that applied game theory to genomic privacy. Considering the reconstruction attack that they proposed [38], Humbert et al. [224] introduced two normal-form games with multiple players to optimize kinship genomic privacy for data sharers. In their setting, there is no data holder (or agent) that collects data and makes decisions for a pool of individuals. Instead, each user (i.e., data owner or data subject) makes a decision on his or her own. The decision to share genomic data affects not only their own utility and privacy but also their relatives' privacy. One set of experiments are conducted on two members in the Utah family with 82,000 SNPs, and the other set of experiments are conducted on the nine members of the Utah family with 1000 randomly chosen SNPs. The results show that misaligned incentives have a negative impact on social welfare.

Their game models have several advantages over my models. First, it highlights the fact that the decision to share one's own genome can affect their relative's privacy. Second, it models multiple players in a sequential manner. Third, it considers how social welfare is affected by the cooperative and altruistic settings.

At the same time, their game models have several limitations compared to mine. First, the adversary is not modeled as a player in the game model. Summarizing the adversary's influence on the game as a fixed probability of successful breach makes the adversarial model unclear and over-simplified. Second, in their game, the protection decisions for all genomic attributes are the same. In all games they considered, each player only has two choices. In the storage-security game, a player chooses to invest in security or not. In the disclosure game, a player chooses to disclose his or her genomes or not. In a user-centric data-sharing model, a user should have the sharing options with more granularity in terms of which portion of the data to disclose. These assumptions may not fit the real situation and may not achieve the optimal payoff. In contrast, in my game models, the protection decision for each attribute can be different. Thus, their games have a much smaller strategy space, such that searching for the Nash equilibrium is not as challenging as in my game models Third, their parameter settings are not based on real-world settings. Although real-world data are used in experiments to verify the analytic solutions, the parameters such as benefits and costs are still expressed as metrics between 0 and 1, irrelevant to real-world scenarios. For the same reason, it is not clear which variable can be controlled to mitigate privacy risk and optimize social welfare.

Table 1.2 A comparison of representative works applying game theory to data privacy

ID	Paper	Players	Actions	Data	Model
1, 2	Chen et al. 2014 (2020)	2 players	Offer/Accept coupon (Randomize) or not	None	Normal
3	Bruckner and Scheffer 2011	1 data generator and 1 learner	Generate training data; select a parameter	Real	Stackelberg
4, 5	Blocki et al. 2013 (2014)	1 auditor and auditees	Allocation of inspection resources, and a continuous punishment rate	None	Stackelberg
6, 7	Yan et al. 2018 (2019)	1 auditor and attackers	Audit or not; choose records as targets	Real	Stackelberg
8	Olteanu et al. 2016	2 online social network users	Share location or not	Real	Normal
9	Freudiger et al. 2009	N mobile nodes	Change pseudonym or not	None	Normal
10	Shokri 2014	2 players	Obfuscation (attacking) mechanism	Real	Stackelberg
11	Kantarcioglu et al. 2010	1 firm and consumers	Invest or not; use service or not	None	Stackelberg
12	Nix and Kantarcioglu 2012	N players	Share secret or not	None	Normal
13	Wang et al. 2014	1 phone user and 1 attacker	Choose the location granularity; choose the sensing source	Real	Stochastic
14	Duong et al. 2010	1 publisher and adversaries	10 anonymization strategy; share background knowledge with peers	None	Normal
15	Li et al. 2016	1 publisher and 1 attacker	Share or not; re-identify or not	Real	Stackelberg
16	Humbert et al. 2015	2 or N family members	Invest or not (disclose or not)	Real	Normal

Table 1.2 summarizes and compares an array of representative works applying game theory to data privacy, which are related to my dissertation research. Paper 1 (Chen et al. 2014) and paper 2 (Chen et al. 2020) both applied game theory to general data privacy; Paper 3 (Bruckner and Scheffer 2011) applied game theory to privacy related to machine learning; Paper 4 (Blocki et al. 2013), paper 5 (Blocki et al. 2014), paper 6 (Yan et al. 2018), and paper 7 (Yan et al. 2019) all applied game theory to privacy in auditing; Paper 8 (Olteanu et al. 2016) applied game theory to privacy related to online social networks; Paper 9 (Freudiger et al. 2009) and paper 10 (Shokri 2014) both applied game theory to location privacy; Paper 11 (Kantarcioglu et al. 2010), paper 12 (Nix and Kantarcioglu 2012), and paper 13 (Wang et al.

2014) all applied game theory to privacy-preserving data mining; Paper 14 (Duong et al. 2010), paper 15 (Li et al. 2016), and paper 16 (Humber et al. 2015) all applied game theory to privacy-preserving data publishing. Papers 1, 2, 3, 8, 10, and 15 have 2 players and others have more players. Papers 3, 6, 7, 8, 10, 13, 15, and 16 have empirical experiments based on real datasets and others do not have. Papers 3, 4, 5, 6, 7, 10, 11, and 15 used specific game models such as Stackelberg game and Stochastic game. Most importantly, players in most games proposed in these works have limited options, especially binary options as those in the prisoner's dilemma game. In contrast, in all my game models, I assume the players can choose from hundreds of thousands of options, due to the high dimensionality of the genomic data. Among all these papers, only paper 16 dealt with genomic data, thus is most related to my work.

1.4.7 Computational Aspects of Privacy Games

As the interface of theoretical computer science and game theory, an area known as the algorithmic game theory has exploded phenomenally over the past several decades [225].

Efforts have been made to handle the computation challenges in searching for optimal solutions in privacy games. Blocki et al. [209] presented a polynomial-time approximation scheme to compute a solution that is arbitrarily close to the optimal solution. Due to the effect of the punishment rate they introduced to their model, the optimization problem has quadratic and non-convex constraints. Although for a fixed punishment rate, the induced problem is a linear programming problem, a binary search over values is not feasible. They presented an additive fully polynomial-time approximation scheme to obtain an additive approximation efficiently. The optimal values partition the constraints into two: i) those that are tight, and ii) those in which the probability variables are zero. The partitioning allows a linear number of iterations in each of which sub-problem with quadratic equality constraints is solved. Their approach to computing the Stackelberg equilibrium is based on the multiple linear programming techniques introduced by Conitzer and Sandholm [226].

Duong et al. [223] computed the PSNE in their game model, but they only selected ten strategies (out of 168,440 strategies) for the publisher due to the huge strategy space.

The strategy set in the game model used by Rajbhandari and Snekkenes [190] is significantly smaller than those I used in my models. Their game was solved using mixed-strategy Nash Equilibrium to compute probabilities with which the players exhaustively play each of their strategies.

Shokri [204] solved his Stackelberg game using linear programming. Because the computation cost is cubic in the cardinality of the set of secrets, the author suggested using approximation techniques to speed up the computation. However, the experimental results were not convincing because of the small size of the dataset: Shokri ran experiments on a simulated dataset with ten users and ten differential privacy thresholds.

1.5. Summary of the Dissertation Research

The main goal of my dissertation research is to establish a general game model that can find the best strategy to share a dataset while preserving the data subjects' privacy. To achieve this goal, I built game theoretic models to provide more accurate and practical risk assessment and risk mitigation, with four applications targeting well-known privacy attacks using real datasets in the context of health and genomic data sharing. In doing so, I addressed the modeling and computation challenges of finding the optimal solution or Nash equilibriums of the game models in a high-dimensional environment.

I have applied game theoretic models to solve several real-world problems. For most problems I investigated, the state-of-the-art research was still in the level of attack demonstration. From a certain perspective, my works delved several levels deeper into the problem. First, I proposed a defense model, which is tested using real-world datasets. Second, I optimized the performance of the defense model by efficiently searching through the solution space. Third, I used a game model that also considered the adversary's behavior, which makes the evaluation more accurate. Fourth, I developed effective algorithms to accelerate the game solving. During the process, I faced the challenges to model complex problems as game models and developed algorithms to accelerate the computation. The results demonstrated that game theoretic models and the corresponding solving approaches could be regarded as useful models and methods like the K-anonymity and Differential Privacy.

The relations among four tasks and corresponding game models are illustrated in Figure 1.5. The games in these tasks target different attacks evolved over time and were built one upon another. Short summaries and reviews for all completed tasks are provided as follows. Part of these reviews has been published in a book chapter [227], where I reviewed all game theoretic models applied to privacy-preserving genomic data sharing problems in two scenarios. In the first scenario, a data center shares de-identified individual-level genomic data, while the adversary re-identified those records by linking them to an external dataset (Tasks 1 and 4). In the second scenario, a data center shares summary statistics to the public, and the malicious recipient utilizes a statistical inference attack to detect the membership of targeted individuals (Tasks 2 and 3).

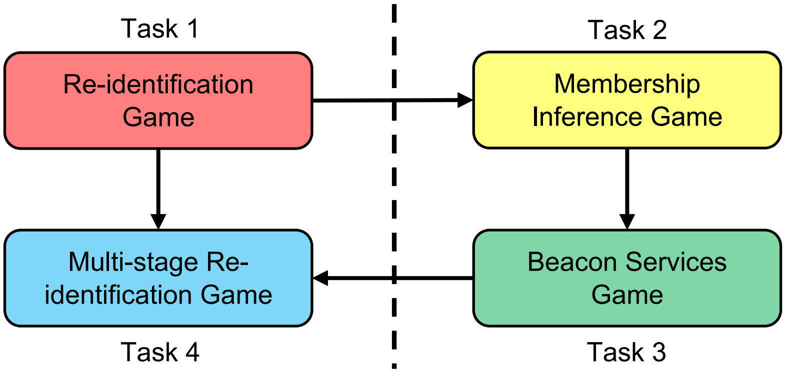


Figure 1.5 The relations among four tasks and corresponding game models

1.5.1 Task 1: Protecting Individual-level Demographic Health Data against Record Linkage Attacks based on a Two-player Stackelberg Game Model (Re-identification Game)

The first task tended to answer the question of how to formalize the interaction between a health data publisher and a potential adversary that intends to re-identify targets from the anonymized dataset. Previous adversary models almost always assume a worst-case scenario where the adversary will always attack. However, I noted that a rational adversary would only attempt to attack if it is profitable. Thus, in this dissertation, I introduced a game theoretic framework to model the cost and benefit relations between a data publisher and an adversary. In the meantime, I investigated the risk of privacy breaches and optimized the data publisher's utility on sharing the health data. Applying the model to a real-world dataset, I compared my publishing policy with existing rule-based policy and other data sharing policies. The generalization strategy space of the data publisher expanded exponentially along with the number and granularity of the attributes in the health data with demographic attributes, which brought computational challenges for me to overcome.

Next, I demonstrate how a game theoretic framework could be utilized to analyze the re-identification risks. The health data sharing process and the re-identification game are illustrated in Figure 1.6.

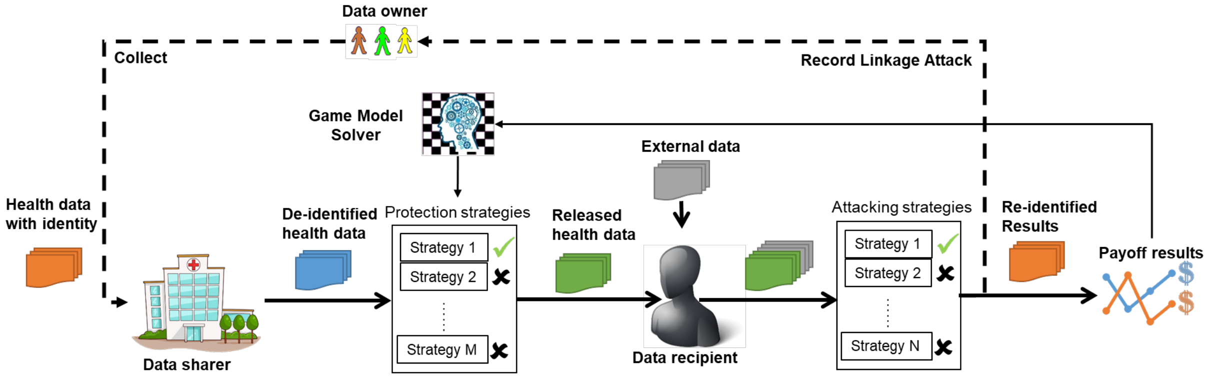


Figure 1.6 An illustration of the health data sharing process and the re-identification game.

The health data sharing process follows a series of steps: 1) The data sharer collects identified health data (including genomic data) from data owners; 2) The sharer releases de-identified health data to a recipient; and 3) The recipient re-identifies targets in the study by linking them to an external dataset upon a set of demographic quasi-identifiers (e.g., age, race, gender). The game components include the following decision points: 1) the recipient selects the optimal attacking strategy given the released data, and 2) the sharer selects the optimal protection strategy by solving the game model.

The NIH requires researchers who are granted funding for genome-based datasets to deposit and share data through online repositories, such as dbGaP. At the same time, the publishers also have an incentive to protect the identities of the individuals who participated in the original research. A malicious recipient of the genomic data would try to re-identify the records in the genetic dataset by linking it to an external dataset and benefit in various ways, such as contacting the participants for marketing purposes, blackmailing the participants using inferred sensitive attributes or publishing the results to gain academic rewards. A two-player Stackelberg game model was used to model this scenario, in which the leader is the publisher, and the follower is the malicious recipient.

The probability that the recipient will successfully re-identify a record is dependent upon the quality (i.e., preciseness and completeness) of the information released by the publisher. A typical technique to sanitize a dataset for anonymization purposes is the generalization technique. Thus, the size of the publisher's strategy space is dependent upon the number of quasi-identifiers (i.e., linking attributes) and generalization levels for each quasi-identifier.

Given a record released at a certain generalization level, the recipient has two options: either attempt the re-identification attack or not. A successful re-identification of a record results in loss to the publisher and gain to the attacker. The recipient will only spend a fixed cost when the attack is attempted. The publisher will always gain a fixed payout for a record, depending on its generalization level.

The solution of the game (i.e., the Nash equilibrium) could be found using two search algorithm: 1) Backward Induction, an exhaustive search algorithm that is more suitable for a relatively small search space, and 2) Lattice-Based algorithm, a heuristic-driven approach that suits a relatively large search space by that pruning nodes.

The experiments are based on a real-world dataset consists of 32,561 US Census records with demographic attributes including age, race, gender, and 5-digit ZIP codes. The generalization levels for each of them are 6, 4, 2, and 6, respectively.

The results show that the publisher's optimal strategy for sharing the genomic dataset with demographic attributes while protecting privacy could always be found using the game theoretic approach. Most importantly, it is actually possible to achieve zero risks. The zero-risk solution shares nearly as much data as the optimal one. It also shares much more data than would be shared under the HIPAA Safe Harbor policy. A publishing strategy that compliant with Safe Harbor yields a higher payoff to the publisher due to the reduced privacy risk. These findings are robust to order-of-magnitude changes in parameters such as gains and losses to the publisher and the recipient. In terms of computational performance, the Lattice-Based algorithm runs much faster than the baseline Backward Induction algorithm and returns almost the same average payoff for the publisher (99.5% of the solutions are optimal).

Moreover, the re-identification game has been implemented and integrated into an open-source toolkit: ARX [228].

1.5.2 Task 2: Protecting Summary-level Genomic Data against Membership Inference Attacks based on a Two-player Stackelberg Game Model (Membership Inference Game)

The second task was still based on the two-player Stackelberg game model. However, in this task, the datasets were extremely high dimensional. One single DNA sequence of each individual human being carries millions of bits of information. Thus, almost every individual's DNA sequence becomes unique, which makes de-identifying the whole genomic sequence very difficult and even impossible. As a result, only summary-level genomic data can be released without the data subject's consent. I examined the state-of-the-art membership inference attack targeting summary-level genomic data introduced by Sankararaman et al. [36]. I proposed a genetic algorithm to search for the huge suppression strategy space for the data publisher. In addition, I used real-world datasets to evaluate the performance of my solutions.

Next, I demonstrate how game theoretic models can be applied to determine the optimal set of genomic data sharing policies against Sankararaman et al.'s membership-inference attack. The genomic data sharing process and the membership inference game are illustrated in Figure 1.7.

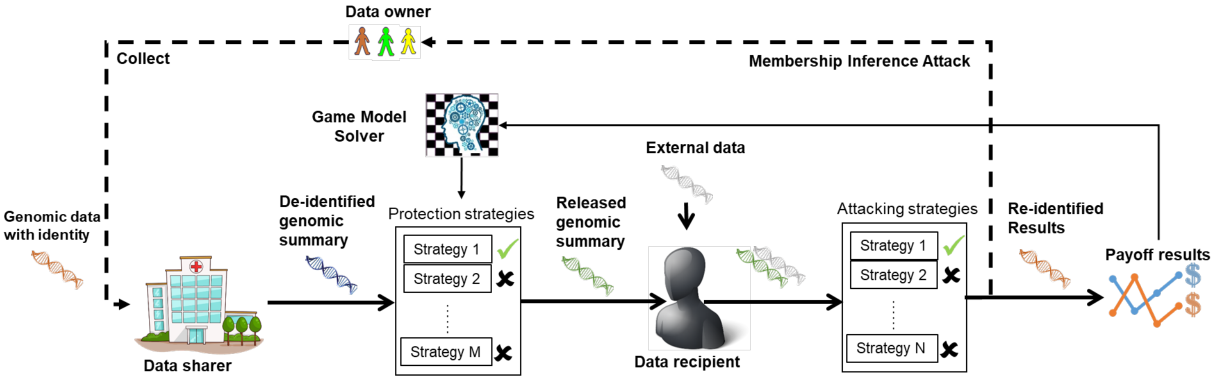


Figure 1.7 An illustration of the genomic data sharing process and the membership inference game. The genomic data sharing process follows a series of steps: 1) The data sharer collects identified genomic data from data owners; 2) The sharer releases de-identified genomic summary data to a recipient; and 3) The recipient infers the membership of targets in the study with the assistance of an external dataset and statistical hypothesis tests (e.g., likelihood ratio test). The game components include the following decision points: 1) the recipient selects the optimal attacking strategy given the released summary data, and 2) the sharer selects the optimal protection strategy by solving the game model.

In this model, there are two players: the sharer, who could be an investigator of a study, and the recipient, who requests and accesses the data. A malicious recipient has the potential to infer the presence of identified genomes in the research study. The sharer gains utility from disseminating data, while the recipient benefits by detecting and exploiting the targets. The costs of an attack, if it is enacted, include the cost of accessing the data and the fines for breaching a DUA (that is, if such deviation is detected). The sharer decides the protections to set in place when releasing information about the single-nucleotide polymorphisms (SNPs), and the recipient chooses to enact an attack (provided it is worth its cost). The interaction between the two players was naturally modeled using a Stackelberg game model. The sharer's optimal strategy, which balances the expected utility and privacy risk, can be found as the equilibrium of the game.

For a large dataset, the strategy space of the sharer can be quite large. As such, the solution to the optimation problem is difficult to discover computationally. Thus, we leveraged a genetic algorithm (GA), which is an optimization approach inspired by evolutionary processes.

The experimental evaluation was conducted on a real dataset of 8,194 individuals from the Sequence and Phenotype Integration Exchange (SPHINX) program in the Electronic Medical Records and Genomics (eMERGE) project [229], with 51,826 genetic variants (e.g., SNPs) being collected. The 1000 Genomes Phase 3 resource [230] was used as the reference population.

The results show that the best policy option is realized in a game theoretic setting that combines an SNP suppression approach with a DUA in terms of the sharer's payoff. In addition, we provided a solution to the game that results in no attack being committed by adding an additional set of constraints to the Stacekberg game model.

To help set a reasonable penalty in the DUA, they further investigated how the results change according to the penalty set in the DUA. As expected, the overall payoff of the sharer increased until the penalty achieves the maximum value the recipient can gain from the attack. However, after a certain point, the growth rate slowed and eventually became very small. This provides an analytical way for the policymaker to choose a preferred penalty level in the DUA.

In another set of experiments, we considered how the prior probability of a target's inclusion in a study influences the results. They set the prior according to four genome sequencing programs: (a) the Precision Medicine Initiative (PMI) (which became the All of Us program) [231], (b) the Million Veteran Program (MVP) [232], (c) the BioVU de-identified DNA repository program of the Vanderbilt University Medical Center [233], and (d) the Rare Diseases Clinical Research Network. Some notable findings include (a) when the prior probability is relatively small, as in PMI, the difference between the DUA policy and the game policies is negligible, (b) the sharer's payoff is negatively correlated with the prior probability, regardless of the policy, and the game policies are the most robust even when the prior probability increases substantially.

These results demonstrate that blending economic, legal, and technical approaches can help strike the right balance between data utility and privacy risk in genomic data sharing and have the potential to revolutionize how policies are designed.

This game model has the potential to be applied to membership inference attacks targeting phenotypic data [234].

1.5.3 Task 3: Practical Privacy Protection of Genomic and Health Data through Beacon Services (Beacon Services Game)

Although my solution performed well in the evaluation in Task 2, it had not been implemented in the real-world yet. In task 3, I had the chance to improve the privacy protection ability of a working genomic data-sharing platform called Beacon service [43] created by the Global Alliance for Genomics and Health (GA4GH), that only allows users to query the presence of any variant in a genomic dataset. An example of the query could be "Does the dataset have any genomes with nucleotide A at position 121,212,028 on chromosome 10?" In this way, the information released from the system is restricted to a minimal level but is still useful because, for instance, the observation of a rare allele in multiple datasets might merit further investigation. A new attack, SB attack, introduced by Shringarpure and Bustamante [86] challenged the security and privacy design of Beacon services. To mitigate the privacy risk posed by SB attack to the Beacon services, the Integrating Data for Analysis, Anonymization, and Sharing (iDASH) National Center for Biomedical Computing designed one of the three tracks of their 2016 Genomic Privacy Protection Challenge to explicitly solve this problem [235]. In the iDASH competition, I developed a practical protection strategy against the SB attack by quantifying the utility and privacy metrics and finding the optimal set of variants for flipping. Later, I introduced a generalized framework that is more representative of the real world. Basically, they designed a measurement that helps to accelerate the search process for the best strategy to protect the genomic data against the SB attack, implicitly using a game theoretic approach. Afterward, I proposed an explicit game theoretic model like the one in Task 2 to defend against the SB attack.

Next, I demonstrate how game theoretic models can be applied to determine the optimal set of genomic data sharing policies against the SB attack. The genomic data sharing process via the Beacon service and the Beacon services game are illustrated in Figure 1.8.

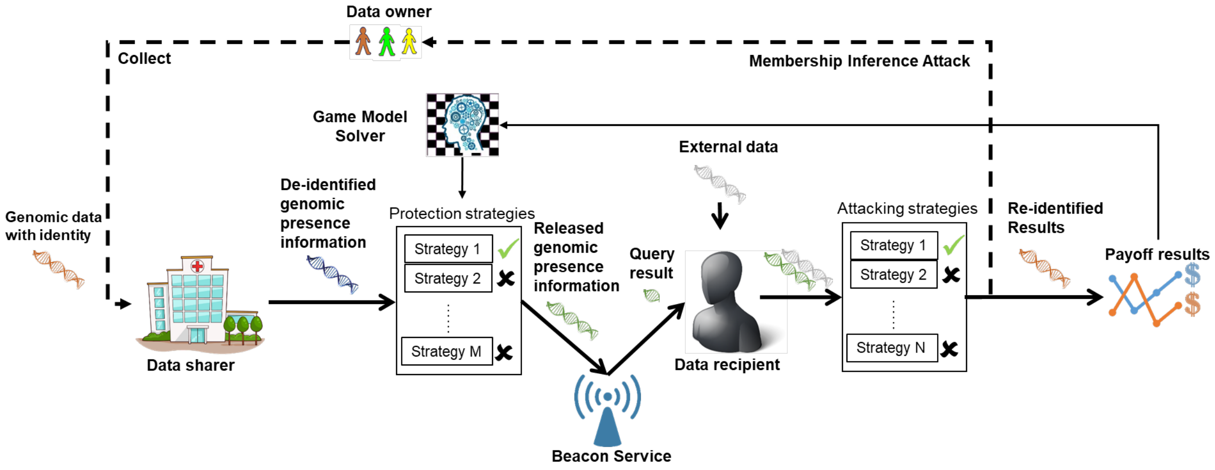


Figure 1.8 An illustration of the genomic data sharing process via the Beacon service and the Beacon services game.

The genomic data sharing process via the Beacon service follows a series of steps: 1) The data sharer collects identified genomic data from data owners; 2) The sharer releases de-identified genomic presence information to Beacon server; 3) The recipient queries the Beacon server for the presence information regarding particular genetic variants of the targets; and 4) The recipient infers the membership of targets in the study with the assistance of an external dataset and statistical hypothesis tests (e.g., likelihood ratio test). The game components include the following decision points: 1) the recipient selects the optimal attacking strategy given the released presence information, and 2) the sharer selects the optimal protection strategy by solving the game model.

As in the membership-inference setting, there are two parties in consideration: the defender and the attacker. The attacker is a malicious user launching the SB attack, while the defender is the data holder sharing the genomic data while mitigating the privacy risk of the SB attack. Thus, a Stackelberg game model is again a natural framing, where the leader is the defender, and the follower is the attacker.

In the SB attack, given a set of Beacon responses, the attacker relies upon a log-likelihood ratio test (LRT) to infer the presence of a targeted genome in the dataset. In the iDASH variation of the SB attack, it is assumed that the attacker knows the alternative allele frequency (AAF) of all single nucleotide variants (SNVs) in the underlying population.

In the iDASH Challenge, it was assumed that the defender is not aware of the attacker's query sequence a priori and does not keep track of the queries. As a result, we need to find a defender's strategy that is independent of the attacker's query sequence. Our experiments for the iDASH Challenge examined a Strategic Flipping strategy which we proposed to optimize the utility-privacy tradeoff. The utility is measured as the number of queries that the defender responds to truthfully. We introduced the notion of discriminative power for each SNV in the pool. The discriminative power represents an SNV's ability to distinguish the records in the pool of individuals behind the beacon from a reference dataset. We defined a differential discriminative power for each SNV in the pool, which represents the difference between its discriminative power before and after a flip. The top k percent of the SNVs in the pool, ranked by their differential discriminative power, will have their query responses flipped. We used a greedy algorithm to search the defender's strategy space for a local optimum. We started from the result of the Top-k Flipping step and keep searching until no strategy with better effectiveness can be found.

The experimental evaluation was conducted with a real dataset based on the first 400,000 SNVs in Chromosome 10. The pool is composed of 250 individuals randomly selected from the 2,504 individuals in Phase 3 of the 1000 Genomes Project [230]. The reference includes 250 individuals randomly selected from the remaining individuals in Phase 3 of the project.

The experimental results showed that the proposed method outperforms all posited baseline and state of the art methods (that were applicable to real-world scenarios) regardless of how key parameters that drive the attack (e.g., the effectiveness measure, the number of records behind the beacon, and the attacker's estimate of allele frequency) vary. In most scenarios, the advantages of the proposed method over other alternative methods are substantial. The effectiveness of our proposed method is larger than all of the alternative methods when the value for the k parameter is smaller than five. It is anticipated that using a game theoretic approach will lead to a better payoff in practice, which represents a better balance between the data utility and privacy risk.

1.5.4 Task 4: Protecting Genomic and Demographic Health Data against Multi-stage Re-identification Attacks based on a Two-player Stackelberg Game Model (Multi-stage Re-identification Game)

The fourth task was still based on the two-player Stackelberg game model. However, in this game, the adversary utilized more data resources and executed multi-stage re-identification attacks. I examined my game model against a state-of-the-art two-stage re-identification attack targeting genomic data introduced by Gymrek et al. [31]. The Gymrek attack was also known as the surname inference attack, which was able to infer surname from a small portion of a de-identified genome sequence. With the successfully inferred surname, the adversary can improve the performance of the record linkage attack. This investigation into a game model against a two-stage attack, rather than a one-stage attack, demonstrated that the modeling and solving of the game with a multi-stage attack were more complex and time-consuming, which is a challenge for the game solver. However, the increased stage makes the attack more vulnerable to my game theoretic strategy because the adversary could be tricked to reach a wrong intermediate result to mitigate the privacy risk. Specifically, I masked individual-level demographic and genomic attributes in the targeted genomic dataset. Additionally, I proposed a pruning algorithm to search the huge data masking strategy space for the data sharer. In addition, I used real-world datasets, as used by Gymrek et al., and large-scale synthetic datasets to evaluate the effectiveness of my model and the efficiency of my method. The results demonstrate that maximal utility can be achieved if sharing partial data is allowed, in which case most data can be shared with little privacy sacrificed. Through extensive sensitivity analyses on essential parameters, I provide insights into risk mitigation directions for corresponding stakeholders.

Next, I demonstrate how game theoretic models can be applied to determine the optimal set of genomic data sharing policies against Gymrek et al.'s surname inference and re-identification two-stage attack. The genomic (and demographic) data sharing process and the multi-stage re-identification game are illustrated in Figure 1.9.

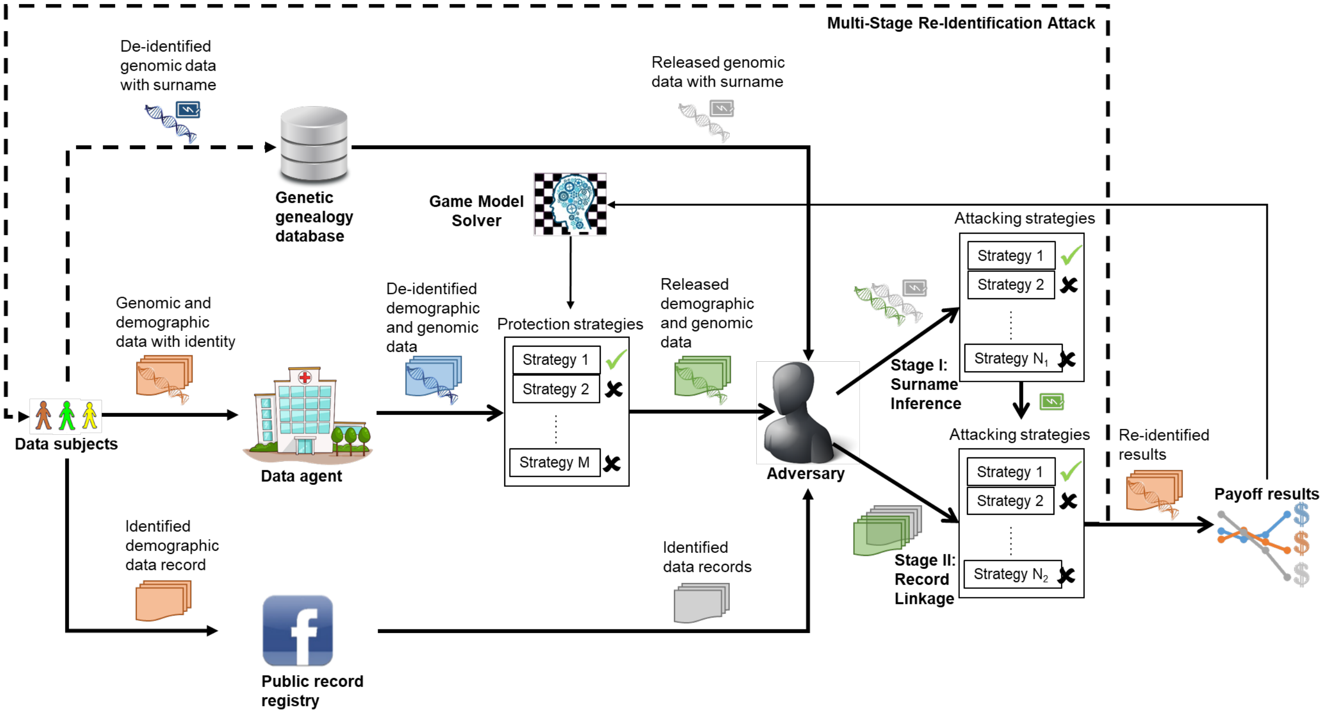

Multi-Stage Re-Identification Attack
De-identified genomic data with surname
Released genomic data with surname
Genetic genealogy database
Game Model Solver
Attacking strategies
Strategy 1
Strategy 2
Strategy N1
Genomic and demographic data with identity
De-identified demographic and genomic data
Protection strategies
Strategy 1
Strategy 2
Strategy M
Released demographic and genomic data
Stage I: Surname Inference
Data subjects
Data agent
Adversary
Attacking strategies
Strategy 1
Strategy 2
Strategy N2
Re-identified results
Payoff results
Identified demographic data record
Identified data records
Stage II: Record Linkage
Public record registry

Figure 1.9 An illustration of the genomic (and demographic) data sharing process and the multi-stage re-identification game.

The data sharing process follows a series of steps: 1) The data subject shares identified demographic record to a public record registry; 2) The data subject may share de-identified genomic data with surname to a genetic genealogy database; 3) the data subject or a data agent shares de-identified genomic and demographic data to the adversary; 4) The adversary infers the surname of the data subject with the help of the released genomic data with a surname from the genetic genealogy database (Stage I); and 5) The adversary re-identifies the data subject by linking to an external dataset upon a set of demographic attributes (e.g., age, state) and the inferred surname (Stage II). The game components include the following decision points: 1) the adversary selects the optimal attacking strategy given the released data, and 2) the sharer selects the optimal protection strategy by solving the game model.

In this model, there are two players: a data subject or their agent (e.g., a healthcare organization that collected their data) chooses how much of their genomic data to share in a public repository, such as 1000Genomes [230], OpenSNP [61], or Personal Genome Projects [60], and the adversary, has the incentive and means to determine the identities of subjects in anonymized shared datasets. Other data about the subject that is already available to the public, possibly at some cost.

The interaction between the two players was naturally modeled using a Stackelberg game model in which the data subject acts as a leader who chooses how much of their genomic data to share, and the adversary is the follower who obtains the shared data and then decides whether to execute an attack. The sharer's optimal strategy, which balances the expected utility and privacy risk, can be found as the equilibrium of the game.

The experimental evaluation was conducted on a real dataset of 156,761 individuals from the Ysearch dataset and Craig Venter's genomic record [31], with 100 Y-STRs being collected. The Intelius.com was used as the public record registry. To evaluate the effectiveness of our methods in a larger and more controllable environment and to facilitate replications of our investigation without privacy concerns, we simulated a genetic genealogical population of 600,000 individuals with 20 attributes.

By comparing variations of our masking game to several baseline models, the results show that the best policy option is realized in a game theoretic setting in terms of the data subjects' average payoff. More specifically, about 86% of data is shared, and fewer than 1% of data subjects are expected to be re-identified in the masking game. In contrast, about 30% of data is shared, and about 6% of data subjects are expected to be re-identified in an alternative model. Notice that, without any protection, about 76% of data subjects are expected to be re-identified. In addition, we provided a solution to the game that results in no attack being committed by adding an additional set of constraints to the Stacekberg game model. At last, we performed extensive sensitivity analyses to show what direction a data subject should care more about for each parameter and show that our findings are robust to fluctuations of most model parameters.

These results demonstrate that although an additional stage can substantially increase the accuracy of the re-identification when there is no protection, it makes the attack more vulnerable to our game theoretic protection. Second, most people would share most of their data if sharing partial data is permitted. This finding is intriguing because it suggests that providing data subjects with options could encourage a greater degree of data sharing while preventing re-identification.

1.6 Dissertation Overview

The remainder of this dissertation is organized as follows: In Chapter 2, I formalize the re-identification game model in Task 1 and find the best generalization strategy for the data publisher in a simulation to protect a real-world demographic dataset. Part of this chapter has been published as a peer-reviewed journal article [236]. In Chapter 3, I formalize the genomic privacy game model in Task 2 and find the best suppression strategy for the data publisher to protect a high-dimensional real-world genomic dataset. Part of this chapter has been published as a peer-reviewed journal article [237]. In Chapter 4, I investigate a practical protection service called Beacon service, improve its protection ability against the SB Attack, and formalize a game model to counteract the attack. Part of this chapter has been published as a peer-reviewed journal article [238]. In Chapter 5, I formalize the multi-stage re-identification game model in Task 4 and find the best masking strategy for the data subject based on large-scale simulated datasets and real-world genomic datasets. Part of this chapter has been submitted as a journal article.

Chapter 2

RE-IDENTIFICATION GAME

Given the potential wealth of insights in personal data the big databases can provide, many organizations aim to share data while protecting privacy by sharing de-identified data, but are concerned because various demonstrations show such data can be re-identified. Yet these investigations focus on how attacks can be perpetrated, not the likelihood they will be realized. This chapter introduces a game theoretic framework that enables a publisher to balance re-identification risk with the value of sharing data, leveraging a natural assumption that a recipient only attempts re-identification if its potential gains outweigh the costs. We apply the framework to a real case study, where the value of the data to the publisher is the actual grant funding dollar amounts from a national sponsor, and the re-identification gain of the recipient is the fine paid to a regulator for violation of federal privacy rules. There are three notable findings: 1) it is possible to achieve zero risk, in that the recipient never gains from re-identification, while sharing almost as much data as the optimal solution that allows for a small amount of risk; 2) the zero-risk solution enables sharing much more data than a commonly invoked de-identification policy of the U.S. Health Insurance Portability and Accountability Act (HIPAA); and 3) a sensitivity analysis demonstrates these findings are robust to order-of-magnitude changes in player losses and gains. In combination, these findings provide support that such a framework can enable pragmatic policy decisions about de-identified data sharing.

2.1 Introduction

Our ability to collect and analyze personal data has grown dramatically over the past decade, a trend that shows no sign of slowing. While this information enables a wide range of institutions to perform novel research in biomedicine and the social sciences, big data has become big business. There is a rapidly expanding market for sharing and selling data for secondary analysis, driven by profits as well as grant funding agencies aiming to support transparency and research productivity [239, 240]. This movement towards data sharing *en masse* must be accomplished by provisions that appropriately protect the privacy expectations of the individuals to whom the data corresponds [241, 242]. Historically, this has been achieved by removing aspects of an individual's identity (e.g., suppression of certain demographics), a practice codified in laws around the world, such as the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [5] and the General Data Protection Regulation [65]. Such laws provide specific guidance on how to share *de-identified* or *anonymized* information.

Table 2.1 The attributes removed, or generalized, to satisfy the HIPAA Safe Harbor policy.

	Attributes
(A)	Names
(B)	All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
(C)	All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
(D)	Telephone numbers
(E)	Fax numbers
(F)	Email addresses
(G)	Social security numbers
(H)	Medical record numbers
(I)	Health plan beneficiary numbers
(J)	Account numbers
(K)	Certificate/license numbers
(L)	Vehicle identifiers and serial numbers, including license plate numbers
(M)	Device identifiers and serial numbers
(N)	Web Universal Resource Locators (URLs)
(O)	Internet Protocol (IP) addresses
(P)	Biometric identifiers, including finger and voice prints
(Q)	Full-face photographs and any comparable images
(R)	Any other unique identifying number, characteristic, or code, except as permitted by paragraph §164.514(c)

Notice that the policy requires the removal of explicit identifiers (e.g., the names of the corresponding individual or of their relatives, employers, or household members), quasi-identifiers (e.g., attributes that could potentially be linked to identify them, such as demographics), and unique codes (e.g., medical record numbers).

The HIPAA Privacy Rule [5] states that only "individually identifiable" information is covered by the regulation. It goes on to state that data is no longer subject to the regulation when it is de-identified. Specifically, this is defined as "information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable." The Privacy Rule then provides alternative implementation specifications, one of which must be followed to ensure that data meets the de-identification definition. The first implementation specification is the Safe Harbor policy, which enumerates 18 attributes that must be generalized or suppressed. The details are provided in Table 2.1. In this study, we focus on Safe Harbor's perspective on demographics, which states that 1) all ZIP codes must be rolled back to their initial three digits and, further, that codes with populations of less than 20,000 individuals must be grouped into a single code of 000** and 2) ages over 90 must be aggregated into a top-coded age group of 90+.

However, a growing collection of investigations demonstrate how de-identified information can be re-identified to the individuals from which it was derived (e.g., via demographics [27, 28, 29], genome sequences [30, 31], mobility patterns [243], and social networks [244, 245]). This phenomenon has led to claims that de-identification fails to adequately protect privacy [246, 247]. It has been suggested that society should adopt new definitions of privacy (e.g., [140]) and new legal mechanisms to mitigate misuse and abuse of identified personal information (e.g., [248, 249, 250]). However, such calls for a revolution are based on demonstrations of what is *possible* and not necessarily what is *probable*. To date, there has been little evidence that such attacks will be realized in practice for any reason other than demonstration [173], and there are many reasons why an adversary may choose to forgo a re-identification attempt in the first place [162]. Rather than viewing de-identification as a dichotomous problem of "broken" or not, it should be considered as a matter of risk over a continuous range.

HIPAA acknowledges this fact by stating in the second implementation specification, which the game theoretic perspective is designed to address, that "information is not individually identifiable health information only if: 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." Here, we highlight the fact that the notion of de-identification is explicitly tied to a risk assessment that accounts for the capabilities of a reasonable adversary. We believe this is a clear justification for the application of the game-based approach to the de-identification problem.

Recently, risk-based approaches have been proposed for decision support in de-identification [174]. However, it should be recognized that a key source of re-identification risk is a decision on the part of the data recipient to *attempt* re-identification. It is natural that the anticipated data recipient, in most practical data sharing settings, will only attempt re-identification if there is a tangible economic benefit (e.g., monetary gain) to doing so. Consequently, we model the data recipient as an attacker who will choose to attempt re-identification if the associated expected benefits outweigh the costs. The costs of re-identification can come from numerous sources, including those associated with purchasing data to execute an attack by linking on common features (e.g., residual demographics), as well as the time and resource utilization necessary to run the attack. We rely on this model of a data recipient who is motivated by the economic gain as a part of a game theoretic framework that 1) computes the best data sharing strategy for the publisher and 2) accounts for the associated incentives of data recipients to attempt re-identification.

We illustrate our framework through a case study using the demographics reported in a publicly available dataset from the U.S. Census Bureau. In this game, the benefits to the data publisher are proportional to the amount of funding provided to investigators via National Institutes of Health (NIH) grants, and the losses are proportional to fines paid to the U.S. Department of Health and Human Services (HHS) for privacy violations in the form of information security breaches. We begin by computing an optimal data sharing solution in this setting. Interestingly, in this solution, the data recipient has the incentive to try to re-identify a significant fraction of records. Nevertheless, the likelihood of a *successful* re-identification is quite low, and the only reason there is any incentive is that it is cheap to attempt it (we set the per-record cost to be only $4, which is roughly the cost of using an online data broker). Our most significant finding, however, is that it is actually possible to achieve *zero* risk, in the sense that the attacker has no incentive to re-identify any record in the published data. Remarkably, this zero-risk solution shares nearly as much data as the optimal, but slightly more risky, data sharing policy. Moreover, the zero-risk solution also shares significantly more data than would be shared under the HIPAA-recommended Safe Harbor de-identification policy—even though Safe Harbor also incurs non-zero re-identification risk. Motivated by the ubiquitous use of Safe Harbor as minimal (and, therefore, safe) HIPAA compliance, we additionally consider publishing strategies that satisfy Safe Harbor standards. In this highly restricted setting, we find a policy which, while compliant with Safe Harbor, actually yields a higher utility to the publisher, primarily due to a significantly reduced re-identification risk. Indeed, we observe that these gains are quite substantial for a non-trivial fraction of individuals. Finally, we execute an extensive sensitivity analysis of our findings and observe that they are robust to order-of-magnitude changes in both gains and losses of the publisher and data recipient.

2.2 Methods

2.2.1 Model

Consider an organization (e.g., an academic medical center) that aims to release data with as much fine-grained information as possible, but simultaneously account for the risk of re-identification and concomitant fallout. The re-identification risk has two sources: first, a decision by the data recipient to *attempt* re-identification, aimed at achieving some goal which the publisher views as undesirable (such as imposing a fine on the publisher, or selling re-identified data), and second, the probability that a re-identification attempt succeeds. We model the data recipient as an intelligent attacker who can access external resources at a fixed cost to perform a *linkage* attack, where the attributes shared by the data sets are leveraged to connect the corresponding records, and who only attempts re-identification if his associated benefits exceed the costs (which can also include linking and curation costs). If re-identification of a record in the data set is attempted, the probability it succeeds derives from the fact that the external resource could contain the corresponding individual's identity. For example, de-identified medical records have been linked to voter registration lists to identify individuals and disclose potentially sensitive test results [166].

Table 2.2 A summary of the notation used in this chapter.

Notation	Meaning
m	The number of shared attributes in both released data and external source
$g = \{g_1, \dots, g_m\}$	The publisher's strategy on generalization levels for one record
h_f	The number of levels in the generalization hierarchy for attribute f
r	The number of the publisher's available strategies for one record
$v(g)$	The benefit that the publisher receives from sharing data using strategy g
V	The benefit that the publisher receives by sharing the record in its original form
L	The publisher's loss for one record due to a successful re-identification
c	The adversary's cost to launch a re-identification attack towards one record
$\pi(g)$	The probability the adversary re-identifies one record successfully given strategy
$U_p(g)$, $U_a(g)$	The publisher's and the adversary's payoffs, given strategy g
$a(g)$	The adversary's strategy given strategy g
$GI(g)$	The generalization intensity of the publisher's strategy given strategy g

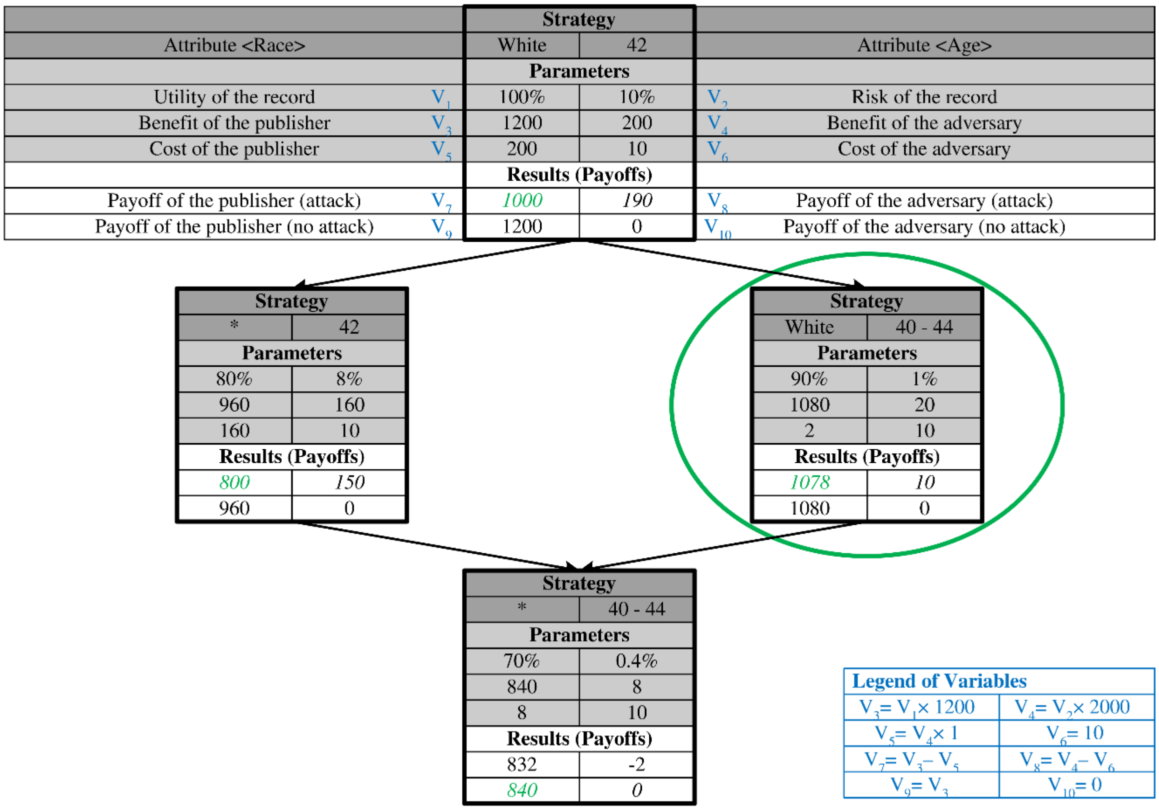
The ability of the data recipient to successfully re-identify a record hinges on the information released to him by the publisher: the more precise and complete the information, the more likely it would be that a re-identification attack succeeds. In formal notation, we let g be the representation of a given record that is released to a recipient, and let $\pi(g)$ be the probability of successful re-identification, should it be attempted. The space of data representations we consider involves *attribute generalization hierarchies*, one of the most common paradigms in data de-identification [132]. For example, let us take an individual's age. An age of 22 can be retained in the most specific form, or generalized to a range [20-25], [20-29], [0-50], or * (i.e., any age). In this case, we say the age generalization hierarchy has 5 levels. We assume that such a hierarchy is given for each attribute, so that the publisher's choice is the level of specificity within it. For attribute f, we use g_f and h_f to denote the specific level and the number of levels in the corresponding hierarchy, respectively. In the above example, if the publisher chooses to release the attributes in the most specific form, our representation will be $g_f = 0$, and $h_f = 5$. Table 2.2 summarizes the notation that is useful in this chapter.

If a record is released at specificity g, the recipient has two options: either attempt re-identification, incurring a fixed cost c, or not. A successful re-identification of a record results in a loss to the publisher which we denote by L, and a gain to the attacker (recipient); to keep things simple, we let L also be the attacker's gain, although generalization is direct. If a re-identification fails, or is not attempted, nothing is gained by the data recipient (besides the data, of course) or lost by the publisher. The latter, however, will always obtain a fixed payout $v(g)$ for a record that is shared at specificity g. The expected payoffs of both the publisher and the data recipient (U_p and U_a, respectively) are shown in Table 2.3.

Table 2.3 Payoff functions of the publisher and adversary for a fixed data sharing strategy.

	no attack	**attack**
$U_p(g)$	$v(g)$	$v(g) - L\pi(g)$
$U_a(g)$	0	$L\pi(g) - c$

To make the game concrete, Figure 2.1 depicts an example of a simplified version. Here, the publisher has four actions (or strategies) to obfuscate an individual's record, which correspond to the amount of detail they are willing to reveal about certain personal characteristics. Given each action, the adversary has two responses: i) attack or ii) not. In Figure 2.1, each node represents an action of the publisher and each edge represents a partial order relation between two nodes. Both the risk and the utility of the publisher's action decreases from top to bottom. To maximize her payoff for each action (marked in green), the publisher's optimal action is to generalize the age attribute (marked by the green ellipse), and while the adversary's best response is to mount an attack.



		Strategy			
Attribute <Race>		White	42		Attribute <Age>
		Parameters			
Utility of the record	V_1	100%	10%	V_2	Risk of the record
Benefit of the publisher	V_3	1200	200	V_4	Benefit of the adversary
Cost of the publisher	V_5	200	10	V_6	Cost of the adversary
		Results (Payoffs)			
Payoff of the publisher (attack)	V_7	*1000*	*190*	V_8	Payoff of the adversary (attack)
Payoff of the publisher (no attack)	V_9	1200	0	V_{10}	Payoff of the adversary (no attack)

Strategy	
*	42
Parameters	
80%	8%
960	160
160	10
Results (Payoffs)	
800	*150*
960	0

Strategy	
White	40 - 44
Parameters	
90%	1%
1080	20
2	10
Results (Payoffs)	
1078	*10*
1080	0

Strategy	
*	40 - 44
Parameters	
70%	0.4%
840	8
8	10
Results (Payoffs)	
832	-2
840	*0*

Legend of Variables	
$V_3 = V_1 \times 1200$	$V_4 = V_2 \times 2000$
$V_5 = V_4 \times 1$	$V_6 = 10$
$V_7 = V_3 - V_5$	$V_8 = V_4 - V_6$
$V_9 = V_3$	$V_{10} = 0$

Figure 2.1 An illustrative example of the re-identification game with four data sharing strategies.

In game theoretic terms, the environment we have constructed is a Stackelberg game, in which the publisher first releases the data represented at specificity g to the data recipient, who subsequently decides whether or not to attempt re-identification. The solution to this game—a Stackelberg equilibrium—entails a publisher's decision about an *optimal* representation g that maximizes her expected utility over all possible representations, balancing the benefits of releasing as much data as possible ($v(g)$) and the risk of re-identification, determined by the data recipient's decision. We play out this game independently for each record in the data set.

The legal ambiguity of the data de-identification policy landscape warrants several extensions to the basic model above, which we will henceforth call the *Basic Game*. The first is a *No-Attack* game. In this version, we constrain the publisher to choose a representation so that a data recipient driven by economic incentives will never choose to attempt re-identification. The second is *SH-Friendly*, in which we constrain representations to be at least as strict as the Safe Harbor (SH) de-identification guideline in HIPAA so that they could easily explain to an Institutional Review Board (or another authority) why they have selected to share such data. Since both these extensions constrain the set of options available to the publisher, it can be proven they are suboptimal in terms of the underlying cost-benefit tradeoffs the publisher's faces, as captured by the model. However, institutional and regulatory oversight may effectively impose these constraints, and our analysis provides decision support for data publishers in such environments, as well as a quantitative assessment of the impact such constraints have on the value of de-identified data.

2.2.2 Solving the Game

We propose two approaches to solving the re-identification game. The first is *backward induction*, in which we consider the adversary's response, and corresponding player utilities, for each possible choice of publisher's generalization level, g. We then select the generalization level which maximizes the publisher's utility. Since backward induction requires an exhaustive search through the combinatorial space of all feasible representations of data, it clearly would not scale when the number of potentially identifying attributes (often termed *quasi-identifiers*) is large. Consequently, we develop a *lattice-based search* heuristic algorithm that takes advantage of the fact that generalization levels of the attributes form a lattice.

Table 2.4 Algorithm 1: Backward Induction Search (BIS) Algorithm.

```
Input: $G = \{g\}$, the set of strategies for the publisher; $L$, the publisher's loss; $c$, the adversary's cost;
$W = \{v\ (g)\}$, the set of benefits for the publisher; $\Pi = \{\pi\{g\}\}$, the set of probability of successful attack
Output: $g^*$, the publisher's best strategy
$g \leftarrow$ 1st strategy in $G$
$g^* \leftarrow g$
if $L\pi(g) \leq c$ then
    $U_m \leftarrow v(g)$
else
    $U_m \leftarrow v(g) - L\pi(g)$
end if
while $g.next \neq NULL$ do
    $g \leftarrow g.next$
    if $L\pi(g) \leq c$ then
        $U_d \leftarrow v(g)$
    else
        $U_d \leftarrow v(g) - L\pi(g)$
    end if
    if $U_d > U_m$ then
        $U_m \leftarrow U_d$
        $g^* \leftarrow g$
    end if
end while
return $g^*$
```

2.2.2.1 Backward Induction Search Algorithm

Algorithm 1 in Table 2.4 reports on the pseudocode for the Backward Induction Search (BIS) method. This is a brute force exploration of the strategy space that guarantees the discovery of the optimal strategy. The time complexity of this algorithm is $O(r) = O(\prod_f h_f)$. The running time increases linearly with the size of the publisher's strategy space r which is a product of h_f the number of generalization levels for each attribute f.

2.2.2.2 Lattice-Based Search Algorithm

The BIS approach requires an exhaustive search of all possible strategies. As the number of attributes and the number of generalization levels for each attribute grow, the number of strategies will overwhelm the search making it impossible to search every possible strategy in an efficient manner. As such, we devised a heuristic-driven approach that takes advantage of a natural order to the strategies. Specifically, the benefit of the publisher $v(g)$ and the probability of success $\pi(g)$ have a partial order on the set of the publisher's strategies along the generalization level. For instance, imagine strategy g^i has at least one attribute that is more general than the corresponding attribute of another strategy g^j while other attributes are in the same generalization level. In this case, the benefit $v(g^i)$ and the probability of success $\pi(g^i)$ will never be larger than the benefit $v(g^j)$ and probability $\pi(g^j)$, respectively.

As such, the illustrative example in Figure 2.1 can be regarded as a part of a lattice. The top node of the lattice indicates a strategy sharing all attributes without generalization, while the bottom node of the lattice indicates a strategy of sharing no data.

For a database with only two attributes Age and Race, let the domain generalization hierarchy (DGH) for the attribute Age be the one shown in Figure 2.2 and the DGH for the attribute Race be the one shown in Figure 2.3, then the whole lattice of the strategy space for publishing record <White, 42> has a structure as shown in Figure 2.4.

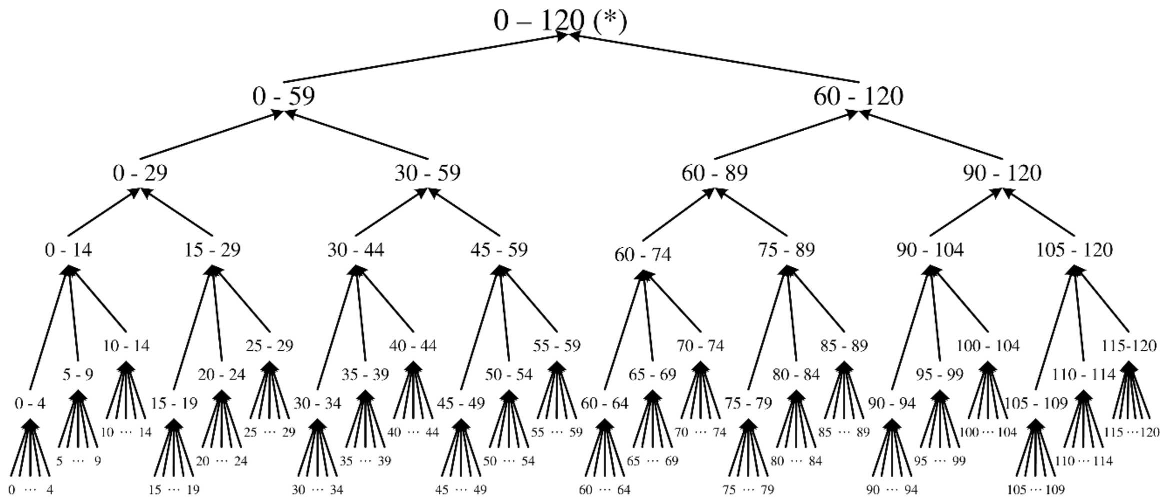


Figure 2.2 The Domain Generation Hierarchy (DGH) for the attribute Age in the case study.

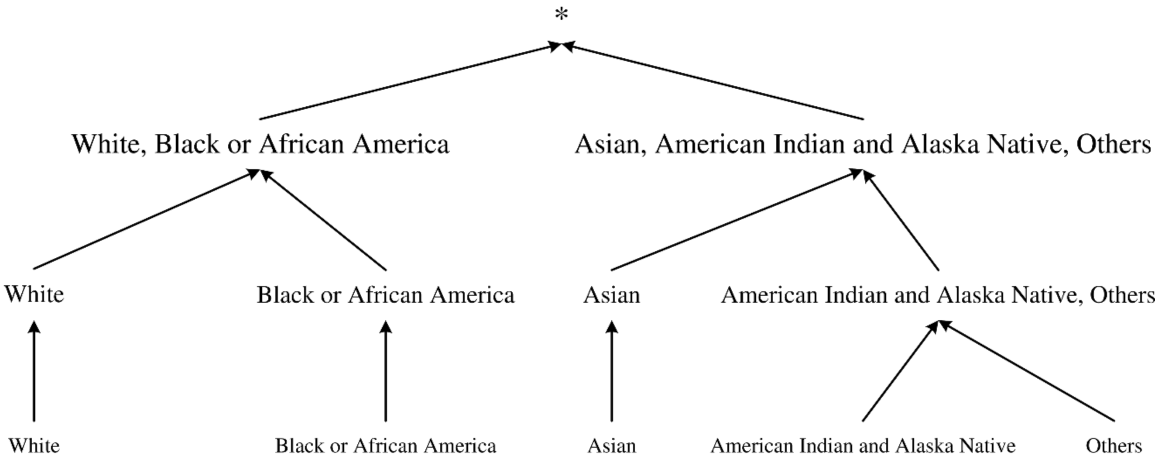

*
White, Black or African America
Asian, American Indian and Alaska Native, Others
White
Black or African America
Asian
American Indian and Alaska Native, Others
White
Black or African America
Asian
American Indian and Alaska Native
Others

Figure 2.3 The Domain Generation Hierarchy (DGH) for the attribute Race in the case study.

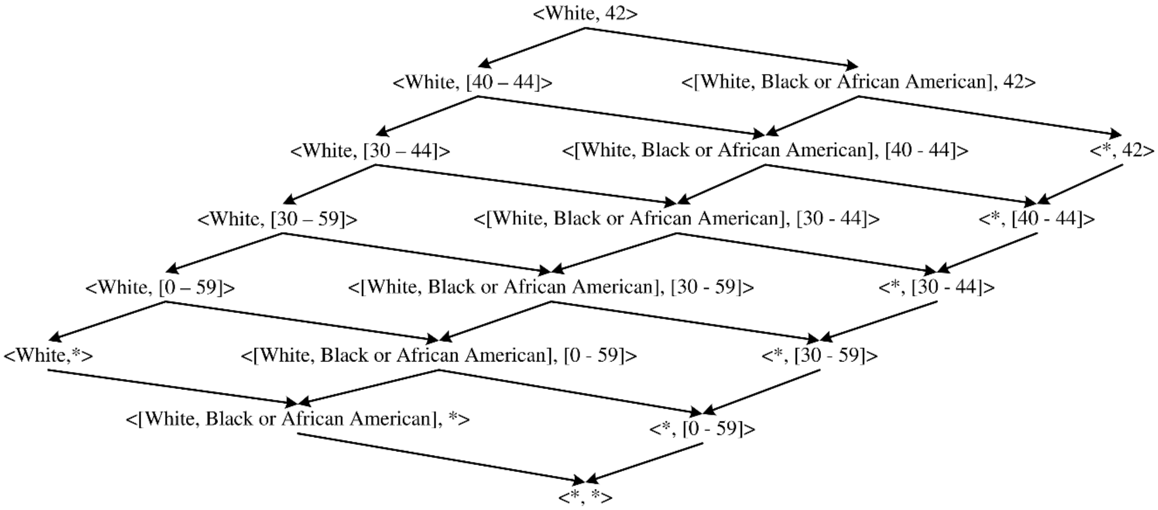

<White, 42>
<White, [40 – 44]>
<[White, Black or African American], 42>
<White, [30 – 44]>
<[White, Black or African American], [40 - 44]>
<*, 42>
<White, [30 – 59]>
<[White, Black or African American], [30 - 44]>
<*, [40 - 44]>
<White, [0 – 59]>
<[White, Black or African American], [30 - 59]>
<*, [30 - 44]>
<White,*>
<[White, Black or African American], [0 - 59]>
<*, [30 - 59]>
<[White, Black or African American], *>
<*, [0 - 59]>
<*, *>

Figure 2.4 The illustrative lattice for a database with two attributes Race and Age.

Table 2.5 Algorithm 2: Lattice-Based Search (LBS) Algorithm.

```
Input: $G = \{g\}$, the set of strategies for the publisher; $L$, the publisher's loss; $c$, the adversary's cost;
$W = \{v\ (g)\}$, the set of benefits for the publisher; $\Pi = \{\pi\{g\}\}$, the set of probability of successful attack
Output: $g$, the publisher's best strategy
$top \leftarrow [0, 0, \ldots, 0]$
$g \leftarrow top$
while $g.children \neq \emptyset$ do
    if $L\pi(g) \leq c$ then
        return $g$
    end if
    $U_m \leftarrow v(g) - L\pi(g)$
    $g_m \leftarrow g$
    for all $g_c \in g.children$ do
        if $v(g_c) - L\pi(g_c) \geq U_m$ then
            $g_m \leftarrow g_c$
            $U_m \leftarrow v(g_c) - L\pi(g_c)$
        end if
    end for
    $g \leftarrow g_m$
end while
return $g$
```

Our lattice-based search (LBS) algorithm is shown in Algorithm 2 in Table 2.5 and described as follows. The algorithm starts the search from the top node. If $L\pi(g) < c$, then the adversary will not attack given this strategy nor any strategy represented by any descendent of this node. Thus, all descendants of this node are pruned and the algorithm returns strategy g. Otherwise, it searches through every child of the current node and continues with the child with the largest payoff, and prunes the other children. The algorithm halts when either 1) the payoff of the current node is larger than the payoffs of its children or 2) the current node has no children. The time complexity of this algorithm is $O(m) = O(\sum_f h_f)$ which is much smaller than the one of backward induction; however, it cannot guarantee to find the global optimum. Nevertheless, the algorithm can be improved to find the global optimum by continuing the search among the set of nodes that have not been searched or pruned until the set is empty.

2.3. Results

2.3.1 Dataset and Experimental Setup

We evaluate the re-identification game framework in two contexts. First, we compare the data sharing policies computed by the framework to the HIPAA Safe Harbor de-identification policy in a case study. This case study parameterizes the system with evidence-based benefits, costs, and loss values. Second, we perform a sensitivity analysis to understand the relationship between the parameters and the strategies played by the publisher and the adversary.

2.3.1.1 Dataset

Our experiments focus on the Adult dataset, a publicly-accessible extract of the 1994 U.S. Census database [251]. The dataset consists of 48,842 records on 14 personal attributes. For this study, we remove all records with missing values, yielding a dataset of 32,561 records. We use traditional demographics (commonly exploited in re-identification studies) – *Age*, *Race*, and *Gender*, but all ages above 90 were published as one group (90+). As such, we use the publicly available U.S. Census data to disaggregate these ages into years 90 through 120. To compare how changes in geographic features influence identifiability, we add an additional attribute, 5-digit ZIP codes in the state of Tennessee, which was obtained from U.S. Census Bureau's dataset [252]. We add such data because, according to Safe Harbor, only the first 3 digits of a ZIP code can be disclosed, provided it has at least 20,000 inhabitants. Specifically, for each <Age, Race, Gender> combination, we assign a ZIP code proportionally to the probability distribution reported in the U.S. Census Bureau's PCT12A through PCT12G tables.

2.3.1.2 Parameters

It should be recognized that the publisher's benefit is affected by the information loss of the released record. To measure information loss, we define an entropy-based metric IL as follows. Given a strategy of the publisher g which generalizes an attribute f to an interval $[q_l(f,g), q_h(f,g)]$, we assume that A_f, the original value of the attribute f, is a uniformly distributed random variable in the interval. The probability that A_f equals q where $q \in [q_l(f,g), q_h(f,g)]$is computed as $\mathrm{P}(q,f,g) = 1/size(q_l(f,g), q_h(f,g))$.

The total information loss is computed as the sum of the entropy for each attribute as

$$IL(g) = \sum_f \sum_q -P(q,f,g)\, log\big(P(q,f,g)\big) = \sum_f -\,log\left(\frac{1}{size(q_l(f,g), q_h(f,g))}\right). \qquad (2.1)$$

We normalize IL by dividing by its maximal value. This corresponds to the scenario where every attribute A_f is generalized to the entire domain D_f. The maximal value is computed as

$$max\big(IL(g)\big) = \sum_f -\,log\left(\frac{1}{size(D_f)}\right). \qquad (2.2)$$

Since the publisher's benefit $v(g)$ will decrease as the information loss increases, we estimate it using a linear function in Equation 2.3, in which V corresponds to the benefit that the publisher receives by sharing the record in its original form.

$$v(g) = V \times \left(1 - \frac{IL(g)}{max\big(IL(g)\big)}\right). \qquad (2.3)$$

To compute the probability of the adversary's success $\pi(g)$, we adapt the disclosure measure described in [253] which is inversely proportional to the size of population group $n_p(g)$ that matches the record's released attribute values.

$$\pi(g) = \frac{1}{n_p(g)}. \qquad (2.4)$$

The cost of the adversary c is the price the adversary pays for accessing a record in the external resource used in the linkage attack.

To relate the game to real-world de-identification policies, we compare the best strategy of the publisher from the game to the HIPAA Safe Harbor standard. We compare the resulting strategies using the following measures: i) adversary's payoff U_a, ii) publisher's payoff U_p, iii) adversary's best strategy a, iv) generalization intensity of the publisher's best strategy, and v) the risk of the dataset (e.g., the probability of re-identification, the probability of success if the attack transpires).

We define the generalization intensity $GI(g)$ as follows:

$$GI(g) = \frac{\sum_f g_f}{\sum_f h_f - m} \qquad (2.5)$$

in which m represents the number of attributes. $GI(g) = 0$ indicates the publisher's strategy g is to release the record without generalization, and $GI(g) = 1$ implies that g completely suppresses the record.

2.3.2 A Case Study in Genomic Data Sharing

In the case study, the publishers are biomedical researchers who are disseminating research datasets. Funding organizations, such as the NIH, require researchers who are granted funding to publish the data generated by their research through websites such as the Database of Genotypes and Phenotypes (dbGaP) [254]. However, while they need to share data, they also have an incentive to protect the identities of the individuals who participated in the original research. The benefit associated with publishing the research dataset can be correlated to the amount of funding received for the project. For example, consider the dataset [255] in dbGaP submitted by the five separate member institutions of the NIH-sponsored Electronic Medical Records and Genomics Network (EMERGE) [54]. According to the NIH [256], the total sum of grant funding provided for the project is \$22,272,084. Since there are 18,663 entries, we estimate that the publisher's benefit associated with each record is $V = \$22{,}272{,}084/18{,}663 \approx \$1{,}200.$

Table 2.6 Recent notable HIPAA breach violation cases as reported by the U.S. Department of Health and Human Services.

Entity	Fine	#Records	Fine/record	Date
New York and Presbyterian Hospital	\$4,800,000	6,800	\$705.9	May 7, 2014
QCA Health Plan, Inc.	\$250,000	148	\$1689.2	Apr. 22, 2014
Skagit County, Washington	\$215,000	118,000	\$1.8	Mar. 7, 2014
Adult and Pediatric Dermatology	\$150,000	2,200	\$68.2	Dec. 26, 2013
Affinity Health Plan, Inc.	\$1,215,780	344,579	\$3.5	Aug. 14, 2013
WellPoint, Inc.	\$1,700,000	612,402	\$2.8	Jul. 11, 2013
Idaho State University	\$400,000	17,500	\$22.9	May 21, 2013
The Hospice of North Idaho	\$50,000	441	\$113.4	Jan. 2, 2013

To the best of our knowledge, there has never been a fine levied for the re-identification of data. Thus, as a proxy, we assume the loss of the publisher for each re-identified record L could be proportional to the fine paid to a federal regulator for a data breach as reported on the Office for Civil Rights' *Wall of Shame* on the U.S. Department of Health and Human Services (HHS)'s website [257]. As such, we set $L = \$300$ according to the average fines per record, which is \$325.95, recognizing that the statutory penalties could be much higher. Table 2.6 provides a summary of the fines and numbers of records involved in recent HIPAA violation breach cases. The sensitivity analysis in the following section provides intuition into how other types of loss, such as that which is incurred through identity theft blackmail, influences the results of the game.

We set the cost of the adversary c for each external identified record to \$4 (based on the \$3.95 price for a basic report from Intelius[9]).

Finally, we set the number of generalization levels h for each of the attributes Age, Race, Gender, and Zip to be 6, 4, 2, and 6, respectively, to ensure the publisher has a relatively large strategy space (a space of 288 strategies), considering the number of distinct values for these attributes are 121, 5, 2, and 628, respectively. We construct the generalization hierarchies such that i) HIPAA Safe Harbor de-identification policy is always a strategy in the publisher's strategy space; ii) for Age, Race, and Gender, each level has a relatively certain number of distinct values (e.g., the number of distinct values in 6 levels of the attribute Age from top to bottom is 121, 60 or 61, 30 or 31, 15 or 16, 5 or 6, and 1, respectively); iii) for Zip, each level has a distinct form of representation (e.g., for Zip 37203, the 6 levels from top to bottom are represented as *****, 3****, 37***, 372**, 3720*, and 37203, respectively) except Zip 42223, 42602, and 72338 which has fewer than 20,000 residents (e.g., for Zip 42223, the 6 levels from top to bottom is represented as *****, [4****, 7****], [42***, 72***], [422**, 426**, 723**], 4222*, and 42223, respectively).

For orientation, Figure 2.5 demonstrates the publisher's and adversary's payoffs across all strategy profiles for the record <48, Asian, Female, 38363> in the *Basic Game*. In this case, the publisher's best strategy is to release the record as <48, *, *, 38363>. The resulting generalization intensity is 0.29 and the adversary's probability of success is 0.0104. The adversary's best response is *no attack*, such that his payoff is \$0 while the publisher's payoff is \$1,049.70. Now, if the publisher had invoked the HIPAA Safe Harbor policy, the record would have been released as <48, Asian, Female, 383**>, which yields a generalization intensity of 0.14 (less intensive) and the adversary's probability of success would have been 0.1429, which is approximately 14 times of the rate of the *Basic Game*. Moreover, unlike the *Basic Game*, the adversary's best response is to attack this record. As a consequence, the adversary's expected payoff increases to \$38.86, and the publisher's payoff decreases to \$776.30. For this record, the best strategy from the *Basic Game* outperforms the Safe Harbor policy.

[9] Intelius. https://www.intelius.com/.

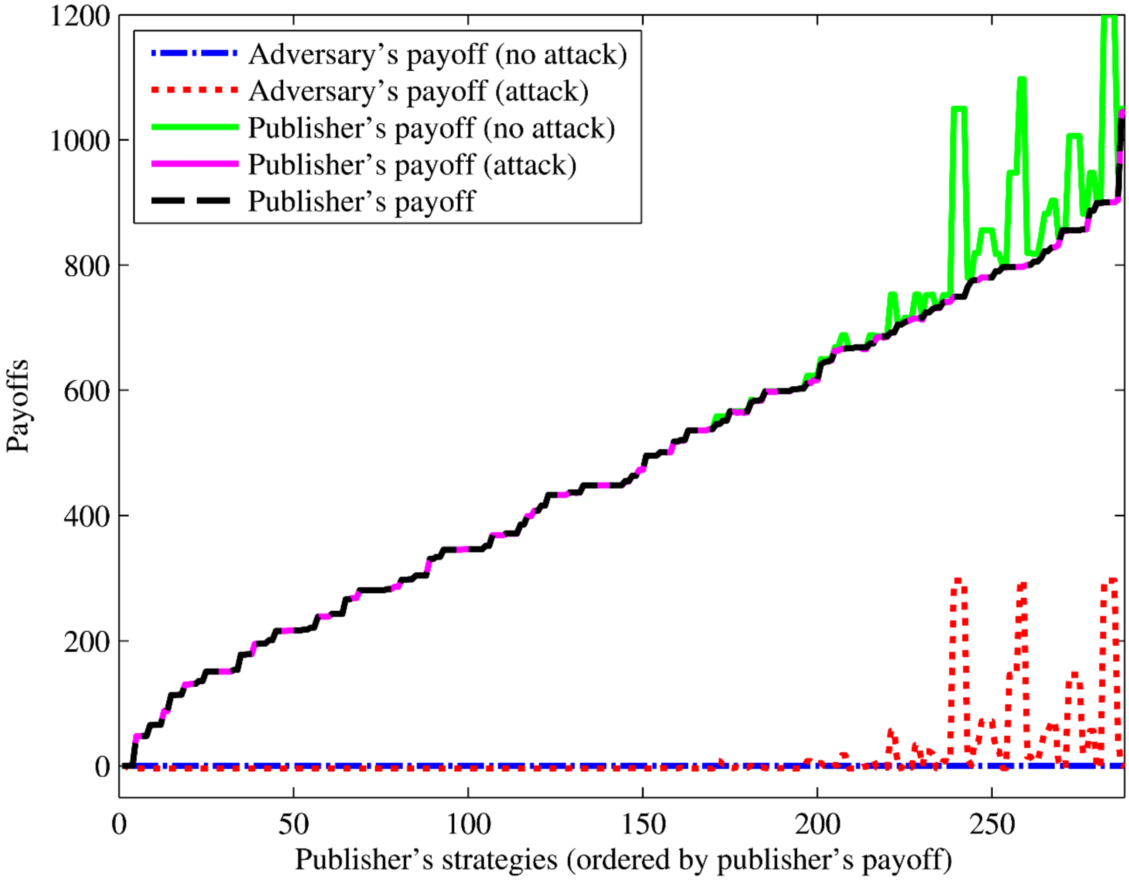


Figure 2.5 Payoffs for the record <48, Asian, Female, 38363> across all strategies.

Table 2.7 A comparison of four de-identification policies for the case study on performance measures.

Performance Measures	**SH**	**Basic**	**SH-Friendly**	**No-Attack**
Publisher's average payoff over all records	$852.06	$1,195.50	$852.51	$1,169.80
Adversary's average payoff over all records	$0.43	$2.43	$0.05	0
Proportion of records with GI = 1 (suppression)	0	0	0	0
Proportion of records with GI = 0 (most specific)	0	62.98%	0	62.73%
Average GI over all records	0.1431	0.0280	0.1453	0.0412
Proportion of records the adversary will attack	2.68%	21.87%	0.91%	0
Average probability of re-id over all records	0.0110	0.0018	0.0003	0
Average prob. of re-id over all attacked records	0.0504	0.0668	0.0310	0

SH: Safe Harbor. GI: Generalization Intensity.

Turning our attention to the entire dataset, Table 2.7 summarizes the results of all variations of the re-identification game and Safe Harbor on several measures. In the *Basic Game*, it can be seen that the average payoff to the publisher and the adversary for each record is $1,195.50 and $2.43, respectively. For every record, the publisher's best strategy is to share data. In 62.98% of the time, records are shared in their most specific form. At the same time, the adversary's best strategy is to attack 21.87%, or 7122, of the records. However, the average probability of a successful re-identification for any record is 0.0110. And, for each of the records attacked, the average probability of success is 0.0504, such that the expected number of re-identified records would be ~359.

Under the Safe Harbor policy, the average payoff to the publisher and adversary for each record is $869.84 and $0.43, respectively. In this scenario, the adversary's best strategy is to attack 2.68%, or 873, of the records. The average probability of a successful re-identification for any record is 0.0018. For each of the attacked records, the average probability of success is 0.0668, such that the expected number of re-identified records is ~58.

In the *No-Attack Game*, it can be seen that the average payoff for the publisher is $1,169.80 (while, of course, the adversary receives a payoff of $0). In this situation, the publisher's best strategy is to share all the records, while 62.73% of the records are shared without any generalization.

In the *SH-Friendly Game*, the average payoff to the publisher and adversary for each record is $852.51 and $0.05, respectively. The publisher's best strategy is to share all records with generalization, while the adversary's best strategy is to attack 0.91%, or 295, of the records. The average probability of a successful re-identification for any record is 0.0003. For each of the attacked records, the average probability of success is 0.0310, such that the expected number of re-identified records is ~9. The best strategies from the *No-Attack Game* and *SH-Friendly Game* both unambiguously outperform the Safe Harbor policy by achieving higher payoffs for the publisher as well as lower payoffs for the adversary.

To investigate these results in greater depth, Figure 2.6 depicts the distribution of payoff gains for the *Basic*, *SH-Friendly*, and *No-Attack* games, relative to the HIPAA Safe Harbor policy (in other words, how much better each of these performs than Safe Harbor from the publisher's and adversary's perspectives). In Figure 2.6, the plot to the left displays the frequency distribution of payoff differences for the publisher. It can be observed that the publisher almost always receives a higher payoff using our framework. Among the three games, the *Basic Game* is the best, and the *SH-Friendly Game* is the worst. The plot to the right displays the frequency distribution of payoffs for the adversary. Here, it can be observed that, at times, the adversary receives a higher payoff only via the *Basic Game*. This is because the publisher optimizes the payoff regardless of the risk and, in certain instances, the risk increases in tandem with the payoff.

Figure 2.7 depicts the detailed distributions of the publisher's and the adversary's payoffs between the games and the HIPAA Safe Harbor de-identification policy for all individuals. First, we consider the viewpoint of the publisher. In the plots on the left, it can be observed that only the *No-Attack Game* has a smaller payoff than Safe Harbor for the publisher. Moreover, among the variations on the de-identification game, the *Basic Game* has the largest publisher's payoff. Next, we turn our attention to the adversary. In the plots to the right, it can be observed that only the *Basic Game* has the capability of achieving a larger payoff than Safe Harbor for the adversary. Among the three games, the *SH-Friendly Game* exhibits the smallest payoff for the adversary.

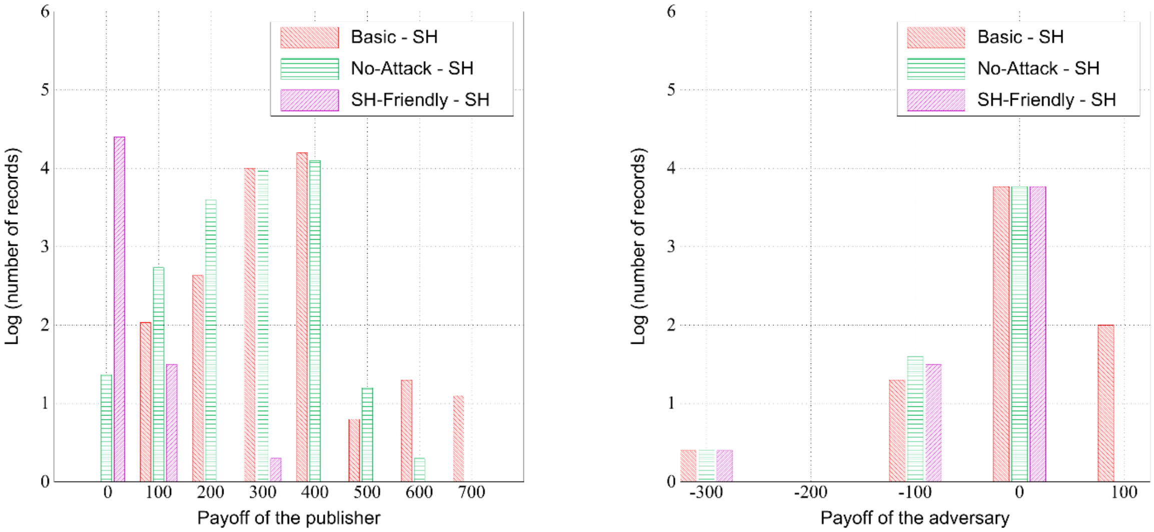


Figure 2.6 Histogram of Payoff Differences.

Distributions of the publisher's payoff differences (left) and the adversary's payoff differences (right) between games and HIPAA Safe Harbor (SH).

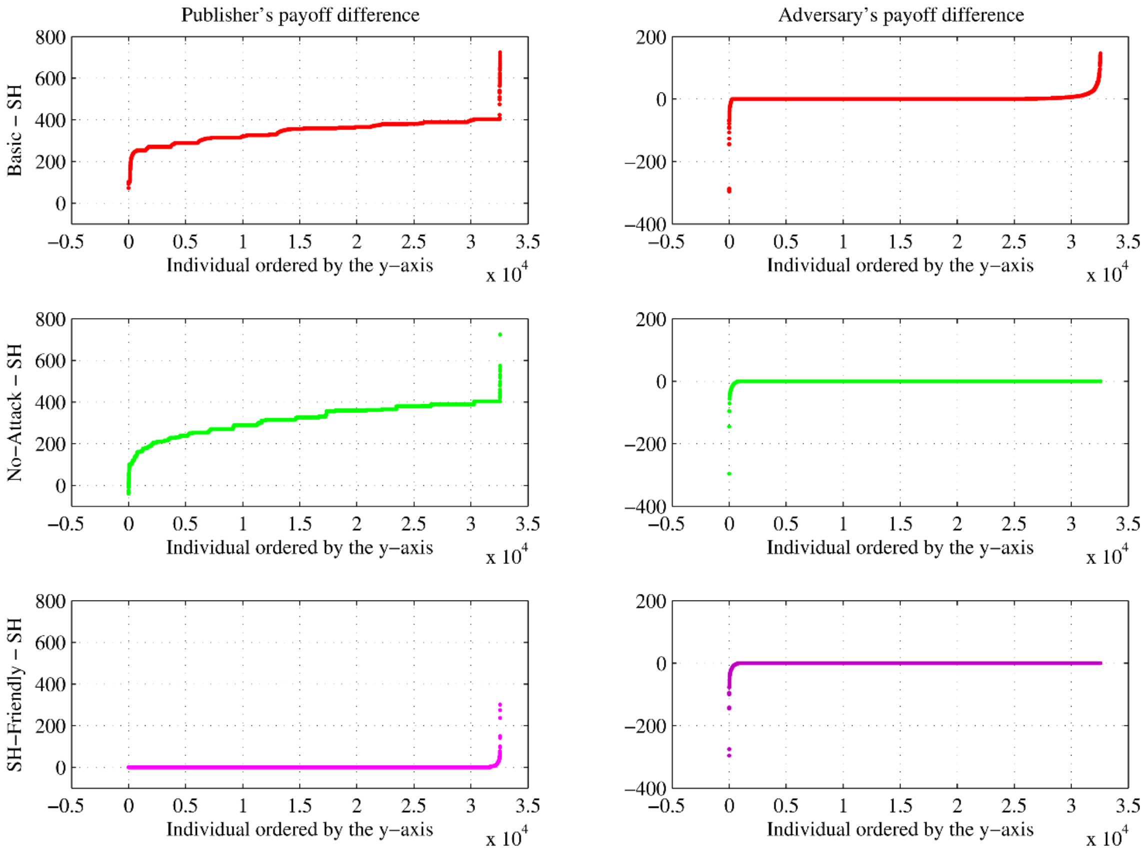


Figure 2.7 Scatter-plot of Payoff Differences.

Detailed distributions of the publisher's payoff differences (left) and the adversary's payoff differences (right) between games and HIPAA Safe Harbor (SH).

2.3.3 Sensitivity Analysis of the Game

In the re-identification game, the Stackelberg equilibrium will remain unchanged if the parameters L, V, c are multiplied by a constant factor. As such, we vary two parameters while holding the third constant.

Figure 2.8 and Figure 2.9 show the best strategy for the publisher (in the form of GI) and the adversary for two records with vastly different adopted strategies. Specifically, Figs. 2.8 and 2.9 correspond to <19, Black or African American, Male, 37208> and <62, White, Female, 37014>, respectively.

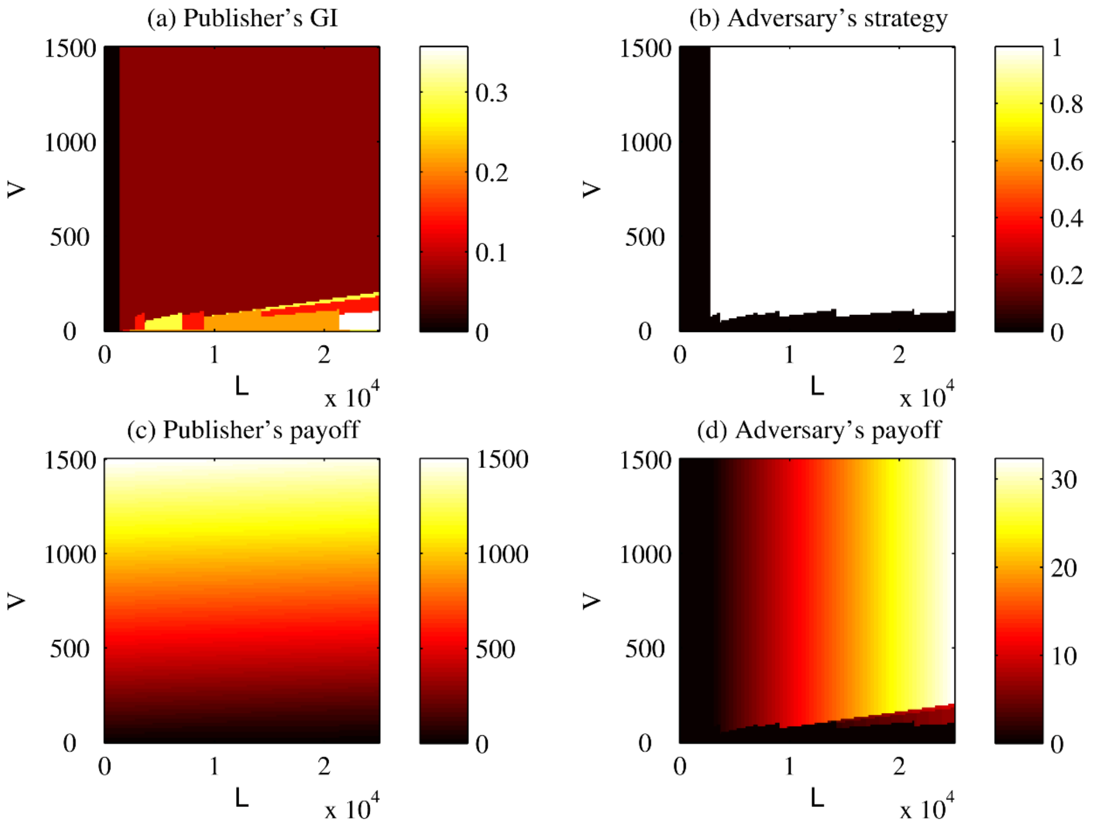


Figure 2.8 Example solution 1 in Basic Game.

Resulting strategies and payoffs for <19, Black or African American, Male, 37208> in the Basic Game. *GI* stands for the generalization intensity. *V* corresponds to the benefit that the publisher receives by sharing the record in its original form. *L* corresponds to the publisher's loss for one record due to a successful re-identification.

For orientation, in Figure 2.8(a), a GI of 1 (white at the top of the color bar) implies that the publisher will not release anything; a GI of 0 (black at the bottom of the color bar) implies that the publisher will release everything; and a GI between 0 and 1 implies that the publisher will generalize part of the record. From Figure 2.8(a), it can be observed that the whole space is separated into several parts that within each of them the publisher's GI keeps the same. Generally, the color of the part representing high V and low L is darker than that of the part representing low V and high L. However, it is not necessary true that the publisher uses a larger amount of generalization for the record when L is high and V is low. In Figure 2.8(b), the white and black areas indicate that the adversary chooses to attack and not attack, respectively. From Figure 2.8(b), it can be observed that a frontier cuts the whole space into two parts represent different strategies for the adversary. The adversary will not attack when L or V is small. In Figure 2.8(c), the brighter the region, the larger the publisher's payoff. In Figure 2.8(c), clearly, the publisher's payoff increases as V increases. In Figure 2.8(d), the brighter the region, the larger the adversary's payoff. From Figure 2.8(d), it can be observed that the whole space is separated into several parts that, within each of them, the adversary's payoff increases only as L increases. The pattern of separation in Figure 2.8(d) is the same as that in Figure 2.8(a).

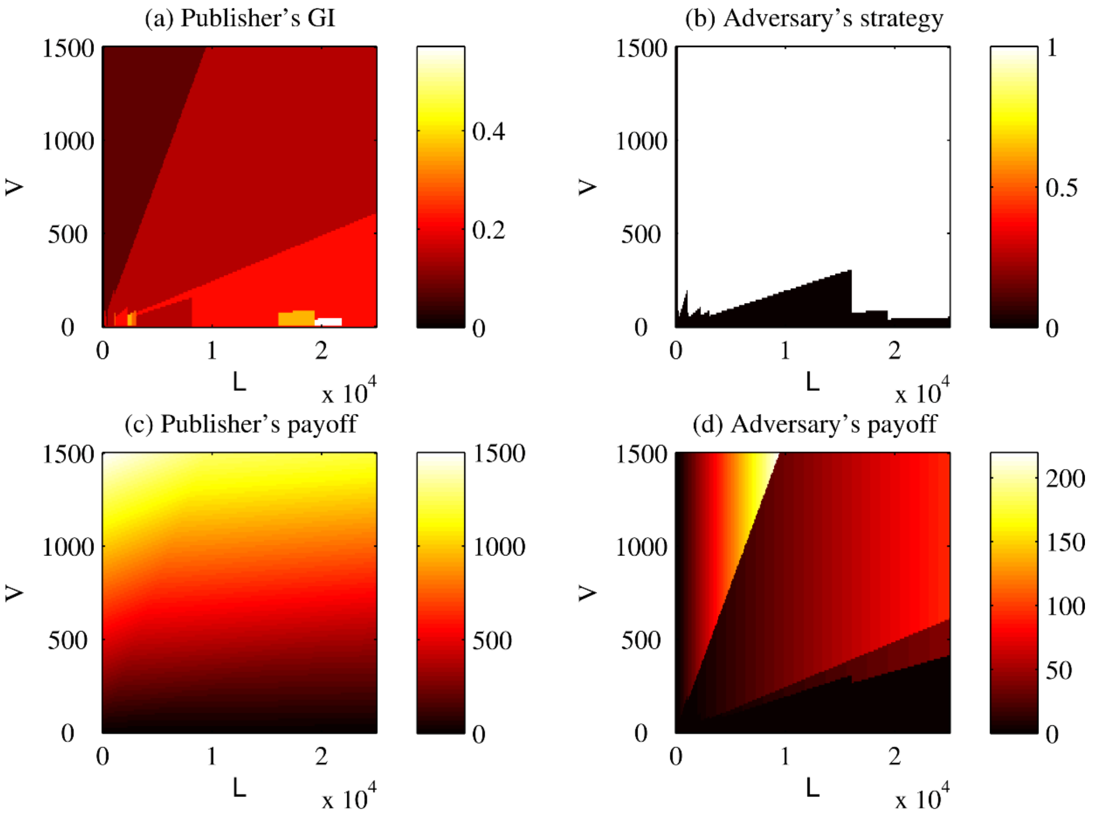


Figure 2.9 Example solution 2 in Basic Game.

Resulting strategies and payoffs for <62, White, Female, 37014> in the Basic Game. *GI* stands for the generalization intensity. *V* corresponds to the benefit that the publisher receives by sharing the record in its original form. *L* corresponds to the publisher's loss for one record due to a successful re-identification.

Similar results can be observed in Figure 2.9. The differences between Figure 2.8 and Figure 2.9 include apparently different patterns of separation for the publisher's GI and different frontiers for the adversary's strategy.

For these two records, we find that both the publisher's strategy and the adversary's strategy are not sensitive to the changes of V and L near the point representing the case study ($V = \$1{,}200$, $L = \$300$, $c = \$4$).

Figure 2.10 depicts the publisher's GI, adversary's strategy, and both payoffs for the entire dataset while varying L and V in the *Basic Game*. In Figure 2.10(a), the brighter the region, the larger the publisher's average GI. It can be observed that the publisher, on average, generalizes less intensively (i.e., shares more data) than the Safe Harbor policy (0.1431) when either $L \leq \$4{,}200$, or $V \geq \$100$; thus, this general finding holds over an order-of-magnitude range of both of these parameters. In Figure 2.10(b), the brighter the region, the larger the chance the adversary will attack. We can see that the adversary will attack fewer than 30% of the records when either $L \leq \$1{,}000$ (giving us an order-of-magnitude range of robustness for L, for any value of V up to \$1,500) or $V \geq \$100$ (independent of L up to \$25,000). In Figure 2.10(c), the brighter the region, the larger the publisher's average payoff. As expected, the publisher's payoff is correlated with the publisher's benefit V, such that when the benefit is large, the payoff is large as well. In Figure 2.10(d), the brighter the region, the larger the adversary's average payoff. The adversary's payoff is large for instances where the publisher's benefit and loss are large.

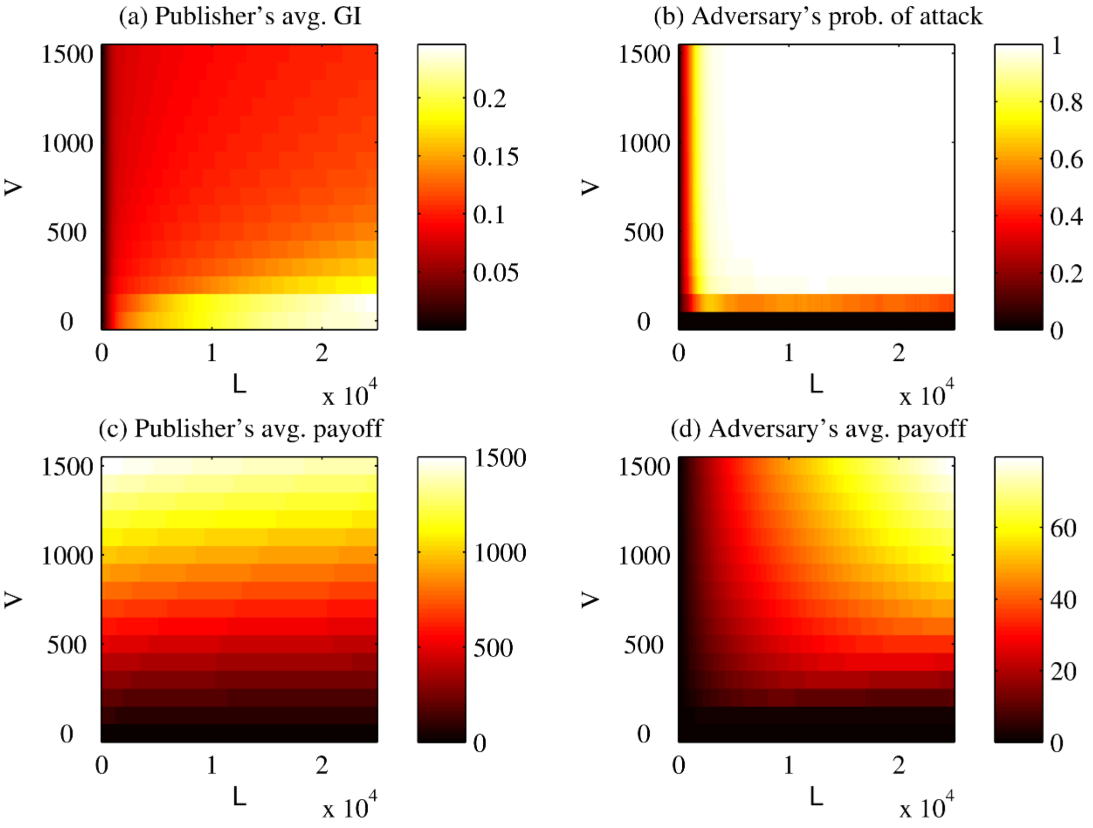


Figure 2.10 The sensitivity of Basic Game to Benefit and Loss.

The sensitivity of the average payoffs and strategies over the dataset to the changes of the publisher's benefit and the publisher's loss in the Basic Game. *GI* stands for the generalization intensity. *V* corresponds to the benefit that the publisher receives by sharing the record in its original form. *L* corresponds to the publisher's loss for one record due to a successful re-identification.

In Figure 2.11, we compare the publisher's and adversary's payoffs for the Safe Harbor policy with the *Basic*, *No-Attack*, and *SH-Friendly* games by changing L, V, or c only. It can be observed from Figure 2.11(a), Figure 2.11(c), and Figure 2.11(e) that all three games lead to better payoffs for the publisher comparing to Safe Harbor. In addition, Figure 2.11(b), Figure 2.11(d), and Figure 2.11(f) illustrate that the *No-Attack* and *SH-Friendly* games protect the data better than Safe Harbor does. Moreover, when V is smaller than \$200, or c is larger than \$90, the *Basic Game* outperforms Safe Harbor as well. From Figure 2.11, it can be observed that the increase of L benefits the adversary only; the increase of c benefits the publisher only; and the increase of V benefits both, though more for the publisher.

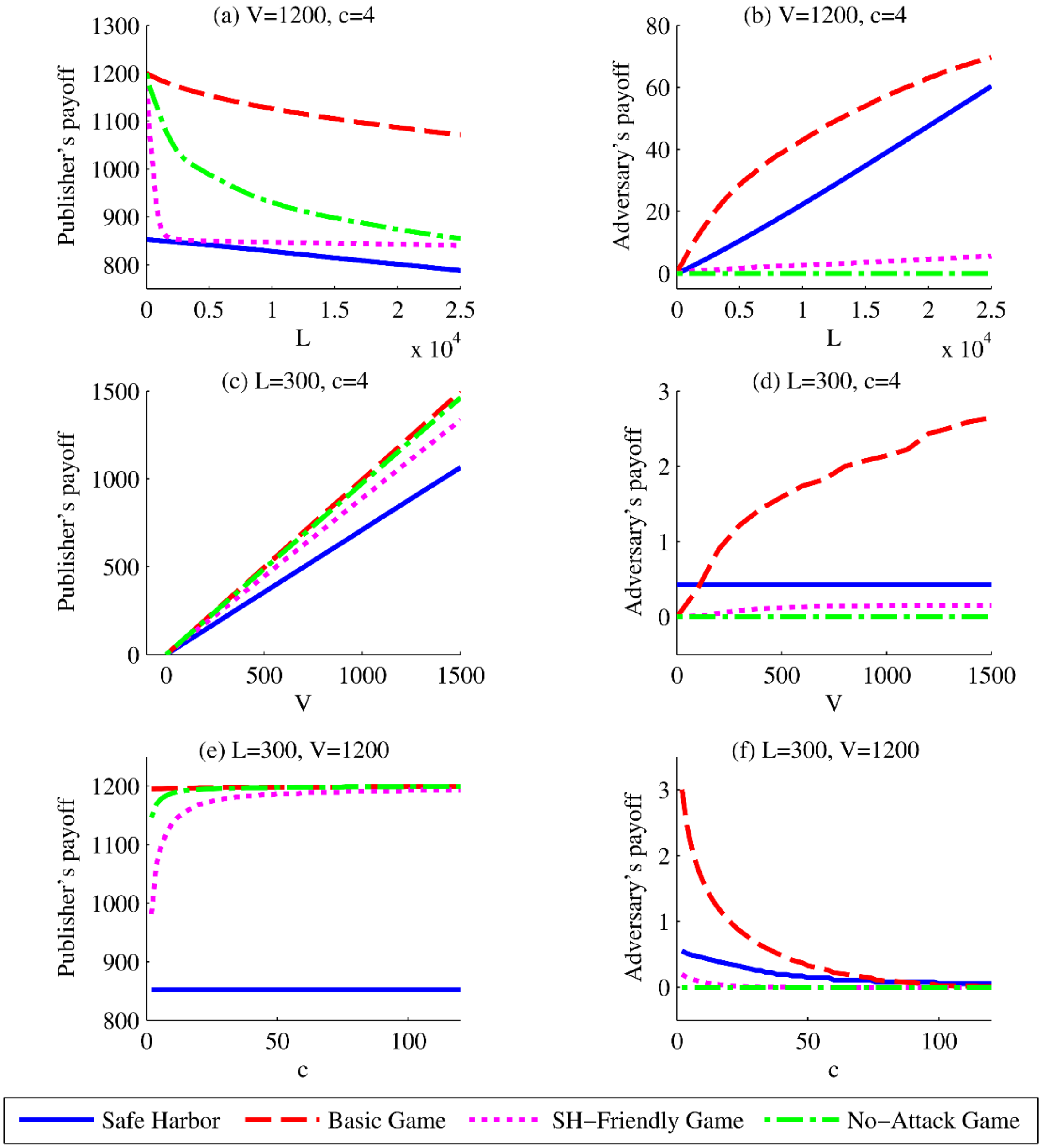


Figure 2.11 Sensitivities of different scenarios to Benefit or Loss.

The sensitivity of the average payoffs over the dataset to the change of the publisher's benefit, the publisher's loss, or the adversary's cost in different de-identification scenarios. *V* corresponds to the benefit that the publisher receives by sharing the record in its original form. *L* corresponds to the publisher's loss for one record due to a successful re-identification. *c* corresponds to the adversary's cost to launch a re-identification attack towards one record.

2.3.4 Performance Comparison of Game Solvers

In this subsection, we compare the performance of the BIS and LBS algorithms to determine if 1) the latter is more efficient, and 2) the results of LBS are sufficiently close to the baseline BIS. To do so, we rely on two categories of evaluation metrics. The first category corresponds to quality measures, which include i) the publisher's average payoff, ii) the adversary's average payoff, iii) the average absolute difference of the publisher's payoff D_p between the two algorithms, and iv) the average absolute difference of the adversary's payoff D_a between the two algorithms. The average absolute difference of the publisher's payoff is calculated as follows. Let the publisher's payoff using BIS and LBS for record i be $U_{BIS}(i)$ and $U_{LBS}(i)$, respectively, and n be the total number of the records. Then, the average absolute difference of the publisher's payoff is calculated as:

$$D_p = \frac{\sum_{i=1}^{n} |U_{BIS}(i) - U_{LBS}(i)|}{n}. \tag{2.6}$$

For instance, if the publisher's payoffs are \$700, \$800, and \$900 in the results of the baseline BIS, and \$700, \$805, \$899 in the results of a heuristic-driven LBS, then the publisher's average payoff will be \$801 and the average absolute difference of the publisher's payoff is \$2. The average absolute difference of the adversary's payoff is calculated similarly.

The second category of measures correspond to the efficiency of the strategy search process: v) the average running time and vi) the average number of strategies (or nodes in the lattice) searched. We retain the same parameter settings introduced in the previous section.

Table 2.8 provides a summary of the comparison results for BIS and LBS. To measure the runtime (in milliseconds, or ms), we use an Intel CORE i5 2.67GHz machine with 4GB RAM. We apply a two-sample t-test at the 5% significance level to show that the running times of BIS and LBS are significantly different. It is clear that the LBS approach runs faster than the baseline approach (BIS). The LBS approach ends up with almost the same average payoff for the publisher and the even less average payoff for the adversary. Additionally, 99.47% of the solutions returned from LBS was the optimal solutions.

Table 2.8 A performance comparison of the de-identification game solving approaches.

Metric	**BIS**	**LBS**
Running time (ms)	268.1	15.8
Searched nodes	288	3.9
Publisher's Average Payoff	\$1195.5	\$1195.2
Adversary's Average Payoff	\$2.43	\$0.12
Publisher's Average Payoff Difference	0	\$0.29
Adversary's Average Payoff Difference	0	\$2.64

BIS: Backward Induction Search. LBS: Lattice-Based Search. Payoff difference means the absolute difference of payoff for one record between a heuristic-driven approach and the baseline BIS approach.

Next, to illustrate the sensitivity of quality metrics with respect to L and V, we plot the average absolute deviation of the publisher's payoff D_p from the baseline solution in Figure 2.12. From Figure 2.12(a), we observe that, when $V = \$1{,}200$, $c = \$4$, as L increases, the LBS solution deviates from the BIS baseline, but with a weakening trend. From Figure 2.12(b), when $L = \$300$, $c = \$4$, the LBS solution deviates most when V is around \$300 and then converges to BIS as V increases.

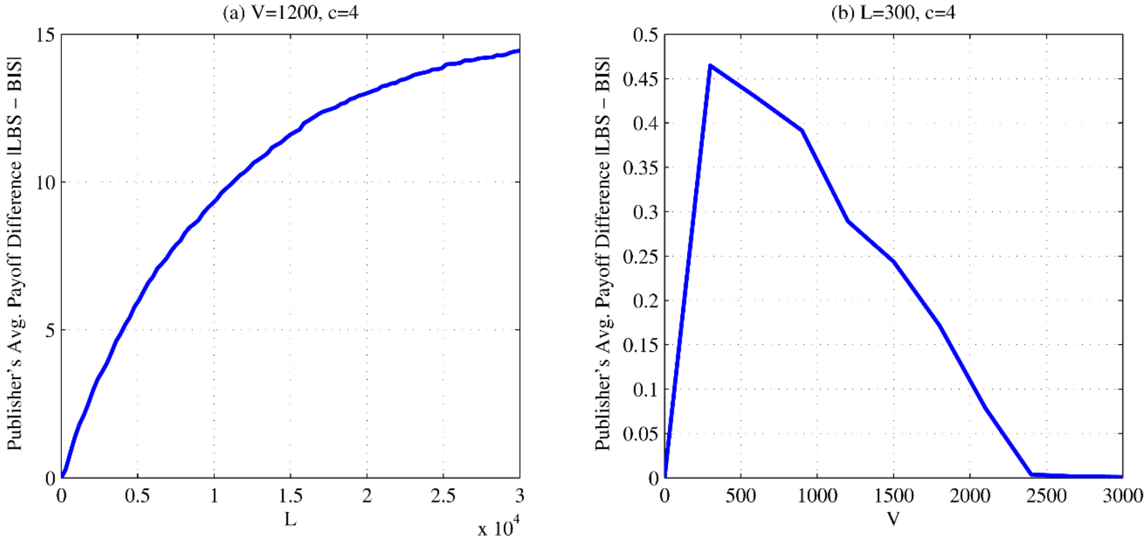


Figure 2.12 The accuracy of Lattice-based Search.

A comparison of the accuracy of the LBS game solving heuristic. V corresponds to the benefit that the publisher receives by sharing the record in its original form. L corresponds to the publisher's loss for one record due to a successful re-identification. c corresponds to the adversary's cost to launch a re-identification attack towards one record.

2.4 Discussion

The results of this study are notable for several reasons. First, they illustrate a new formal framework for analyzing data de-identification risk, which builds on game theory in order to capture the motivations of a data recipient for re-identifying such data. This framework enables us to identify and rigorously support a de-identification policy, which optimally trades off the value of shared data and the associated re-identification risk. Second, our results demonstrate that it is actually possible to achieve zero risk, in the sense that a data recipient would have no incentive to attempt re-identification (with limitations discussed below). Remarkably, this "zero-risk" policy shares nearly as much data as an optimal policy and far more data than the popular HIPAA Safe Harbor de-identification policy. Third, even if we impose

Safe Harbor guidelines as a strict constraint on the kinds of de-identification policies we can consider, our results demonstrate that a number of individuals are served poorly by Safe Harbor, in the sense that it releases *too much* information about them. Our model significantly improves outcomes by clamping down on the information release for these individuals.

The legal and institutional complexity of data sharing risk, as well as uncertainty of enforcement guidelines, make rigorous risk analysis for data sharing particularly challenging. Our general framework, as well as a case study illustration, suggest that a formal approach based on a game theoretic model is defensible in the light of regulatory oversight (it is motivated by, and compliant with, the HIPAA notion of an anticipated data recipient) and sufficiently flexible to offer data publishers real alternatives to balance their data sharing needs with institutional and legal requirements, as well as their own risk preferences.

2.4.1 Limitations

This study has several limitations that can serve as guideposts for future research. The first limitation is that our case study assumed a single source of side data. Thus, the results could potentially change if other related side data can be utilized. However, we emphasize that this is not a limitation of our model, which captures arbitrary side data (the cost and probability of re-identification would then correspond to the most efficacious re-identification policy). Second, our study imposes a fixed generalization hierarchy for all records, thereby significantly limiting the granularity of the publisher's decision space. Our framework does not readily admit to adding noise to published data release, and this is an important problem to consider in future work. The third limitation is that our model assumes a single adversary (data recipient). In future work, it would be desirable to generalize the model to capture the scenario of multiple data recipients, the uncertainty about the payoffs and information of data recipients, as well as the possible constraint that the same representation of data is shared with all recipients, which would give rise to a multi-follower Bayesian Stackelberg game like the one in [179] and [209]. When multiple types of adversaries are involved, cooperation and competition among them may make the game similar to the prisoner's dilemma game or the public goods game, for which there has been recent research in solving the problem in a scalable fashion [258, 259]. The fourth limitation is that the model makes no provision for other sources of data re-identification risk, such as another party breaking into a data recipient's systems and stealing the data. This third party may well have strong non-economic motivations to re-identify data. The clear implication is that our measure of risk (including settings with "zero" risk) underestimates the true risk of re-identification. Capturing this background risk of data sharing is also an important subject for future work.

2.4.2 Conclusions

De-identification has been a hallmark of data protection for years, but there is a fear that this technique is insufficient given mounting re-identification evidence. However, before we throw out de-identification in favor of other technical and legal mechanisms, it should be recognized that it is fundamentally a problem of risk management and that data will be compromised only if adversaries are sufficiently incentivized to do so. To enable pragmatic discussion about de-identification and potential concerns, we introduced a novel formalization of the re-identification problem as a Stackelberg game between a data publisher and recipient. We translated the risk and the utility of the released dataset into the data publisher's benefit and cost and proposed several methods for computing the optimal data sharing policy for the publisher. We illustrated our model using a real case study, showing that we can typically achieve much better outcomes for the publisher than HIPAA Safe Harbor, a popular real-world de-identification policy, often at lower re-identification risk. Indeed, our results indicate that publishers can choose strategies that allow for sharing a significant amount of data (far more than under Safe Harbor) while ensuring that it is never beneficial for the data recipient to attempt re-identification, thus ensuring zero risk within the context of our modeling framework.

Chapter 3

MEMBERSHIP INFERENCE GAME

Emerging scientific endeavors are creating big data repositories from millions of individuals. Sharing data in a privacy-respecting manner could lead to important discoveries, but high-profile demonstrations show that links between de-identified genomic data and named persons can sometimes be reestablished. Such re-identification attacks have focused on worst-case scenarios and spurred the adoption of data sharing practices that unnecessarily impede research. To mitigate concerns, organizations have traditionally relied upon legal deterrents, like data use agreements, and are considering suppressing or adding noise to genomic variants. In this chapter, we use a game theoretic lens to develop more effective, quantifiable protections for genomic data sharing. This is a fundamentally different approach because it accounts for adversarial behavior and capabilities and tailors protections to anticipated recipients with reasonable resources, not adversaries with unlimited means. We demonstrate this approach with a new public resource with genomic summary data from over 8000 individuals - the Sequence and Phenotype Integration Exchange (SPHINX) – and show risks can be balanced against utility more effectively than traditional approaches. We further show the generalizability of this framework by applying it to other genomic data collection and sharing endeavors. Recognizing that such models are dependent on a variety of parameters, we perform extensive sensitivity analyses to show that our findings are robust to their fluctuations.

3.1 Introduction

Emerging large-scale scientific endeavors [260], such as the Precision Medicine Initiative (PMI) [231], will collect a wide variety of data, including genomic and phenomic records on millions of participants, enabling unprecedented opportunities for discovery and eventual treatment [9, 261, 3, 262]. Broad sharing and reuse of data from such programs are endorsed by many people, but participants often expect that their privacy, and particularly their anonymity, will be preserved [263]. To meet these expectations, organizations have adopted various legal protections, such as data use agreements (DUAs) that explicitly prohibit re-identification [264] and technical controls, such as the suppression [36] or adding noise [265, 127] to genomic variants that have a high likelihood of distinguishing an individual.

However, some worry that the preservation of privacy in such settings may be impossible to realize, even when sharing only summary statistics [76, 266]. This concern is driven, in part, by high-profile demonstrations over the past decade showing how de-identified genomic data can be tracked back to named persons [7, 86], leading to public apologies and dramatic policy changes [66]. Unfortunately, though re-identification attacks on genomic data have been perpetrated, they focused on worst-case scenarios [172] and, in doing so, have spurred the adoption of data sharing practices that may unnecessarily impede research [267].

Genomic data sharing policies should tailor protections to anticipated recipients with reasonable resources, not adversaries with unlimited means. The precedent for this perspective exists in the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 [5], the National Institutes of Health (NIH) Genomic Data Sharing Policy [268], and the General Data Protection Regulation of the European Union [65]. In this chapter, we posit that a grounded model for genomic data dissemination can be achieved through an approach based on game theory that accounts for adversarial behavior and capabilities. Our approach is fundamentally different to reasoning about genomic data privacy - though such techniques have already been used to analyze the re-identification risk [236] and proven successful in other risk inherent domains, such as airport security and coast guard patrols [269, 270].

To illustrate the potential of this strategy for the design of genomic data sharing policies, this chapter examines known attacks on genome-phenome summary statistics and shows that levels of risk that have been claimed may be substantially larger than is necessary and risks can be balanced against utility more effectively using a game-based approach. We demonstrate how such models can be applied to determine the optimal set of protections for emerging genomic data sharing initiatives, leveraging a new public resource, the Sequence and Phenotype Integration Exchange (SPHINX) system[10] for a case study, which reports single nucleotide polymorphism (SNP) summary statistics (i.e., minor allele frequencies) on data collected from the NIH-sponsored Electronic Medical Records and Genomics Pharmacogenomics (eMERGE-PGx) network [229].

3.2 Methods

3.2.1 Model

[10] Sequence and Phenotype Integration Exchange. https://www.emergesphinx.org/

To gain intuition into these findings, we begin by pointing out that a grounded model of the genomic data privacy problem requires an explicit definition of the actors involved and their interactions, as illustrated in Figure 3.1. In the basic case, there are two actors: (i) the *sharer*, who could be an investigator of a study or an organization, such as an academic medical center that manages genomic data on their behalf, and (ii) the *recipient*, who requests and accesses the data for some purpose (e.g., replication of published findings or discovery of new associations). The overwhelming majority of recipients are unlikely to misuse the data, but the privacy concern is on those with the potential to exploit named genomes (or targets) by determining their presence in the research study. A core motivation for both sharing and attacking genomic records is the belief that the data has intrinsic value. Specifically, the sharer benefits by gaining utility from disseminating data, while the recipient benefits by detecting (and exploiting) the targets. It is important to recognize that attacks entail costs, such as obtaining identified data needed for linkage, as well as the human capital or computational power necessary to run the attack [31]. Additional costs to a recipient may further accrue from the consequences of breaching a DUA, should such an agreement be in place, and the recipient is found to be in breach.

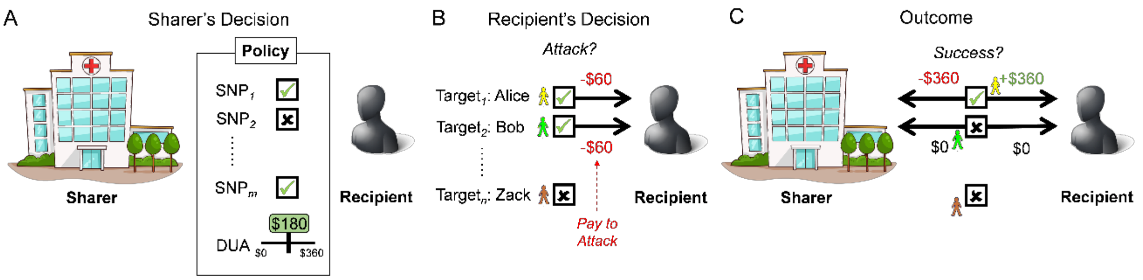


Figure 3.1 The genomic data sharing process.

In this process, (**A**) a genomic data sharing policy is made by the sharer, (**B**) a recipient chooses to attack targets in received data, and (**C**) the overall payoffs as a consequence. SNP stands for single nucleotide polymorphism. DUA stands for data use agreement.

The NIH Genomic Data Sharing Policy [268] states that a potential NIH grant awardee needs to define a data-sharing plan. Thus, if the investigators believe that sharing data is too risky, they may choose not to apply for a grant or may state that only a limited amount of data will be shared in a restricted fashion (which will be taken into account before a proposal for funding is adjudicated). The sharer's decision about a combination of instituting a DUA and technical protection measures (e.g., suppressing information on certain SNPs), as well as the recipient's choice of whether an attack is worth its cost, constitute a Stackelberg game [177], a natural model of this interaction. In this model, the sharer is a leader who may 1) require a DUA with liquidated damages in the event of a breach of contract and 2) share a subset of SNP summary statistics from a specific study (suppressing the rest). The recipient of the data, whom we assume to be self-interested, then follows by determining whether or not the benefits gained by attacking each target outweigh the costs. Crucially, the sharer chooses the policy that optimally balances the anticipated utility and privacy risk.

To be precise, we define g as a set of genomic variants (or SNPs) to be shared, a as a set of individuals to be attacked, $B_S(g)$ as the benefit and $\hat{C}_S(a)$ as the estimated cost to the sharer. The sharer's goal is to maximize the following payoff function by selecting the best strategy g^*:

$$g^* = \underset{g}{\operatorname{argmax}} \overbrace{\left(B_S(g) - \hat{C}_S\left(a^*(g)\right)\right)}^{Sharer's\ Payoff} \qquad \text{(\textit{Sharer's Optimal Strategy})} \ (3.1)$$

in which

$$a^*(g) = \underset{a}{\operatorname{argmax}} \overbrace{\left(\hat{B}_R(g,a) - \hat{C}_R(a)\right)}^{Recipient's\ Payoff}. \qquad \text{(\textit{Recipient's Optimal Strategy})} \ (3.2)$$

In this model, the recipient aims to maximize his or her own payoff (achieved through exploiting the data) and thus determines the subset of targets for whom they believe can successfully be attacked. We use $\hat{C}_R(a)$ to define the estimated cost to the recipient for attacking targets, which includes the expected prefixed liquidated damage penalty for breach of DUA. We denote the estimated benefit to the recipient as $\hat{B}_R(g,a)$, which is the benefit the recipient expects to gain from attacking targets.

We defer the complete details of the benefit and cost representations below in Subsection 3.2.2 and highlight the connection between the sharer's cost and the recipient's benefit. Specifically, for any successfully attacked target, the recipient gains a fixed amount G_R, while the sharer loses a fixed amount L_S. As a result, both the sharer's cost and the recipient's benefit are proportional to the number of successfully attacked targets.

To simulate the recipient's uncertainty, we adopt the framework introduced by Sankararaman *et al.* [36], which compares genomic variant rates between the shared data and a public reference such as data from 1000 Genomes [230] (which assumes the shared data is drawn from a reference population), so that the estimated number of successfully attacked targets is computed as:

$$\widehat{N}_R(g,a) = \sum\nolimits_{i \in I} \pi(i,g)a[i]\tau[i] \ = \sum\nolimits_{i \in I} p[i]l(i,g)a[i]\tau[i] \qquad (3.3)$$

where $\pi(i,g)$ is the posterior probability that target i is in the study, I is the set of individual available for attack, $p[i]$ is the prior probability that target i is in the study, $l(i,g)$ is the likelihood ratio that compares the likelihood that target i is in the study versus that it is in a reference population, $\tau[i]$ is the probability that individual i is targeted, and $a[i]$ is a binary variable which is 1 if target i is attacked and 0 otherwise

(See **Appendix A.1** for the detailed derivation). One of the notable aspects of this model is that it incorporates the prior probability. It has been recognized that this prior probability of inclusion in a study has a significant impact on the likelihood of a successful attack [170], yet it is often neglected in re-identification risk assessments, including the popular framework upon which we build [36]. We illustrate the importance of this factor in our experimental analysis below in Section 3.3. The game is computationally solved using an optimization algorithm, as outlined below in Subsection 3.2.3.

3.2.2 Payoff functions

Table 3.1 The notation used throughout this chapter, ordered by their chronological introduction.

Notation	Description
J	The set of single nucleotide polymorphisms (SNPs) available for sharing
g	A possible combination of SNPs for sharing, represented by a binary vector of size m
I	The set of individuals in the study dataset available for attack
$\hat{Y}_S$	The sharer's payoff, as estimated by the sharer
a	A possible combination of individuals in the study, represented by a binary vector of size n
B_S	The sharer's benefit
$\hat{C}_S$	The sharer's cost estimated by the sharer
H	The worth of all the data available to the sharer
U	The utility score of shared data
L_S	The loss to the sharer per successfully attacked targets
$\widehat{N}_S$	The number of successfully attacked targets, as estimated by the sharer
D	The study dataset for sharing (i.e., the pool dataset)
$d[i,j]$	The number of minor alleles in SNP j of individual i in the study data (SNP j = j^{th} SNP, individual i = i^{th} individual)
$w[j]$	The utility weight of the SNP j
$f[j]$	The minor allele frequency of SNP j in the study dataset
$\hat{f}[j]$	The minor allele frequency of SNP j in the public reference dataset
$\tau[i]$	The probability that individual i in the study dataset is targeted, or the targeting rate
$\hat{Y}_R$	The recipient's payoff, as estimated by the recipient, as estimated by the sharer (regarding the study dataset)
$\widehat{B}_R$	The recipient's benefit, as estimated by the recipient, as estimated by the sharer (regarding the study dataset)
$\hat{C}_R$	The recipient's cost, as estimated by the sharer (regarding the study dataset)
G_R	The gain to the recipient per successfully attacked target
c_a	The cost of access to a target
c_p	The expected cost of an imposed penalty per attack
$\widehat{N}_R$	The number of successfully attacked targets, as estimated by the recipient, as estimated by the sharer
$\pi(g,i)$	The posterior probability, as inferred by the recipient, that target i is in the study dataset
$p[i]$	The prior probability that target i is in the study dataset
$l(g,i)$	The likelihood ratio comparing the likelihood target i is in the study data versus in the reference
$d[i,j]$	The number of minor alleles in SNP j of individual i in the study dataset
$a^*(g)$	The recipient's optimal attacking strategy, given the sharer's strategy g, on targets in the study only
g^*	The sharer's optimal sharing strategy in the Stackelberg game

This subsection provides intuition into the derivation and representation of the payoff functions in the context of the Stackelberg game for genomic summary data sharing. Table 3.1 provides a summary of the notation commonly used throughout this chapter, ordered chronologically.

Figure 3.2 provides an illustration of a strategy profile (*i.e.*, the sharer's strategy and the recipient's strategy), as a part of the payoff matrix in the game model. Both the sharer's and the recipient's payoffs are functions of both their strategies, which we present below.

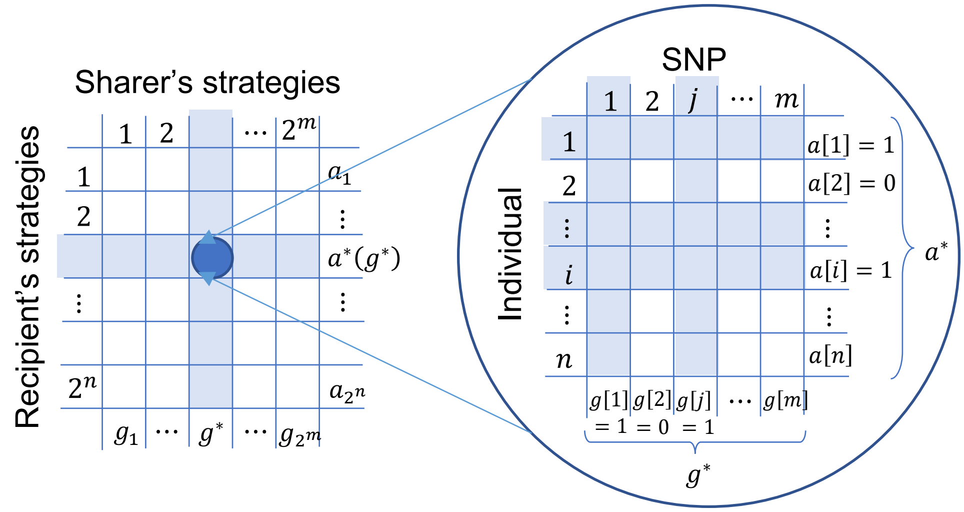


Figure 3.2 An illustration of a strategy profile (on the right) as a part of a payoff matrix (on the left) in our game model.

SNP stands for single nucleotide polymorphism. m is the number of SNPs available for sharing. n is the number of individuals in the study. a^* indicates the recipient's best strategy. g^* indicates the sharer's best strategy.

To begin, let J represent the set of genomic variants, which, in this case, we assume corresponds to the single nucleotide polymorphisms (SNPs), available for sharing. The set of all possible sharing strategies is represented by the power set of J, 2^J, where each strategy is referred to as g and indicates a possible combination of SNPs. We let I be the set of individuals available for attack. Similarly, the set of all possible attacking strategies is represented by the power set of I, 2^I, where each strategy is denoted by a and indicates a possible combination of attacked individuals.

3.2.2.1 Payoff for the Sharer

Generally, the sharer's estimated payoff $\hat{Y}_S$ is represented by a function that indicates the interplay of the sharing strategy g and the attacking strategy a:

$$\hat{Y}_S(g,a) = B_S(g) - \hat{C}_S(a). \tag{3.4}$$

Notice that, in this representation, the sharer's benefit B_S is a function of g, while the sharer's estimated cost C_S is a function of a.

More specifically, based on the assumptions that 1) the sharer's benefit is proportional to the utility of the data and 2) the sharer's cost is proportional to the number of successfully attacked (i.e., identified) targets, the sharer's estimated utility $\hat{Y}_S$ can be represented as:

$$\hat{Y}_S(g,a) = H \cdot U(g) - L_S \cdot \widehat{N}_S(a) \tag{3.5}$$

where H corresponds to the worth of all the data available to the sharer, $U(g)$ is the utility score for shared data, L_S is the loss to the sharer per successfully attacked individual, and $\widehat{N}_S(a)$ is the expected number of successfully attacked targets (i.e., the number of attacked targets that are in the study dataset D) estimated by the sharer given a. The fact that this is estimated, rather than actual, number of successfully attacked targets stems from the data sharer's uncertainty about the target set (i.e., which individuals will actually be targeted by the recipient).

We assume the utility score of the shared data $U(g)$ is additive across all SNPs, such that it is calculated as:

$$U(g) = \sum\nolimits_{j \in J} w[j]g[j] \tag{3.6}$$

where $w[j]$ represents the utility weight of the SNP j, and $\sum_{j \in J} w[j] = 1$. We represent each sharing strategy g as a vector of SNPs, such that $g[j]$ is 1 if the allele frequency for SNP j is shared and 0 otherwise.

We calculate the utility weight $w[j]$ as the normalized absolute difference between $f[j]$ and $\hat{f}[j]$, which correspond to the minor allele frequency of SNP j in the study dataset X and a public reference dataset, respectively. Formally, this is calculated as:

$$w[j] = \frac{\left|f[j] - \hat{f}[j]\right|}{\sum_{j \in J}\left|f[j] - \hat{f}[j]\right|}. \tag{3.7}$$

The sharer estimates the number of successfully attacked targets $\widehat{N}_S(a)$ as:

$$\widehat{N}_S(a) = \sum\nolimits_{i \in I} a[i]\tau[i] \tag{3.8}$$

where $\tau[i]$ is the probability that individual i is targeted. Each attack strategy a is represented as a vector, where $a[i]$ is 1 if individual i is attacked when targeted and 0 otherwise.

3.2.2.2 Payoff for the Recipient from the Sharer's Perspective

The recipient's estimated payoff $\hat{Y}_R$ is a function of the sharing strategy g and the attacking strategy a:

$$\hat{Y}_R(g,a) = \hat{B}_R(g,a) - \hat{C}_R(a) \tag{3.9}$$

where the recipient's estimated benefit $\hat{B}_R$ is a function of both g and a and the recipient's cost $\hat{C}_R$ is a function of a. Notice that the recipient's estimated payoff is only the partial estimated payoff resulting from attacking targets in the study dataset, as the same as the recipient's estimated benefit and cost.

More specifically, based on the assumption that the recipient's benefit is proportional to the number of successfully attacked targets, the recipient's estimated payoff is:

$$\hat{Y}_R(g,a) = G_R \cdot \widehat{N}_R(g,a) - \hat{C}_R(a) \tag{3.10}$$

where G_R is the gain to the recipient per successfully attacked target, while the number of successfully attacked targets $\widehat{N}_R(g,a)$, as estimated by the recipient, is a function of g and a. It should be recognized that, since the recipient does not know the ground truth of who is in the study dataset, the number of successfully attacked targets can only be estimated using shared data.

If we assume that the recipient's cost is proportional to the number of attacked targets, and the recipient's cost per attacked target includes both the cost of access c_a and the expected cost of penalty c_p per attacked target, the recipient's cost $\hat{C}_R$ is:

$$\hat{C}_R(a) = \left(c_a + c_p\right)\sum_{i \in I} a[i]\tau[i]. \tag{3.11}$$

Additionally, the number of successfully attacked targets, as estimated by the recipient $\widehat{N}_R(g,a)$, is:

$$\widehat{N}_R(g,a) = \sum_{i \in I} \pi(i,g)a[i]\tau[i] = \sum_{i \in I} p[i]l(i,g)a[i]\tau[i] \tag{3.12}$$

where $\pi(i,g)$ is the posterior probability, inferred by the recipient given g, that individual i is in the study dataset. Note that, in this representation, $p[i]$ is the prior probability that individual i is in the study dataset, $l(g,i)$ is the likelihood ratio comparing the likelihood that target i is in the study dataset with the likelihood it is in a reference population. Further notice that $\hat{Y}_R(g,a)$ is additive across all individuals: $\hat{Y}_R(g,a) = \sum_{i \in I} \hat{Y}_R[i](g,a[i])$, in which $\hat{Y}_R[i](g,a[i])$ is called the recipient's estimated quantile payoff resulting from attacking individual i, if individual i is targeted. For individual i, this is calculated as:

$$\hat{Y}_R[i](g,a[i]) = \left(G_R \cdot p[i]l(i,g) - c_a - c_p\right)a[i]\tau[i], \quad \forall i \in I. \tag{3.13}$$

We leverage the model introduced by Sankararaman *et al.* [36], such that the likelihood ratio $l(i,g)$ is calculated as:

$$l(i,g) = \frac{\prod_{j \in J} f[j]^{d[i,j]g[j]}(1 - f[j])^{(2-d[i,j])g[j]}}{\prod_{j \in J} \hat{f}[j]^{d[i,j]g[j]}\left(1 - \hat{f}[j]\right)^{(2-d[i,j])g[j]}} \tag{3.14}$$

where $d[i,j] \in \{0,1,2\}$ is the number of minor alleles in SNP j of individual i. It is often easier (because it is more numerically stable) to deal with the log-likelihood ratio, which is calculated as:

$$\log\big(l(i,g)\big) = \sum_{j \in J} d[i,j]g[j]\big(\log f[j] - \log \hat{f}[j]\big) + (2 - d[i,j])g[j]\big(\log(1 - f[j]) - \log\big(1 - \hat{f}[j]\big)\big). \quad (3.15)$$

3.2.2.3 Most Specific Representation of Payoffs

The most specific representation of the sharer's estimated payoff $\hat{Y}_S$, as a function of sharing strategy g and attacking strategy a, is:

$$\hat{Y}_S(g,a) = H \cdot \frac{\sum_{j \in J} g[j]\big|f[j] - \hat{f}[j]\big|}{\sum_{j \in J}\big|f[j] - \hat{f}[j]\big|} - L_S \cdot \sum_{i \in I} a[i]\,\tau[i]. \quad (3.16)$$

The most specific representation of the recipient's estimated quantile payoff $\hat{Y}_R[i]$, as a function of sharing strategy g and attacking strategy a, is:

$$\hat{Y}_R[i](g,a[i]) = \left(G_R \cdot p[i] \frac{\prod_{j \in J} f[j]^{d[i,j]g[j]}(1 - f[j])^{(2-d[i,j])g[j]}}{\prod_{j \in J} \hat{f}[j]^{d[i,j]g[j]}\big(1 - \hat{f}[j]\big)^{(2-d[i,j])g[j]}} - c_a - c_p \right) a[i]\tau[i], \quad \forall i \in I. \quad (3.17)$$

3.2.2.4. Optimization Problem

In a Stackelberg game model, the sharer moves first, while the recipient moves second. In this setting, given a sharing strategy g, an economically-motivated recipient would choose the attacking strategy $a^*(g)$ that maximizes his or her estimated payoff function:

$$a^*(g) = \underset{a}{\operatorname{argmax}}\, \hat{Y}_R(g,a). \quad (3.18)$$

The optimal attacking strategy for a particular target is independent of the attacking strategy invoked for other targets. Thus, the recipient's optimal attacking strategy for individual i, if they are in fact targeted, is:

$$a^*[i](g) = \underset{a[i]}{\operatorname{argmax}}\, \hat{Y}_R[i](g,a[i])\,, \quad \forall i \in I. \quad (3.19)$$

Similarly, an economically-motivated sharer would choose the sharing strategy g^* that maximizes his or her own payoff:

$$g^* = \underset{g}{\operatorname{argmax}}\, \hat{Y}_S\big(g,a^*(g)\big). \quad (3.20)$$

In summary, this subsection provided the equations for calculating the payoffs. To do so, it explained how to discover the optimal strategy for the sharer in the Stackelberg game. **Appendix A.1** provides a detailed derivation of these equations (and the inherent functions) along with their justification.

3.2.3 Solving the Game

The best global solution to the genomic data sharing Stackelberg game can be discovered through an exhaustive search of all possible data sharing policies. However, this is computationally expensive. When it is not computationally practical to search for the optimal solution, heuristics such as a hill-climbing algorithm based on gradient estimation or a genetic algorithm (GA) could be applied to find a locally optimal solution.

In this subsection, we provide globally and locally optimal solutions to the Stackelberg game via a backward induction algorithm (BIA) and a GA, so that other researchers can reproduce our results and validate our methods. To provide some intuition into when to apply the globally optimal BIA or the locally optimal GA, we close this subsection with a comparison of the two approaches in terms of computational complexity and running time in several test cases.

In general, we recommend using GA when the numbers of individuals in the study and the number of SNPs are both large.

3.2.3.1 Common Modules

There are three modules in both algorithms: 1) the filtering module as shown in Figure 3.3, 2) the module for computing LR statistics as shown in Figure 3.4, and 3) the module for computing the sharer's expected payoff as shown in Figure 3.5.

```
$[f, \hat{f}, u, D']$ = Filter $(D, R, \bar{\theta}$, mafcutoff, ldcutoff)
Input:
    The pool dataset (D),
    the reference dataset (R), where each row represents an individual, each column represents a SNP, each cell
    represents the genotype using integer numbers from -1 to 2, where 2 represents minor-minor, 1 represents minor-
    major, 0 represents major-major, and -1 represents missing genotype values,
    the maximal allowable missing rate ($\bar{\theta}$),
    a threshold on minor allele frequency (mafcutoff), and
    a threshold on the p-value indicating linkage disequilibrium (ldcutoff).
Output:
    The minor allele frequencies (MAFs) in pool ($f$),
    the MAFs in reference ($\hat{f}$),
    the utilities for each SNP ($u$), and
    the remaining pool dataset ($D'$).
Main Body:
    Handle missing values:
        FOR EACH SNP in the pool dataset ($D$)
            IF the portion of individuals with missing data is smaller or equal to maximal allowable missing rate ($\bar{\theta}$)
                Remove the SNP from both datasets;
    Compute MAFs:
        FOR EACH SNP in the pool dataset ($D$)
            Compute its MAF in pool ($f$) as the sum of all individuals' values, divided by number of individuals with no
            missing data, divided by 2;
        FOR EACH SNP in the reference dataset ($R$)
            Compute its MAF in reference ($\hat{f}$) as the sum of all individuals' values, divided by number of individuals with
            no missing data, divided by 2;
    Compute utilities associated with each SNP:
        FOR EACH SNP in the pool dataset ($D$)
            Compute its utility ($u$) as the absolute difference between its MAF in pool ($f$) and its MAF in reference ($\hat{f}$);
    Remove SNPs with MAF smaller than mafcutoff:
        FOR EACH SNP in the pool dataset ($D$)
            IF its MAF in pool is smaller than mafcutoff or larger than (1 - mafcutoff)
                Remove the SNP from both datasets ($D, R$), the utility vector ($u$), both MAF vectors ($f, \hat{f}$);
    Let $m$ be the number of remaining SNPs;
    Sort the utility vector in descending order and adjust both datasets, both MAF vectors accordingly;
    Compute the correlation matrix:
        FOR EACH SNP i from 1 to m
            FOR EACH SNP j from 1 to m
                Compute the Chi-square correlation $r^2$ and corresponding p-value for each SNP-pair;
                IF p-value is smaller than ldcutoff
                    SNP i and SNP j are correlated;
    Exclude the SNP outliers:
        Initialize a Boolean vector Selection of all TRUE values;
        Compute the standard deviation of differences between MAF in pool and MAF in reference, $\sigma$;
        Find $n_{drop}$ , the number of SNPs that have differences larger than $6\sigma$;
        Let the first $n_{drop}$ of the vector be FALSE;
    Select a subset of independent SNPs for the sharer:
        FOR EACH SNP i from 1 to (m-1)
            IF Selection(i) is true
                FOR EACH SNP j from (i+1) to m
                    IF SNP i and SNP j are correlated
                        Let Selection(i) be FALSE;
    Trim both datasets, both MAF vectors, and the utility vector according to vector Selection;
    RETURN $f, \hat{f}, u, D$;
```

Figure 3.3 Pseudocode of the filtering module.

```
$LR$ = Compute_LR ($D, f, \hat{f}$)
Input:
    The remaining pool dataset (D),
    the MAFs in pool ($f$), and
    the MAFs in reference ($\hat{f}$),
Output:
    The LR statistics ($LR$).
Main Body:
    Let m be the number of remaining SNPs;
    Let $n$ be the number of individuals in the pool dataset;
    Initialize a n-by-m log-likelihood ratio matrix LR;
    FOR EACH SNP j from 1 to m
        FOR EACH individual i from 1 to n
            IF $D[i,j]$ == 0
                $LR[i,j]$ = 2 * log ((1- $f[j]$)/(1- $\hat{f}[j]$));
            ELSE IF $D[i,j]$ == 1
                $LR[i,j]$ = log ((1- $f[j]$)/(1- $\hat{f}[j]$)) + log ($f[j]$/ $\hat{f}[j]$);
            ELSE IF $D[i,j]$ == 2
                $LR[i,j]$ = 2 * log ($f[j]$/ $\hat{f}[j]$);
            ELSE
                $LR[i,j]$=0;
    RETURN LR;
```

Figure 3.4 Pseudocode of the module for computing LR statistics.

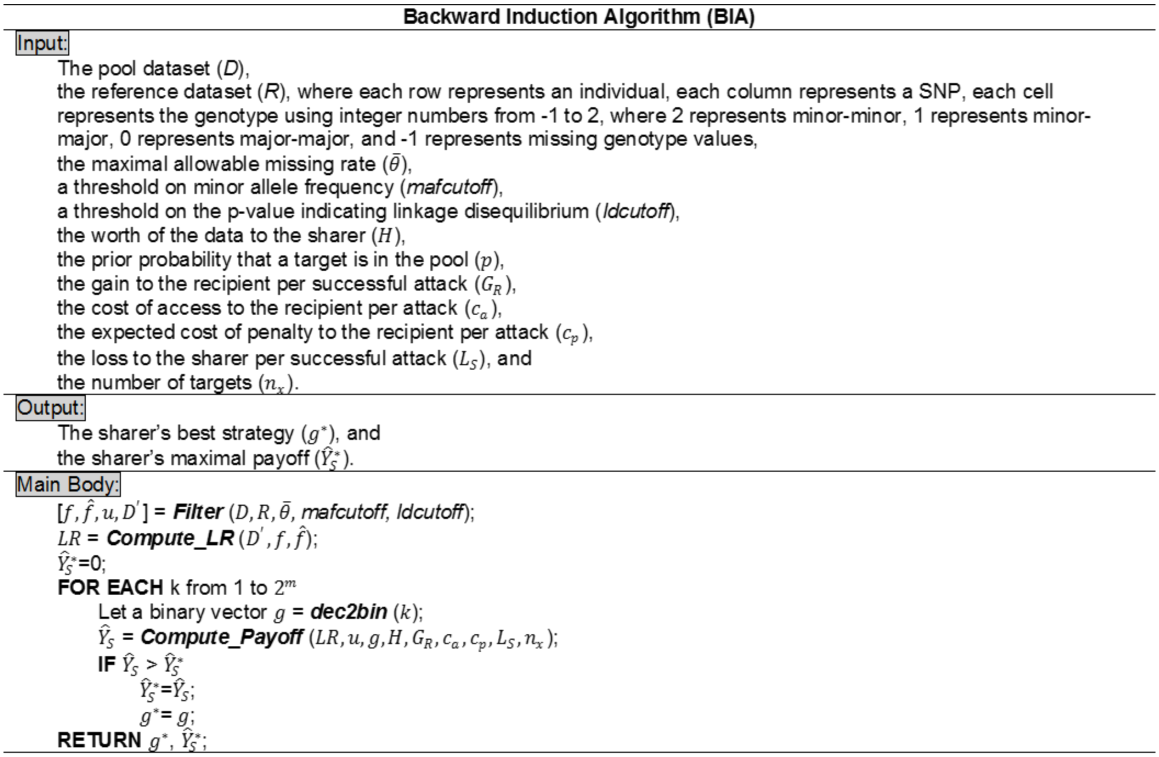
```
$\hat{Y}_S$ = Compute_Payoff ($LR, u, g, H, G_R, c_a, c_p, L_S, n_x$)
Input:
    The LR statistics ($LR$),
    the utilities for each SNP ($u$),
    the sharer's strategy ($g$),
    the worth of the data to the sharer ($H$),
    the prior probability that a target is in the pool ($p$),
    the gain to the recipient per successful attack ($G_R$),
    the cost of access to the recipient per attack ($c_a$),
    the expected cost of penalty to the recipient per attack ($c_p$),
    the loss to the sharer per successful attack ($L_S$), and
    the number of targets ($n_x$).
Output:
    The sharer's expected payoff ($\hat{Y}_S$).
Main Body:
    Let m be the number of remaining SNPs;
    Let $n$ be the number of individuals in the pool dataset;
    Let sum_utility be the sum of the utility vector ($u$);
    Benefit =H*(sum of shared SNPs' utilities)/sum_utility;
    Sum_TP=0;
    FOR EACH individual i from 1 to n
        Sum_LR =0;
        FOR EACH SNP j from 1 to m
            IF SNP j will be shared
                Sum_LR=Sum_LR+$LR[i,j]$;
        Posterior_prob= exp(Sum_LR)*p;
        IF Posterior_prob>1
            Psterior_prob = 1;
        IF $G_R$*Posterior_prob > $(c_a + c_p)$
            Sum_TP=Sum_TP+1;
    Select_rate=$n_x$*p/n;
    Cost = $L_S$* Sum_TP* Select_rate;
    $\hat{Y}_S$ = Benefit – Cost;
    RETURN $\hat{Y}_S$;
```

Figure 3.5 Pseudocode of the module for computing the sharer's expected payoff.

3.2.3.2 Backward Induction Algorithm

In the Stackelberg game, the sharer needs to evaluate his or her payoff for each available strategy. For each of the sharer's strategies, the recipient can play any of their own available strategies. The sharer will choose the strategy that maximizes his or her own payoff. Given the large space of possible strategy combinations, we apply BIA to facilitate the search. BIA is a brute force approach that systematically evaluates all of the possible strategies to discover the one with the maximal payoff. Note that we narrow the strategy space when evaluating the existing SNP suppression protection approach introduced by Sankararaman *et al*. [36].

A detailed description of backward induction is beyond the scope of this subsection. However, for replication purposes, we provide pseudocode designed specifically for our Stackelberg game in Figure 3.6.



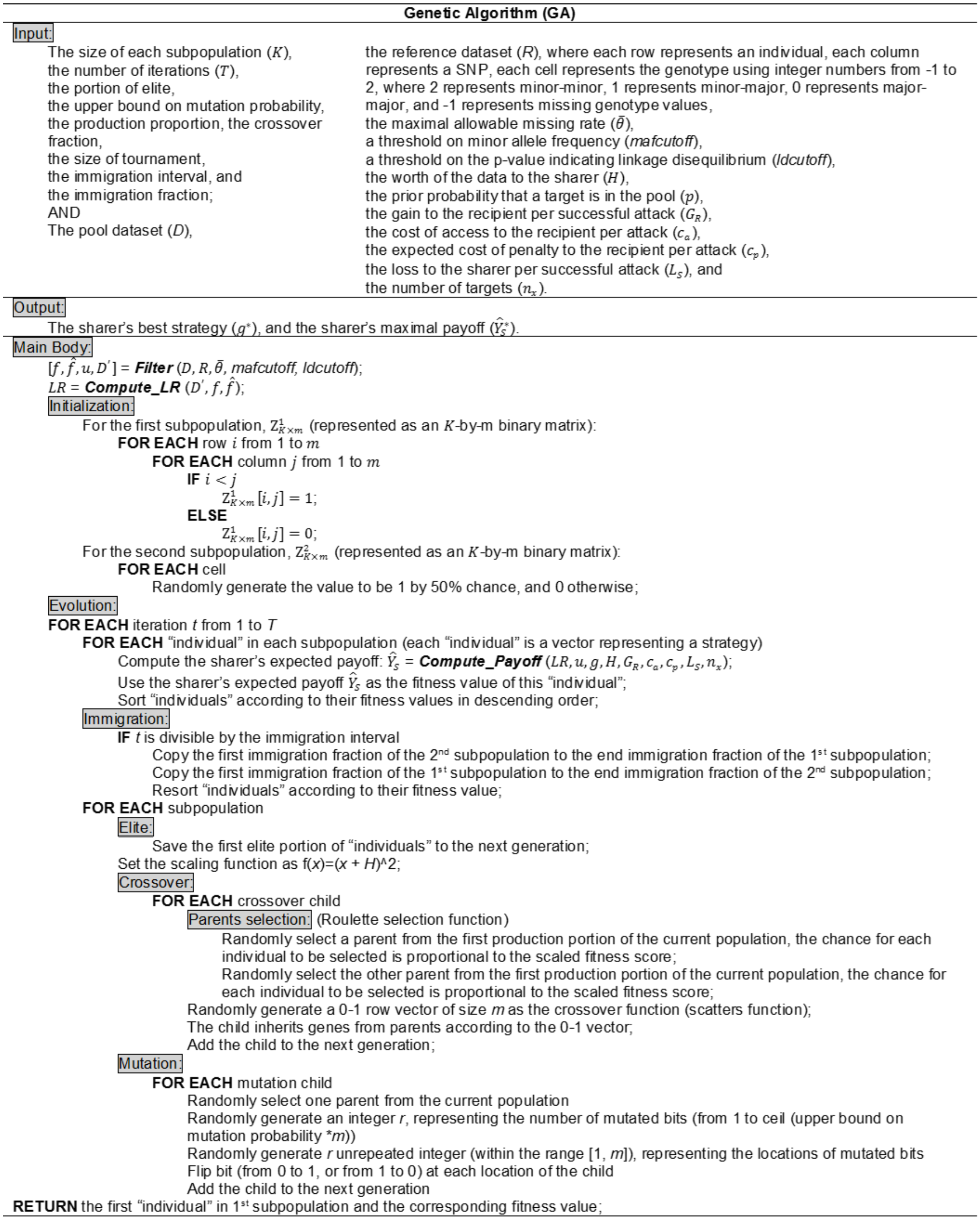
```
Backward Induction Algorithm (BIA)
Input:
    The pool dataset ($D$),
    the reference dataset ($R$), where each row represents an individual, each column represents a SNP, each cell
    represents the genotype using integer numbers from -1 to 2, where 2 represents minor-minor, 1 represents minor-
    major, 0 represents major-major, and -1 represents missing genotype values,
    the maximal allowable missing rate ($\bar{\theta}$),
    a threshold on minor allele frequency (mafcutoff),
    a threshold on the p-value indicating linkage disequilibrium (ldcutoff),
    the worth of the data to the sharer ($H$),
    the prior probability that a target is in the pool ($p$),
    the gain to the recipient per successful attack ($G_R$),
    the cost of access to the recipient per attack ($c_a$),
    the expected cost of penalty to the recipient per attack ($c_p$),
    the loss to the sharer per successful attack ($L_S$), and
    the number of targets ($n_x$).
Output:
    The sharer's best strategy ($g^*$), and
    the sharer's maximal payoff ($\hat{Y}_S^*$).
Main Body:
    $[f, \hat{f}, u, D']$ = Filter ($D, R, \bar{\theta}$, mafcutoff, ldcutoff);
    $LR$ = Compute_LR ($D', f, \hat{f}$);
    $\hat{Y}_S^*$=0;
    FOR EACH k from 1 to $2^m$
        Let a binary vector $g$ = dec2bin ($k$);
        $\hat{Y}_S$ = Compute_Payoff ($LR, u, g, H, G_R, c_a, c_p, L_S, n_x$);
        IF $\hat{Y}_S$ > $\hat{Y}_S^*$
            $\hat{Y}_S^*$=$\hat{Y}_S$;
            $g^*$= $g$;
    RETURN $g^*$, $\hat{Y}_S^*$;
```

Figure 3.6 Pseudocode of the backward induction algorithm (BIA).

3.2.3.3 Genetic Algorithm

A genetic algorithm (GA) is an optimization approach inspired by evolutionary processes. In each iteration, the candidates with higher fitness scores have greater probabilities of being selected. Upon selection, these candidates are subject to crossover and mutation processes to produce a population of the next generation. The crossover assists in the exchange of discovered knowledge (much like genes exchanged between individuals), while the mutation helps in restoring lost or unexplored regions in search space (much like the accumulation induces loss or gain of function in a gene).

The detailed steps in our GA implementation are shown as pseudo-codes in Figure 3.7.



```
Genetic Algorithm (GA)
Input:
    The size of each subpopulation ($K$),
    the number of iterations ($T$),
    the portion of elite,
    the upper bound on mutation probability,
    the production proportion, the crossover fraction,
    the size of tournament,
    the immigration interval, and
    the immigration fraction;
    AND
    The pool dataset ($D$),
    the reference dataset ($R$), where each row represents an individual, each column represents a SNP, each cell represents the genotype using integer numbers from -1 to 2, where 2 represents minor-minor, 1 represents minor-major, 0 represents major-major, and -1 represents missing genotype values,
    the maximal allowable missing rate ($\bar{\theta}$),
    a threshold on minor allele frequency (mafcutoff),
    a threshold on the p-value indicating linkage disequilibrium (ldcutoff),
    the worth of the data to the sharer ($H$),
    the prior probability that a target is in the pool ($p$),
    the gain to the recipient per successful attack ($G_R$),
    the cost of access to the recipient per attack ($c_a$),
    the expected cost of penalty to the recipient per attack ($c_p$),
    the loss to the sharer per successful attack ($L_S$), and
    the number of targets ($n_x$).
Output:
    The sharer's best strategy ($g^*$), and the sharer's maximal payoff ($\hat{Y}_S^*$).
Main Body:
    $[f, \hat{f}, u, D']$ = Filter ($D, R, \bar{\theta}$, mafcutoff, ldcutoff);
    $LR$ = Compute_LR ($D', f, \hat{f}$);
    Initialization:
        For the first subpopulation, $Z_{K\times m}^1$ (represented as an $K$-by-m binary matrix):
            FOR EACH row $i$ from 1 to $m$
                FOR EACH column $j$ from 1 to $m$
                    IF $i < j$
                        $Z_{K\times m}^1[i,j] = 1$;
                    ELSE
                        $Z_{K\times m}^1[i,j] = 0$;
        For the second subpopulation, $Z_{K\times m}^2$ (represented as an $K$-by-m binary matrix):
            FOR EACH cell
                Randomly generate the value to be 1 by 50% chance, and 0 otherwise;
    Evolution:
    FOR EACH iteration $t$ from 1 to $T$
        FOR EACH "individual" in each subpopulation (each "individual" is a vector representing a strategy)
            Compute the sharer's expected payoff: $\hat{Y}_S$ = Compute_Payoff ($LR, u, g, H, G_R, c_a, c_p, L_S, n_x$);
            Use the sharer's expected payoff $\hat{Y}_S$ as the fitness value of this "individual";
            Sort "individuals" according to their fitness values in descending order;
        Immigration:
            IF $t$ is divisible by the immigration interval
                Copy the first immigration fraction of the 2nd subpopulation to the end immigration fraction of the 1st subpopulation;
                Copy the first immigration fraction of the 1st subpopulation to the end immigration fraction of the 2nd subpopulation;
                Resort "individuals" according to their fitness value;
        FOR EACH subpopulation
            Elite:
                Save the first elite portion of "individuals" to the next generation;
            Set the scaling function as f(x)=(x + H)^2;
            Crossover:
                FOR EACH crossover child
                    Parents selection: (Roulette selection function)
                        Randomly select a parent from the first production portion of the current population, the chance for each individual to be selected is proportional to the scaled fitness score;
                        Randomly select the other parent from the first production portion of the current population, the chance for each individual to be selected is proportional to the scaled fitness score;
                    Randomly generate a 0-1 row vector of size m as the crossover function (scatters function);
                    The child inherits genes from parents according to the 0-1 vector;
                    Add the child to the next generation;
            Mutation:
                FOR EACH mutation child
                    Randomly select one parent from the current population
                    Randomly generate an integer r, representing the number of mutated bits (from 1 to ceil (upper bound on mutation probability *m))
                    Randomly generate r unrepeated integer (within the range [1, m]), representing the locations of mutated bits
                    Flip bit (from 0 to 1, or from 1 to 0) at each location of the child
                    Add the child to the next generation
RETURN the first "individual" in 1st subpopulation and the corresponding fitness value;
```

Figure 3.7 Pseudocode of the genetic algorithm (GA) implementation.

3.2.3.4 Performance Analysis

We compare BIA and GA on two types of performance: 1) computational complexity and average running time and 2) accuracy, to determine if GA is 1) more efficient and 2) close enough to BIA in terms of computational results. To facilitate the comparison, we need to simulate the input datasets on large scales.

In a large-scale simulation, the population dataset has m SNPs and n_p individuals. The minor allele frequency (MAF) for each SNP is uniformly distributed in the range [0.1, 0.3]. Each SNP is generated according to the MAFs independently. The missing rate for each SNP is 0.05. For each individual in the population dataset, if and only if the number of minor alleles in the first 20% SNPs is larger than $m/10$, the probability of illness is 1. The pool dataset consists of $n = 0.2 \times n_p$ individuals (a sampling rate of 0.2) randomly selected from the ill individuals, and all the SNPs. The reference dataset consists of $n_r = 0.3 \times n_p$ individuals randomly selected from the population, and all the SNPs. The target dataset consists of $n_x = 0.25 \times n_p$ individuals randomly selected from the population, and all the SNPs.

Table 3.2 Parameter settings for both the backward induction algorithm and the genetic algorithm in the performance analysis.

Parameter	**Setting**
The loss to the sharer per successful attack (L_S)	\$360
The gain to the recipient per successful attack (G_R)	\$360
The worth of the data to the sharer (H)	$\$90 \times n$
The cost of access to the recipient per attack (c_a)	\$60
The expected cost of penalty to the recipient per attack (c_p)	\$100
The prior probability that a target is in the study (p)	0.2
The threshold on minor allele frequency (*mafcutoff*)	0.05
The threshold on the *p*-value indicating linkage disequilibrium (*ldcutoff*)	0.05
The maximal allowable missing rate ($\bar{\theta}$)	0.2
The number of targets (n_x)	$1.25 \times n$

n is the number of individuals in the study.

In GA, T is the number of iterations, and K is the size of each subpopulation. The computational complexity of each algorithm is represented as a function of four key parameters: n, m, T, and K. We set the parameters used by both BIA and GA, as shown in Table 3.2, and set the parameters used only by GA as shown in Table 3.3. To measure the runtime (in milliseconds, or ms), we use an Intel Core i7 3GHz machine with 8 GB random-access memory (RAM). The average running time is the world clock time averaged across 37 repeated experimental runs.

Table 3.3 Parameter settings for the genetic algorithm in the performance analysis.

Parameter	**Setting**
The size of each subpopulation (K)	$m+1$
The number of iterations (T)	50
The portion of the elite	0.2
The upper bound on mutation probability	0.05
The production proportion	0.8
The crossover fraction	0.4
The size of the tournament	10
The immigration interval	10
The immigration fraction	0.2

m is the number of SNPs available for sharing.

Table 3.4 Comparison of algorithms on the computational complexity and the running time

	Computational complexity	**Average running time (ms)**				
		m=5, n=20	m=5, n=200	m=15, n=200	m=15, n=2000	m=20, n=200
Backward Induction Algorithm (BIA)	$O(mn^2 2^m)$	3	4	4,685	19,831	149,465
Genetic Algorithm (GA)	$O(mnTK)$	781	781	783	789	791
BIA / GA	$O(n2^m/TK)$	0.0038	0.0051	5.9834	25.1343	188.9570

n is the number of individuals in the study. m is the number of SNPs available for sharing. T is the number of iterations. K is the size of each subpopulation.

Table 3.5 Comparison of algorithms on computational results.

	Sharer's payoff ($)				
	m=5, n=20	m=5, n=200	m=15, n=200	m=15, n=2000	m=20, n=200
Backward Induction Algorithm (BIA)	755	17,280	16,560	169,020	15,660
Genetic Algorithm (GA)	755	17,280	16,560	169,020	15,660
GA / BIA	1	1	1	1	1

n is the number of individuals in the study. m is the number of SNPs available for sharing.

Table 3.4 shows the performance comparison of two algorithms on computational complexity and running time, based on the study dataset with varying size. It can be seen that initially, when there are only 5 SNPs and 20 individuals, GA is approximately 263x slower than BIA. However, as the size of the search space grows, the running time for BIA quickly outpaces GA. By the time there are 20 SNPs and

200 individuals, GA is 189x faster than BIA. Given that the dataset used in our case study contains hundreds of SNPs and thousands of individuals, we applied GA in all of our empirical investigations.

Table 3.5 shows the performance comparison of two algorithms on the computational results of the sharer's payoff. It can be seen that the results of the two algorithms are the same.

3.3 Results

3.3.1 Dataset and Experimental Setup

This formalized adversarial setting enables us to evaluate various types of genomic data sharing policies in terms of balancing the data utility against the privacy risk. We specifically consider several types of policies, which build on the conventional legal and technical approaches identified above. The first type of policies is the traditional DUA model, whereby the sharer invokes an all-or-nothing publication approach, controlling for the level of liquidated damages associated with a breach of contract. The second type of policies is the traditional SNP suppression model, whereby the sharer reveals summary statistics of the top-ranked SNPs (sorted in a descending order according to their utility weights, which can correspond to the normalized absolute difference between the minor allele frequency in the study and the underlying reference population), but does not invoke a DUA and therefore lacks the ability to penalize the recipient for attempting to identify individuals. The third type of policies allows for a combination of SNP suppression and penalties derived from DUA violations in the Stackelberg game. Finally, we note that focusing solely on the payoff may be considered ethically unsound for institutional review boards that seek to minimize risk. Thus, the fourth type of policies is a special case of the Stackelberg game in which no risk exists because the recipient chooses not to attack, judging that the total cost exceeds any potential benefit he or she could gain.

3.3.1.1. Dataset

We evaluate these policies with the data of 8,194 individuals from the SPHINX program[11]. The genomic data includes 82 genes identified as important for pharmacogenomics, with 38,112 variant regions (i.e., areas of the genome with any type of notable variation across individuals) including 51,826

[11] Sequence and Phenotype Integration Exchange. https://www.emergesphinx.org/.

SNPs, while the phenomic data includes various clinical factors extracted from the electronic medical records of these participants (e.g., diagnosis codes, procedural codes, and medications). SPHINX publishes all summary statistics and requires an end user licensing agreement that prohibits re-identification attempts. To simulate the attack, we assume the recipient is provided with 2,500 named genomic records, 125 (or 5%) of them are in SPHINX, and the others are only in a reference population, for which the 1000 Genomes Phase 3 resource [230] is representative. Figure 3.8 illustrates the relationships between the population and subpopulations associated with the datasets in this experiment. Table 3.6 summarizes the basic information about the two datasets used in our experiments.

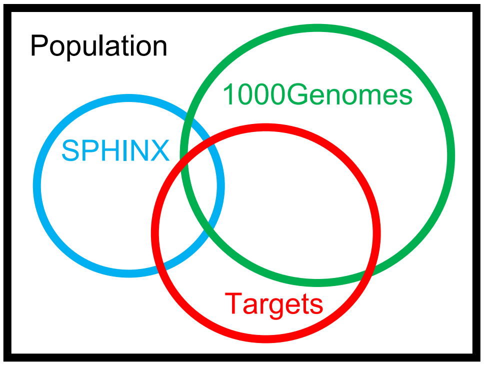


Figure 3.8 Relationships between the population and subpopulations associated with the datasets in the Sequence and Phenotype Integration Exchange (SPHINX) case study.

Table 3.6 Datasets used in our experiments.

Dataset	Role	Number of Individuals	Number of SNPs
SPHINX	Study (pool) dataset	8,194	51,826
1000 Genomes Phase 3	Reference dataset	2,504	> 2 millions

SNP stands for single nucleotide polymorphism. SPHINX stands for Sequence and Phenotype Integration Exchange.

3.3.1.2 Parameters

The parameters with their descriptions, settings, and justifications are provided in Table 3.7. We set the values of these parameters in the case study according to real-world situations. First, the worth of data to the recipient is set to $360 (based on an industry study in 2016) [271]. (The worth of data has been reduced to $150 according to latest version of the industry study in 2020 [272].) Second, the value of the data to the sharer is set to the grant dollars received to share all the genomic summary statistics. However, in this case study, we do not use a specific grant as an example but set the value of the data to the maximal cost the sharer will have. The maximal cost of the sharer is \$45,000 for two reasons: 1) we assume, in the case study, the sharer's loss per successful attack is equal to the recipient's gain per successful attack which is \$360. 2) The prior probability a target is in the study is set to 0.05 such that at most 125 study subjects will be attacked. Therefore, the maximal cost is the product of \$360 and 125 which is \$45,000. An advantage of this setting is that it makes the measurements of the privacy and the utility equally weighted. Whenever the maximal data utility exceeds the maximal privacy cost, the utility will be weighted more than the privacy because the sharer makes decisions according to the payoff which is proportional to the difference between the data utility and privacy cost. Third, the expected penalty is set to be the half of the worth of the data to the recipient, for several reasons: 1) the recipient will never attack if the expected penalty exceeds the worth of the data to the recipient. 2) \$180 is the expected value if the expected penalty is randomly selected within the range from 0 to \$360. 3) The expected penalty is the actual penalty times a crime detection rate which is around 0.2 [273]. Fourth, the expected cost to access the target includes the payment to an online data broker such as Intelius.com for the target's contact information in the background report, which is around \$60.

Table 3.7 Parameter settings in the SPHINX case study.

Parameter	Description	Setting	Justification
p	Prior probability (i.e., sampling rate)	0.05	Averaged sampling rate across all eMERGE sites
m	Number of SNPs	267	Remain after the filtering process
n_x	Number of targets	2500	Approximate size of the 1000Genomes dataset
L_S	Sharer's loss per successful attack	\$360	$L_S = G_R$
G_R	Recipient's gain per successful attack	\$360	According to an industry report [271]
c_p	Recipient's expected cost of penalty	\$180	$c_p = L_S/2$
c_a	Cost of access to the recipient	\$60	Common cost of marketing
H	Maximal benefit of the sharer	\$45000	$H = n_x \cdot p \cdot L_S$
$mafcutoff$	Minimal allowable MAF	0.0001	It is small because "rare variants are extant with minor allele frequencies" are in SPHINX
$ldcutoff$	Maximally allowable p-value on Pearson chi-squared correlation	0.05	Common value for p-value threshold
$\bar{\theta}$	maximal allowable missing rate	0.1	As used by Sankararaman *et al.* [36]

SNP stands for single nucleotide polymorphism. eMERGE stands for Electronic Medical Records and Genomics. MAF stands for minor allele frequency. SPHINX stands for Sequence and Phenotype Integration Exchange.

Additionally, we specifically highlight that the sampling rate p in the SPHINX dataset is set to 0.05. As noted in the main manuscript, the SPHINX program is based on the NIH-sponsored Electronic Medical Records and Genomics (eMERGE) network [54], and each of 10 participating sites has a biobank from which samples were selected for the SPHINX program, the rate of which are shown in Table S22. The sampling rate of 0.05 corresponds to the average rate at which individuals were selected.

Table 3.8 Sampling rate in the SPHINX dataset from the biorepositories of the member sites of the eMERGE network.

eMERGE Sites	Individuals in SPHINX	Individuals in eMERGE Biorepositories	Sampling Rate
Boston Children's Hospital	109	3,372	0.0323
Group Health Cooperative of Puget Sound	990	5,859	0.1689
Cincinnati Children's Hospital	765	8,472	0.0903
Northwestern University Medical Center	731	12,000	0.0609
Marshfield Clinic	747	20,000	0.0374
Mount Sinai School of Medicine	888	25,000	0.0355
Geisinger Health System	1,086	35,000	0.0310
Mayo Clinic	1,013	36,000	0.0281
Children's Hospital of Philadelphia	1,783	160,000	0.0111
Vanderbilt University Medical Center	903	155,000	0.0058
Average	**901.5**	**46,070.3**	**0.0502**

SPHINX stands for Sequence and Phenotype Integration Exchange. eMERGE stands for Electronic Medical Records and Genomics.

For the SNP suppression model, we assume the maximal allowable false positive rate is 0.1 (or 10%) and the maximal allowable true positive rate is 0.5 (or 50%), respectively as set by Sankararaman et al. [36]. The sharer relies on common biallelic autosomal SNPs among the SPHINX dataset, the 1000 Genomes resource, and the dbSNP database[12]. Nevertheless, the sharer filters out SNPs to ensure the missing rate for each SNP below 0.1, the variant rate above 0.0001, and the p-value of the Pearson χ^2 correlation test above 0.05 (see **Appendix A.2** for details of the filtering process). The sharer prioritizes the remaining 267 SNPs according to the absolute difference in their variant rates between the study and the reference population.

3.3.2 Case Study

To set the stage, let us take a moment to provide an example of how a game theoretic approach can improve data sharing (see **Appendix A.3** for details of the case study). Figure 3.9 illustrates how the system can be mapped into two dimensions: (i) utility of the shared data, where higher is better, and (ii)

[12] The NCBI Short Genetic Variations database. http://www.ncbi.nlm.nih.gov/SNP/.

privacy protection of the data – again, where higher is better. In this case, the utility is directly related to the absolute difference between the minor allele frequencies of shared SNPs in the study and their known minor allele frequencies in the underlying reference population. A *utility* score of 1 is achieved when information about all SNPs is shared. Additionally, *privacy* is inversely related to *risk*, which, in this setting, is the likelihood that a recipient achieves success in compromising the privacy of targeted individuals. A *privacy* score of 1 is achieved when no attacks are successful, in other words, when no risk exists (i.e., *privacy* equals one minus *risk*). For orientation, the ideal policy would be realized when *utility* and *privacy* are both maximized (a score of 1). However, in practice, there is a fundamental tradeoff between utility and privacy. A *payoff* score (which is akin to the net benefit) is defined as the difference between *utility* and *risk*. The figure reveals several notable findings. First, policies based on SNP suppression (of which there are a variety of options that suppress subsets of SNPs) achieve the lowest payoffs, among which the existing "best" solution, according to Sankararaman et al.'s approach [36], increases the level of risk substantially by around 300% but only improves the utility by less than 5% when compared to the game theoretic policy. Second, the policy that relies solely on a DUA (with an expected penalty of $180 per violation) is the next best option. Third, the best policy option is realized in a game theoretic setting that combines an SNP suppression approach with a DUA, which has the ability to institute penalties for contractual violations.

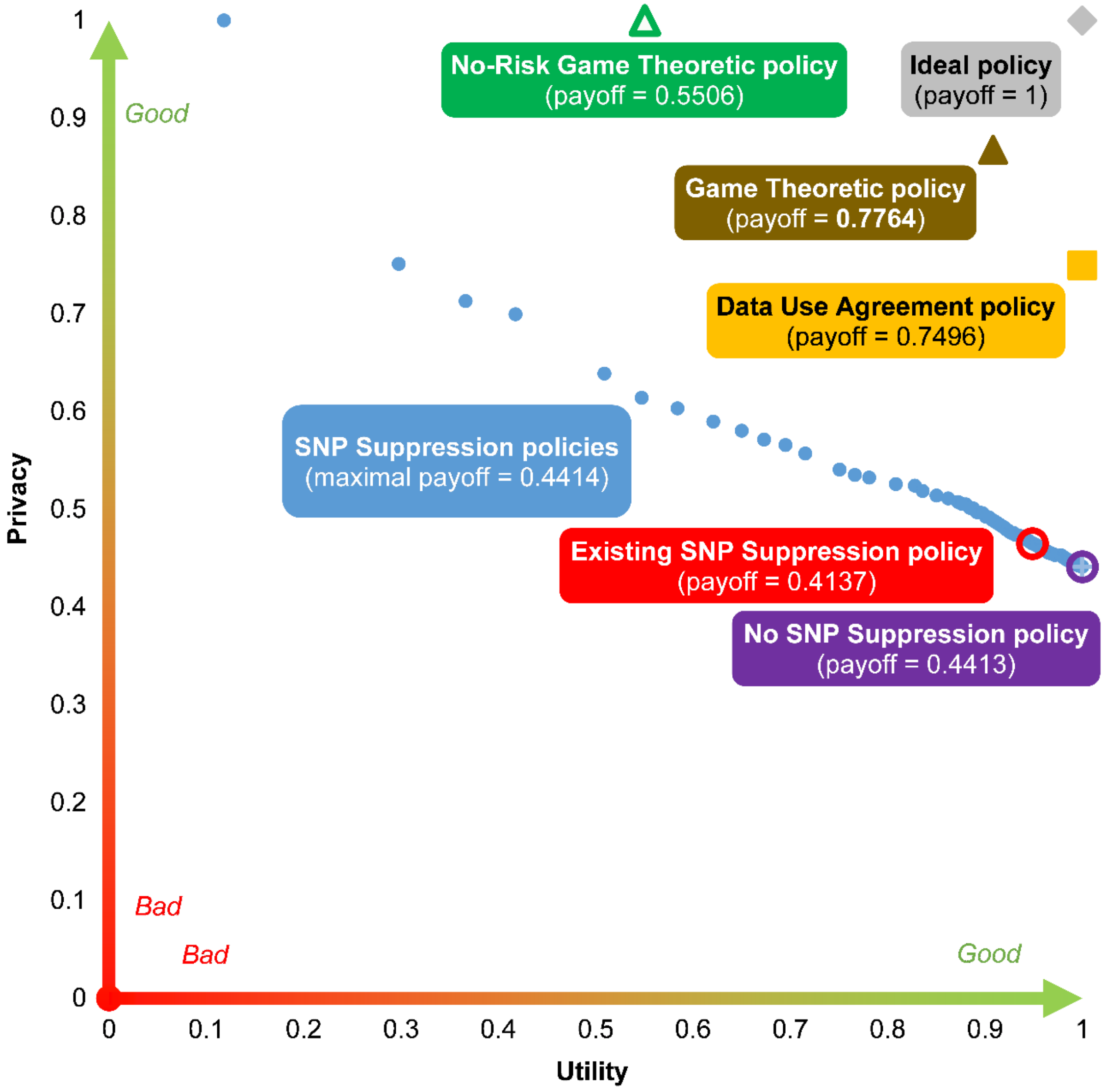


Figure 3.9 A comparison of genomic summary data sharing policies for participants in the SPHINX program.

The compared policies include: 1) the single nucleotide polymorphism (SNP) Suppression policies rely only on hiding of genomic regions (blue dots), 2) the Existing SNP Suppression policy according to Sankararaman et al.'s approach (red circle), 3) the Data Use Agreement (DUA) policy relies only on a legally enforceable contract (gold square), 4) the Game Theoretic policy allows for a combination of a DUA and SNP suppression in a Stackelberg framework (brown triangle), 5) the No-Risk Game Theoretic policy ensures no attack is committed by the recipient (green outlined triangle) and 6) the No SNP Suppression policy illustrates what transpires when no DUA or SNP suppression is applied (purple circle). *Utility* is directly related to the absolute difference between the minor allele frequencies of shared SNPs in the study and their known minor allele frequencies in the underlying reference population (a utility score of 1 is achieved when all SNPs are shared). *Privacy* is inversely related to risk, the likelihood a recipient achieves success in compromising the privacy protection of targeted individuals (a privacy score of 1 is achieved when no attacks are successful, in other words, when no risk exists). A higher *payoff* value represents a more desirable option. SPHINX stands for Sequence and Phenotype Integration Exchange of the eMERGE Network. eMERGE stands for Electronic Medical Records and Genomics.

3.3.3 Sensitivity Analysis

Figure 3.10 depicts the expected payoff of the genomic data sharing policies with respect to a range of liquidated damages associated with the expected penalty of the DUA. The overall payoff (the main graph on the right) is the result of combining (i) the privacy protection afforded to the targeted individuals (the upper graph on the left) and (ii) the utility in the set of SNPs that are shared (the lower graph on the left). The policy derived from the Stackelberg game always achieves the highest payoff to the sharer, regardless of the penalty level. There are several additional findings that we wish to highlight. First, the SNP suppression policy is almost always the worst in terms of privacy. Second, the no-attack game policy leads to a considerably lower utility for the sharer than the other policies, but it still enables a substantial quantity of genomic data to be revealed with a guarantee of full protection. Overall, these results show that, when appropriately managed, deterrents can provide peace of mind when sharing genomic data in the face of re-identification threats.

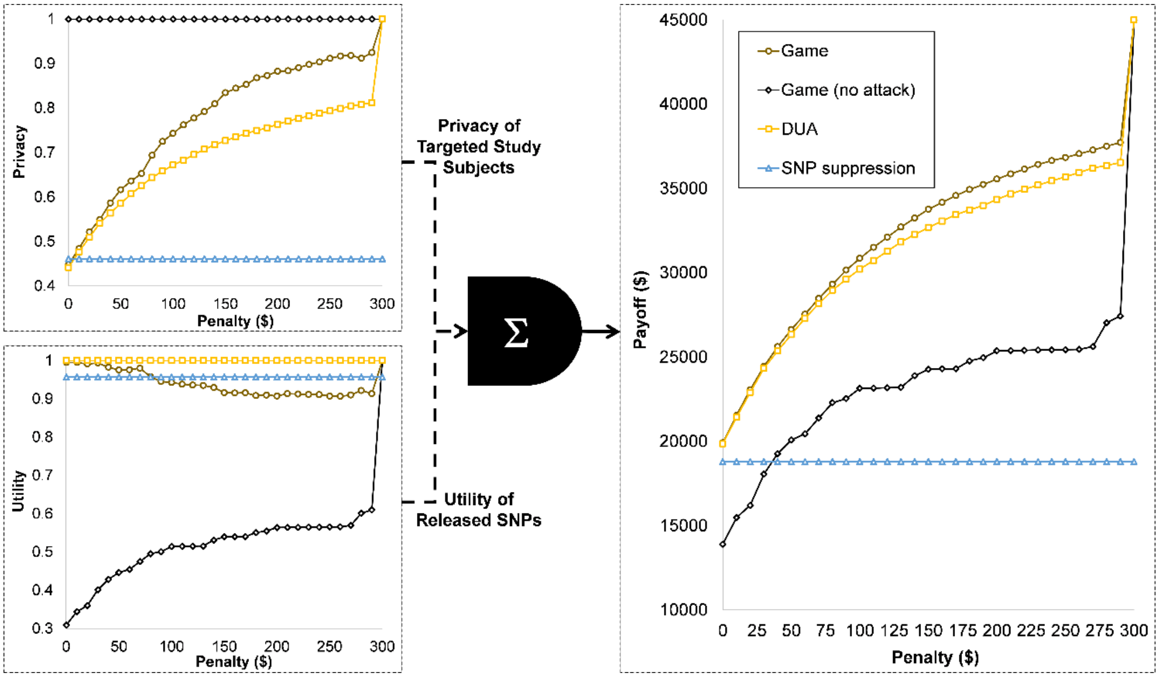


Figure 3.10 Comparisons of four protection policies for the SPHINX program with a varying penalty against the genomic inference attack.

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (yellow lines), and 4) the SNP suppression solution (blue lines) with no penalty. The overall payoff (the main graph on the right) is the result of combining (i) the privacy protection afforded to the targeted individuals (the upper graph on the left) and (ii) the utility in the set of SNPs that are shared (the lower graph on the left). SPHINX stands for Sequence and Phenotype Integration Exchange of the eMERGE Network. eMERGE stands for Electronic Medical Records and Genomics.

Next, we consider how the prior probability influences the data sharing policies. Specifically, we set the prior to simulate four genome sequencing programs: (i) the Precision Medicine Initiative or PMI [231], which will have a prior of 0.003 for 1 million participants out of 318 million US citizens, (ii) the Million Veteran Program or MVP [232], which has a prior of 0.002 for 400,000 participants out of 21 million U.S. military veterans, (iii) the BioVU de-identified DNA repository program of the Vanderbilt University Medical Center [233], which has a prior of 0.1 for 200,000 participants out of 2 million individuals with electronic medical records at the institution, and (iv) the Rare Diseases Clinical Research Network or RDCRN, which we assume has a prior of 0.5 for coverage of half of the possible population. For context, we compare these with the 0.05 prior probability relied upon in the experiments for SPHINX described above.

Figure 3.11 depicts the payoff of the genomic policies as a function of the prior probability. There are several notable findings. First, regardless of the prior probability, the game policies always outperform those relying upon a DUA or SNP suppression in isolation with respect to overall payoff (the main graph on the right). Additionally, the DUA policy always outperforms the SNP suppression policy. Second, when the prior probability is relatively small, as in PMI, the difference between the DUA and the game policies is negligible. Third, the sharer's payoff is negatively correlated with the prior probability, regardless of the policy. Yet unlike the DUA and SNP suppression policies, the game policies are robust because they provide a higher payoff to the sharer even when the prior probability increases substantially.

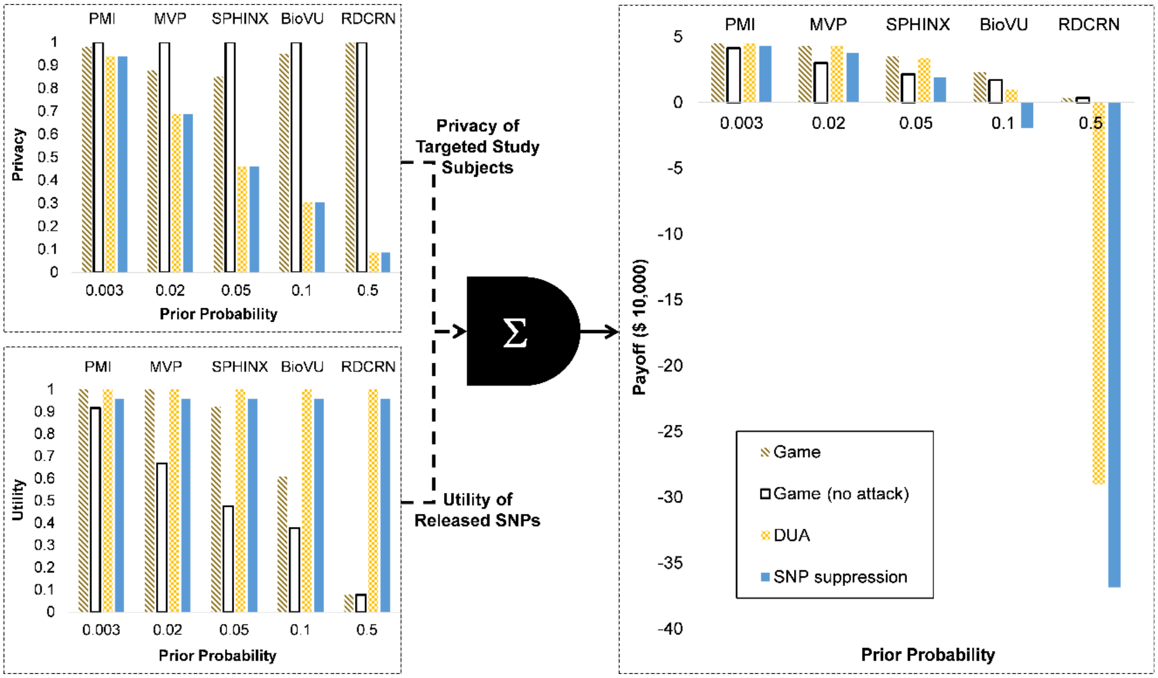


Figure 3.11 Comparisons of four protection policies for a range of genomic data sharing programs with varying prior probabilities against the genomic inference attack.

The compared policies include 1) the optimal game theoretic solution (brown bars filled with downward diagonal pattern), 2) the game theoretic solution that ensures no attack is successful (black bars with no fill), 3) the data use agreement (DUA) (gold bars filled with checkerboard pattern), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue bars with a solid fill). The overall payoff (the main graph on the right) is the result of combining (i) the privacy protection afforded to the targeted individuals (the upper graph on the left) and (ii) the utility in the set of SNPs that are shared (the lower graph on the left). PMI stands for Precision Medicine Initiative. MVP stands for Million Veteran Program. SPHINX stands for Sequence and Phenotype Integration Exchange of the eMERGE Network. eMERGE stands for Electronic Medical Records and Genomics. BioVU stands for de-identified biorepository of Vanderbilt University Medical Center. RDCRN stands for Rare Diseases Clinical Research Network.

3.4 Discussion

These results are remarkable because they demonstrate that blending economic, legal, and technical approaches can dramatically improve our ability to strike the right balance between data utility and privacy risk in genomic data sharing. While game theoretic approaches have the potential to revolutionize how policies are designed, there are several challenges that need to be addressed.

First, the model introduced for this investigation assumes the sharer knows the size of the recipient's target set. This is unlikely to be the case in practice, such that the game simulated by the sharer could yield a suboptimal payoff and even overestimate the privacy risk. There are, however, various approaches to address uncertainty in the target set. One conservative approach would be to assume the recipient is aware of the names of every participant in the study (as well as additional people). Simulating the game in this scenario would likely lead to a genomic data sharing policy that favors privacy over utility.

The second challenge is that the reliability of the game-based results is contingent upon our ability to estimate the worth of data, whether it is phenomic or genomic. Nevertheless, the value of data is likely dependent on its anticipated use. In certain settings, such valuation may be readily accessible. Consider, in the event that data is used to discriminate against an individual with respect to insurability or employment (even though using genomic data could violate federal laws, such as the Genetic Information Nondiscrimination Act [274] and the Americans with Disabilities Act) [275], estimates of the worth of the data might be affected by the value of the expected federal penalties. However, in other settings, the worth may be more difficult to ascertain. For instance, in the event that data is used in more black-market-like settings, such as for blackmail or medical identity fraud, the value of the data might be proportional to the monetary worth (or assets) of the targeted individual. Here as well, modeling several different sets of values may provide guidance for setting protection models.

We recognize that this approach depends on numerous parameters and have thus empirically evaluated the game under a wide range of settings. The findings indicate that the sharer's expected payoff (see **Appendix A.4** for details of an extensive sensitivity analysis) and the superiority of the game policy (see **Appendix A.5** for details of a robustness analysis) are resilient to changes in many model parameters. This suggests that rough estimates of the parameters may be sufficient to put such policies into practice. Still, we acknowledge that a recipient's capabilities may evolve over time. From a technical perspective, the recipient's ability to discern the presence of an individual (e.g., via inference) may grow due to advances in biological knowledge. From an economic stance, the worth of the data may change (e.g., increase) over time. If future-proofing of data sharing scenarios is required, it may be prudent for data sharers to inflate the capabilities of the recipient's prediction model or their valuation of data. At the same time, concerns over future adversaries are conditioned on the expectation that there is only a one-time release of data or summary statistics. As an alternative (or complement to inflating future adversarial capability) to reinforce accountability, data-sharing efforts could be performed in a manner that requires recipients to enter into DUAs that are subject to modification over time to ensure accountability. Such an approach, however, could only be realized in a setting where data has not been put into the public domain.

Our approach is naturally expandable to other genomic summary data sharing scenarios and corresponding adversarial models. For instance, by updating the way of computing likelihood ratios, our approach has the potential to address Shringarpure and Bustamante's attack [86] which is a more powerful attack targeting genomic data-sharing beacons that counteract Sankararaman et al.'s attack.

Chapter 4

BEACON SERVICES GAME

Genomic data is increasingly collected by a wide array of organizations. As such, there is a growing demand to make summary information about such collections available more widely. However, over the past decade, a series of investigations have shown that attacks, rooted in statistical inference methods, can be applied to discern the presence of a known individual's DNA sequence in the pool of subjects. Recently, it was shown that the Beacon Project of the Global Alliance for Genomics and Health, a web service for querying about the presence (or absence) of a specific allele, was vulnerable. The Integrating Data for Analysis, Anonymization, and Sharing (iDASH) Center modeled a track in their 3rd Privacy Protection Challenge on how to mitigate the Beacon vulnerability. We developed the winning solution for this track.

This chapter describes our computational method to optimize the tradeoff between the utility and the privacy of Beacon services. We generalize the genomic data sharing problem beyond that which was introduced in the iDASH Challenge to be more representative of real-world scenarios to allow for a more comprehensive evaluation. We then conduct a sensitivity analysis of our method with respect to several state-of-the-art methods using a dataset of 400,000 positions in Chromosome 10 for 500 individuals from Phase 3 of the 1000 Genomes Project. All methods are evaluated for utility, privacy, and efficiency.

Our method achieves better performance than all state-of-the-art methods, irrespective of how key factors (e.g., the allele frequency in the population, the size of the pool, and utility weights) deviate from the original parameters of the problem. We further illustrate that it is possible for our method to exhibit subpar performance under special cases of allele query sequences. However, we show our method can be extended to address this issue when the query sequence is fixed and known a priori to the data custodian so that they may plan stage their responses accordingly.

I further provide additional descriptions of how our method can be explicitly combined with game theoretic approaches to protect the genomic data sharing environment better.

This research shows that it is possible to thwart the attack on Beacon services without substantially altering the utility of the system, using computational methods. The method we initially developed is limited by the design of the scenario and evaluation protocol for the iDASH Challenge; however, it can be improved by allowing the data custodian to act in a staged manner.

4.1 Introduction

Genomic data is increasingly collected by a wide array of organizations [51], ranging from direct-to-consumer genomics companies [52] to clinical institutions [276, 53]. This data serves as the basis of discovery-driven research [277, 54] and, more recently, for personalized medicine programs [231, 55]. However, as the quantity and coverage of genomic data grow, so too does the chance for the discovery and reporting of rare alleles [56, 57]. This is challenging for researchers and clinicians who aim to discern if such an allele (or combination of alleles across the genome) is meaningful with respect to an individual's phenotypic status or should influence the design of a personalized treatment regimen. To mitigate uncertainty, there is a desire to open data held by one organization to those who may need it elsewhere [58, 59]. While there are some initiatives, like the Personal Genome Project [60], that freely and publicly share genomic data linked to phenomic data, the existence of such systems does not necessarily translate into a large number of participants [278]. There are numerous reasons why individuals may not contribute their data to such programs, one of which is a privacy concern that the data would be misused or abused in some way [279, 280]. To mitigate privacy risks, data custodians have turned towards sharing summary level data about the pool of individuals who were in a study or were treated in a clinical setting.

The practice of summary data sharing began on a large scale in the mid-2000s, with programs like the Database of Genotypes and Phenotypes (dbGaP) at the National Institutes of Health [69], which aimed to standardize and centralize genomic data, making it easier to access. Summary statistics about the allele rates were made publicly accessible over the Internet because it was assumed that the privacy risks for such data were minimal. Yet, in 2008, Homer and colleagues [30] demonstrated that an adversary could apply a statistical inference attack to discern the presence of a known individual's DNA sequence in the pool of subjects. This was specifically accomplished by measuring the distance between an individual's sequence to the allele rates exhibited by the pool versus some reference population, such as the International Haplotype Mapping Program [281] or 1000 Genomes [230]. When the target was deemed to be sufficiently biased towards the pool, the adversary could assign the hallmarks of the pool, such as membership in a specific group for case-control study (e.g., individuals positively diagnosed with a sexually transmitted disease). As an artifact of this demonstration, the NIH, Wellcome Trust, and other genomic data custodians restricted access to summary-level genomic data [282, 66]. Since the initial attack, there have been a number of advancements in pool detection methodology (e.g., [36, 79, 170, 31, 88]).

As such inference methods evolved, the Global Alliance for Genomics and Health (GA4GH) formed to facilitate the sharing of genomic and health data in a federated manner [283]. In light of the known attacks, GA4GH created the Beacon Project, which enables data custodians to respond to queries through a web service (i.e., a beacon) about the presence/absence of a specific allele [284]. For instance, a Beacon service could respond yes/no to a question like, "Do you have any genomes with nucleotide A at position 121,212,028 on chromosome 10?" When the answer was affirmative, the requesting system user would learn that the variant in question may not be unique (i.e., because it was observed in a genome collected elsewhere) and that it might be worth pursuing further investigation into its meaning (possibly with the assistance of the answering data custodian).

Though it obscures allele rates, in late 2015, Beacon services were also shown to be vulnerable to a statistical inference attack. Specifically, Shringarpure and Bustamante (SB) described the statistical theory behind the attack and illustrated how it might require no more than 5,000 responses to infer an individual's, or their relatives', membership in the pool [86].

Given the increasing adoption of Beacon, the Integrating Data for Analysis, Anonymization, and Sharing (iDASH) National Center for Biomedical Computing allocated one of the three tracks of their 2016 Genomic Privacy Protection Challenge to explicitly focus on this vulnerability. The organizers formulated the problem as, "Given a sample Beacon database, we will ask [the] participating team to develop solutions to mitigate the Bustamante attack. We will evaluate each algorithm based on the maximum number of correct queries that it can respond [to] before any individual can be re-identified by the Bustamante attack."[13] A subset of the authors of this chapter developed the winning solution to this challenge. While this chapter provides the details behind this solution, we have further extended our initial analysis to illustrate its limits as well as introduce alternative formulations of the problem to evolve the investigation into a setting closer to the real world.

The specific contributions of this chapter include:

(1) We introduce a method that simultaneously optimizes the privacy (based on the SB attack as augmented by the iDASH Challenge organizers) and the utility of the system;

(2) We generalize the genomic data sharing problem to be more representative of scenarios in which beacons will actually be deployed; and

(3) We provide a sensitivity and robustness analysis of our method under various parameterizations of the variables relied upon by the iDASH Challenge.

(4) I further proposed game theoretic models that can be used to protect Beacon services against the SB attack.

[13] iDASH Privacy and Security Workshop. http://www.humangenomeprivacy.org/2016/.

4.2 Methods

The goal of the first track of the 2016 iDASH Challenge was to mitigate an augmented version of the SB attack. The problem was how to find such a strategy for the genomic data custodian. This section begins with a description of the attack model and then models the data custodian's strategy as an optimization problem. In this setting, the attacker is defined as a malicious user launching the SB attack. The defender, by contrast, is defined as the data custodian sharing the genomic data while mitigating the SB attack.

4.2.1 iDASH Variation of the SB Attack

Given the binary genomic summary statistics of a pool of genomes (i.e., the beacon), the attacker relies upon a likelihood ratio test (LRT) to infer whether a targeted genome is in the pool or not. The null hypothesis, H_0, is that the targeted genome is not in the beacon, While the alternative hypothesis, H_1, is that the targeted genome is in the beacon. The attack model used in the iDASH Challenge is based on the SB attack, but it is amended to allow the attacker to know the alternative allele frequency (AAF) of all single nucleotide variants (SNVs) in the underlying population of the beacon. Here, AAF is the frequency at which the alternative allele occurs in a given population. The alternative allele is defined as the second most common allele in a commonly recognized global population (e.g., 1000 Genomes Project). We refer to this scenario as the augmented SB attack (ASBA).

Formally, the log-likelihood of a set of beacon responses $x = \{x_1, \cdots, x_m\}$ and a set of SNVs $d_i = \{d_{i,1}, \cdots, d_{i,m}\}$ for target i is:

$$L(d_i, x) = \sum_{j=1}^{m} d_{ij}\big(x_j \, log \, P(x_j = 1) + (1 - x_j) \, log \, P(x_j = 0)\big), \tag{4.1}$$

where d_{ij} and x_j are binary variables. Specifically, $d_{ij} = 1$ when SNV j of target i has at least one alternative allele and $d_{ij} = 0$ when SNV j of target i has no alternative alleles. Additionally, $x_j = 1$ when SNV j has at least one alternative allele in the beacon; $x_j = 0$ if SNV j has no alternative alleles in the beacon.

The null hypothesis can be formulated as:

$$L_{H_0}(d_i, x) = \sum_{j=1}^{m} d_{ij}\big(x_j \, log\big(1 - D_n^j\big) + (1 - x_j) \, log\big(D_n^j\big)\big). \tag{4.2}$$

And the alternative hypothesis can be formulated as:

$$L_{H_1}(d_i, x) = \sum_{j=1}^{m} d_{ij}\left(x_j\, log\left(1 - \delta D_{n-1}^{j}\right) + \left(1 - x_j\right) log\left(\delta D_{n-1}^{j}\right)\right), \tag{4.3}$$

where δ is the sequencing error rate (which is usually set to 10^{-6}), D_n^j represents the probability that none of the n genomes in the beacon have an alternative allele at SNV j:

$$D_n^j = \left(1 - f_j\right)^{2n} \tag{4.4}$$

Here, f_j is the AAF of SNV j in the population.

The LRT statistic for target i can thus be stated as:

$$\begin{aligned} \Lambda(d_i, x) &= L_{H_0}(d_i, x) - L_{H_1}(d_i, x) \\ &= \sum_{j=1}^{m} d_{ij}\left(x_j\, log\frac{1 - D_n^j}{1 - \delta D_{n-1}^{j}} + \left(1 - x_j\right) log\frac{D_n^j}{\delta D_{n-1}^{j}}\right). \end{aligned} \tag{4.5}$$

Given this statistic, a threshold is selected, such that only targeted genomes with a test statistic below the threshold are regarded as being in the beacon. We assume that the attacker will always select the threshold according to a maximal allowable false positive rate (FPR), which is usually set to 0.05.

To illustrate the ASBA, we present an example of the entire attack and defense process of the iDASH challenge in Figure 4.1. For this example, we selected 8 SNVs from chromosome 10 and populated a pool of 100 records in a beacon repository according to their global AAFs. Now, let us say the attacker has access to a set of genomes with known identities, which we refer to as a target set. The target set can be divided into two mutually exclusive subsets: i) a set of targets that are actually in the pool and ii) a set of other targets. The attacker will query a Beacon service about whether each alternative allele is in the pool behind the beacon and make their attack decision based on all of the returned answers. If the defender in control of the beacon server answers truthfully, the attacker is likely to achieve a high detection rate. However, if the defender invokes some data protection method (as we introduce later on), then the risk will be mitigated. A flipping action of “T” and “F” for a particular SNV position represents answering the query regarding this position truthfully and untruthfully, respectively. In this example, as shown in the bottom right corner of Figure 4.1, it can be seen that the risk has been mitigated substantially.

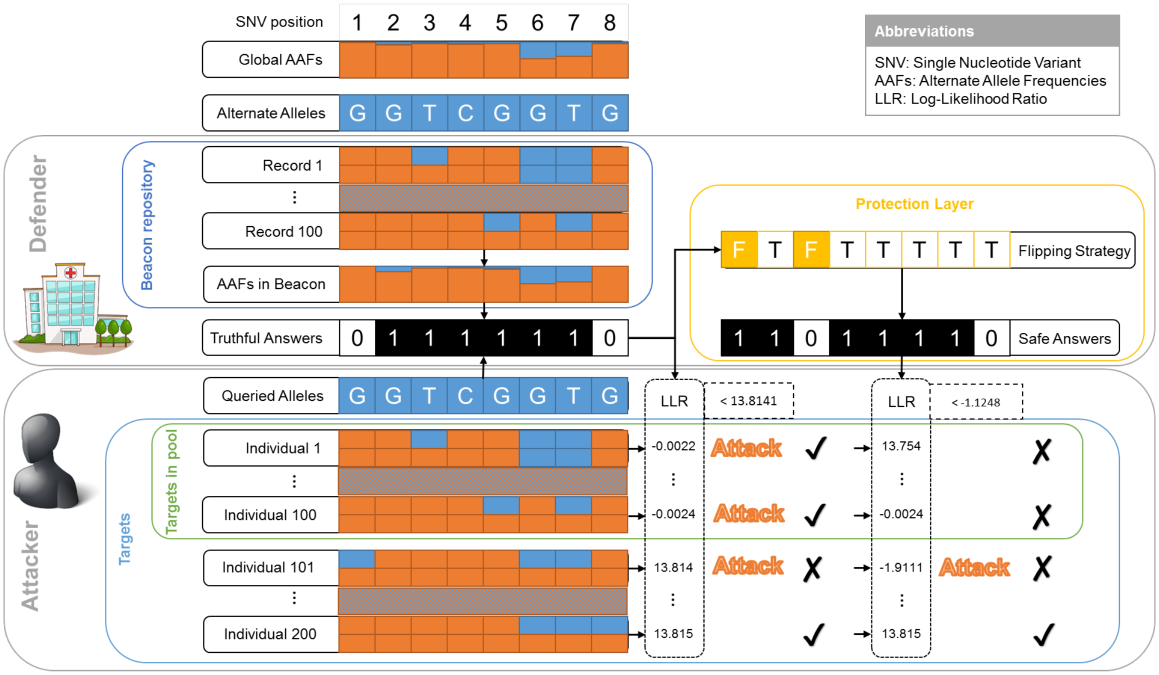


Figure 4.1 An illustration of the ASBA attack and defense process of the iDASH challenge.

4.2.2 Optimization Problem

4.2.2.1 Strategies Available to the Defender

The above problem description implies that the only action the defender can take is to lie in their answers to the attacker's queries.

In practice, the answer to a query from the attacker for a particular allele in the beacon can be 1, 0 and *null*, the latter of which means that the answer to the query is not applicable (e.g., when the defender does not have any records that cover the SNV of interest to the attacker). Henceforth, for simplicity, and alignment with the data analyzed in the iDASH Challenge, we assume that all data accessible through the beacon is a single nucleotide variant (SNV). In the attack model, the 0 and *null* answers can be treated differently. The contributions from the SNV j to the final LRT statistics, according to Equation 4.5, for answers 1, 0 and *null* are $log\,\frac{1-D_n^j}{1-\delta D_{n-1}^j}$, $log\,\frac{D_n^j}{\delta D_{n-1}^j}$, and 0, respectively. Lying about the answer to a query means two alternatives: 1) flipping or 2) masking. We define flipping as changing the answer from 1 → 0 or 0 → 1. We define masking as changing the answer from 1 → *null*, or from 0 → null. It should be recognized that we only consider the former type of lies in our experiment, as what the iDASH challenge set for simplicity. In other words, we choose to flip each SNV or not in our experiment. As an extension, I consider both types of lies in additionally proposed game models. Now, let us say that $\mathrm{S}_d = \{\mathrm{s}_d\}$ is the set of the strategies available to the defender, where each strategy represents a set of SNVs. Then, the number of all available strategies is $|S_d| = 2^m$.

4.2.2.2 Query Sequences Available to the Attacker

The effectiveness of the defender's strategy is influenced by the attacker's query sequence. We further assume the attacker has a pre-determined query sequence and has the potential to query all SNVs. Let us say that $\mathrm{S}_a = \{s_a\}$ is the set of query sequences over all of the SNVs available to the attacker. Then the number of all possible query sequences is $|S_a| = m!$.

We also assume that the only uncertain action raised by the attacker for the defender is the SNV query sequence. All of the other parameters are fixed and known to the defender before he chooses the best strategy.

4.2.2.3 Objective Function

Given this formulation, the iDASH Challenge scenario can be modeled as an optimization problem for the defender. Specifically, we wish to find a set of SNVs to flip that maximizes the utility and the privacy of the data simultaneously.

The effectiveness of a defender's strategy, $Y(s_d, s_a)$, considering both the utility and the privacy, is a function of both the defender's strategies and the attacker's query sequences, which is defined below. This creates a dependency on the definitions of the utility and privacy measures, the manner by which utility and privacy are combined, and the attack model.

As the attacker proceeds through the ordered set of SNVs, he runs a hypothesis test based on the responses for the subset of SNVs queried so far. Now, we assume the defender does not know the ground truth of the attacker's query sequence. As a consequence, the defender's best strategy is the one that maximizes his or her own expected effectiveness:

$$s^*{}_d = \underset{s_d}{\mathrm{argmax}}\, E\big(Y(s_d)\big) = \underset{s_d}{\mathrm{argmax}} \frac{1}{|S_a|} \sum_{s_a \in S_a} Y(s_d, s_a) \tag{4.6}$$

This implies that the attacker will choose any of the query sequences with equal probability. So long as this effectiveness function is defined and known to the defender, the simplest solution for the defender is to examine all available strategies. However, such a brute force approach is computationally challenging because the size of the defender's strategy space, $|S_d|$, increases exponentially with the number of SNVs.

Similarly, calculating the exact valuation of a defender's strategy, $E\big(Y(s_d)\big)$, is difficult because of the large number of query sequences available to the attacker, $|S_a|$. However, the valuation of a defender's strategy can be estimated based on a subset of the attacker's strategies, $\mathrm{S}_a{}'$, under a random selection model.

4.2.2.4. Evaluation Criteria

There are many alternative definitions of effectiveness that could be invoked to evaluate the defender's strategies. Here, we first introduce the approach that was relied upon in the iDASH Challenge and then consider several alternatives. As will become evident, this is important because it influences how the defender will search for their best strategy.

In the description of the iDASH Challenge task, the effectiveness of a strategy was defined as the number of answers that can be served correctly before any targeted individual's presence is successfully detected. However, no matter what the defender's strategy may be, the detection power of the LLR test is always larger than 0 for the first several SNVs unless queries for these SNVs are not answered truthfully. This implies that at least one individual will be successfully detected if the first several SNVs are answered truthfully, which is highly likely because we assume the defender does not know the position of a specific query in the entire query sequence.

However, such a definition dictates towards a worst-case privacy scenario. Specifically, it assumes that a system is considered vulnerable if any record can be breached. As noted earlier, there are alternative definitions that could be applied. For instance, a more pragmatic definition of the effectiveness of a protection model may be the proportion of correct answers that are returned before the presence of a certain number of individuals is detected. Under this formulation, when a method is evaluated, the iDASH Challenge organizers stated that 60%, instead of 0%, would be applied as the threshold for the detection power.

To define several alternative evaluation criteria, let us limit the utility and privacy measures in the [0, 1] range. For the purposes of the iDASH Challenge, the utility can be regarded as the proportion of queries that are answered truthfully. By contrast, privacy is defined as a binary variable that is 1 if, and only if, a certain portion of the targeted individuals are never detected in the beacon, and 0 otherwise. Now, there are numerous ways to combine the utility and privacy measures into an effectiveness measure. In the iDASH Challenge, the effectiveness of the defense was defined as the utility for the proportion of SNVs shared before a certain portion of targeted individuals are re-identified.

Alternatively, the utility can be defined as a weighted sum of correct answers. In this scenario, each SNV can be weighted according to its importance (e.g., correlation with some phenotype). On the other hand, privacy can be defined as the expected false-negative rate when the number of used SNVs is uncertain. The effectiveness of the defense can thus be defined as a weighted sum of the utility and privacy.

4.2.3 Protection Method

In this subsection, we start with a description of the solution we submitted to the iDASH Challenge. To perform a comprehensive empirical analysis, we then provide a description of alternative methods that could be applied to this problem. At last, we describe a game theoretic model that can be applied to Beacon Services.

4.2.3.1 Our iDASH Submission

The solution we submitted to the iDASH Challenge entails searching through a collection of possible strategies that the defender can invoke to protect the system. Each of these strategies utilizes the same method, in the form of flipping some SNV query answers. In this section, we illustrate the principles by which such answers are flipped and how the strategy space is prioritized and searched.

4.2.3.1.1 Flipping responses

In the iDASH Challenge, it was assumed that the defender is not aware of the attacker's query sequence *a priori* and does not keep track of the queries. As a result, we need to find a defender's strategy that is independent of the attacker's query sequence. Since utility is basically measured as the number of queries that the defender responds to truthfully, we introduce the notion of discriminative power for each SNV in the pool. The *discriminative power* represents an SNV's ability to distinguish the records in the pool of individuals behind the beacon from a reference dataset. We define a *differential discriminative power* for each SNV in the pool, which represents the difference between its discriminative power before and after a flip. The top k percent of the SNVs in the pool, ranked by their differential discriminative power, will have their query responses flipped.

Here, we take a moment to define discriminative power formally. We assume that the defender knows the attacker's target set (because, otherwise, the defender's strategy could not be directly measured). If this is not the case, then there are various ways for the defender to estimate the target set (the details of which are deferred to below). Let us say that the targeted pool with n individuals and m SNVs is represented as a binary matrix with n rows and m columns $D_{n\times m} = \{d_{ij}\}$ in which d_{ij} is 1 when individual i in the pool has the alternative allele for SNV j and 0 otherwise. Let us further say that the targeted reference dataset with n' individuals and m SNVs (the same as the SNVs in the pool) is represented by a binary matrix with n' rows and m columns $R_{n' \times m} = \{r_{ij}\}$ in which r_{ij} is one when individual i in the reference has the alternative allele in SNV j and zero otherwise.

Given the truthful answer x_j for the query regarding SNV j, the ability for SNV j to indicate if an individual is behind the beacon, according to the pool data, is defined as the average LLR for all the individuals in the pool if only SNV j is queried:

$$A_j(x_j) = \frac{1}{n} \sum_{i=1}^{n} d_{ij} \left(L'_{H_1}(x_j) - L'_{H_0}(x_j) \right) \tag{4.7}$$

where

$$L'_{H_0}(x_j) = x_j \, log\left(1 - D_n^j\right) + \left(1 - x_j\right) log\left(D_n^j\right) \tag{4.8}$$

$$L'_{H_1}(x_j) = x_j \, log\left(1 - \delta D_{n-1}^j\right) + \left(1 - x_j\right) log\left(\delta D_{n-1}^j\right) \tag{4.9}$$

Similarly, the ability for an SNV to indicate if an individual is behind the beacon, according to the reference dataset, is defined as the average LLR for all the individuals in the reference if only SNV j is queried:

$$A'_j(x_j) = \frac{1}{n'} \sum_{i=1}^{n'} r_{ij} \left(L'_{H_1}(x_j) - L'_{H_0}(x_j) \right) \tag{4.10}$$

The more similar these two values, the less powerful the LLR test will be. Based on this formulation, the discriminative power for SNV j becomes:

$$D_j(x_j) = A_j(x_j) - A'_j(x_j) = \left(\frac{1}{n} \sum_{i=1}^{n} d_{ij} - \frac{1}{n'} \sum_{i=1}^{n'} r_{ij} \right) \left(L'_{H_1}(x_j) - L'_{H_0}(x_j) \right) \tag{4.11}$$

The difference of the discriminative power before and after flipping the SNV is:

$$\Delta D_j(x_j) = D_j(x_j) - D_j(1 - x_j) \tag{4.12}$$

As a result, the first step of our Strategic Flipping method will flip the top k percent of the SNVs sorted according to the differential discriminative power. Note that, when a set of SNVs have the same differential discriminative power, the SNVs with the highest discriminative powers are selected first. When a set of SNVs have the same discriminative power, the SNVs with the lowest AAFs are selected first. A random selection is applied to break AAF ties.

4.2.3.1.2 Searching the strategy space

We use a greedy algorithm to search the defender's strategy space for a local optimum. To do so, we begin by randomly selecting q query sequences. Next, we traverse the l nearest neighbors of the current best strategy and find the neighbor with the best average measure of effectiveness, which is averaged across the q query sequences. Two strategies are regarded as neighbors if they only have one answer that is different. The distance between two neighbors is calculated as the absolute difference of the average number of answers provided truthfully by the two strategies and the rank of the different SNVs of these two strategies, in descending order, sorted by the differential discriminative power. In other words, if the number of answers provided truthfully by these two strategies is t and t' and the rank of the different SNV is τ, then the distance between two neighbors is $|\tau - (t + t')/2|$. When l equals to 2, only the top SNVs, in terms of differential discriminative power, are flipped.

We start from the result of the aforementioned Top-K Flipping step and keep searching until no strategy with better effectiveness can be found. In the case where two measures need to be optimized simultaneously (such as utility and privacy measures), we search the neighborhoods for a Pareto-optimal strategy. In a multiple-objective optimization problem, a strategy is Pareto optimal if no other strategies dominate it (i.e., is better than it in terms of both measures in consideration).

4.2.3.2 Alternative Methods

We selected five alternative methods beyond the one we proposed above. The first three are baseline methods that set the upper and lower bounds on the measures. The last two are state-of-the-art methods in the literature that address the beacon detection problem. For reference purposes, we name our iDASH solution as the Strategic Flipping Method or M_{SF}.

4.2.3.2.1 Truthful Method (marked as M_T)

The defender simply responds to all queries truthfully. This sets the lower bound for the privacy measure and the upper bound for the utility measure.

4.2.3.2.2 Baseline Method (marked as M_B)

The defender flips the k percent of the SNVs with the lowest AAF in the underlying population from which the pool is sampled. This method was used by the organizers of the iDASH Challenge to establish a lower bound for the effectiveness measure.

4.2.3.2.3 Greedy Accountable Method (marked as M_{GA})

We assume the users' queries are accountable, whereby each user has an account such that the defender documents the attacker's queries and results as they are processed. Upon submission of the next

SNV query, the defender will flip the answer for this SNV if, and only if, the power of the resulting LLR test would be smaller than when no flip is applied.

4.2.3.2.4 Random Flips (marked as M_{RF})

This method was recently proposed by Raisaro et al. [159]. In this method, the defender flips ε portion of SNVs that exhibit unique alleles in the beacon.

4.2.3.2.5 Query Budget (marked as M_{QB})

This method was also proposed by Raisaro et al. [159]. In this method, a privacy budget is assigned to each individual in the beacon. Each time a record contains the queried allele, the budget for that user is reduced by a certain amount. Once a record's budget is exhausted, their genome will no longer contribute to responses provided by the beacon. Similar to M_R, this method requires the users' queries to be documented.

4.2.3.3 Game Theoretic Models

4.2.3.3.1 Rational attack model

With this attack model corresponding to Equation 4.5 in mind, the data holder can find the best protection strategy by solving an optimization problem. However, a rational attacker will do better by making decisions based on the payoff function instead of adopting a fixed strategy. If we use a game model similar to the one mentioned in the last chapter, according to Equation A.50, the attacker's best strategy is as follows:

$$a_i{}^*(x) = \begin{cases} 1, & \Lambda(d_i, x) > \log(c) - \log(b \cdot p_i) = \theta_i, \\ 0, & \Lambda(d_i, x) \leq \log(c) - \log(b \cdot p_i) = \theta_i, \end{cases} \quad \forall i \in I \tag{4.13}$$

where c is the cost per attack, b is the benefit per successful attack, and p_i is the prior probability that target i is in the dataset. In other words, the attacker maximizes his payoff so long as the threshold θ_i is set accordingly. Note that the thresholds for different targets are not necessarily the same.

4.2.3.3.2 Additional strategies available to the defender

In the context of the iDASH challenge, the only choice the defender can select is to answer "Yes" or "No" to the attacker's query. Since it is assumed that the attacker's queries are not accountable (i.e., not in a registered system), the defender needs to select a strategy beforehand, which makes the game a one-shot game. Thus, the number of actions available to the defender is 2^m. In each strategy, for the alternative allele in each position, the defender could choose to answer truthfully or not. If we define the strategy for each SNV as choosing one of these two functions both from the set {“Yes”, “No”} (or $\{1, 0\}$) to the set itself, then the number of strategies available to the defender is also 2^m. The first function is disclosing (or D), such that $D(0) = 0$, $D(1) = 1$, and the second function is flipping (or F), such that $F(0) = 1$, $F(1) = 0$.

In a more practical scenario, the defender has a third choice to select for the attacker’s query, which is to answer “Not Applicable” (i.e., “NA” or “Null”). This happens, for instance, when the defender does not have any records that cover the SNV of interest. In this case, the number of available actions to the defender increases to 3^m. If we define the strategy for each SNV as choosing one of these three functions all from the set {“Yes”, “No”} (or $\{1, 0\}$) to the set {“Yes”, “NA”, “No”} (or $\{1, 0.5, 0\}$), then the number of available strategies to the defender is also 3^m. The first function is disclosing (or D), such that $D(0) = 0$, $D(1) = 1$, and the second function is flipping (or F), such that $F(0) = 1$, $F(1) = 0$ and the third function is masking (or M), such that $M(0) = 0.5$, $M(1) = 0.5$.

Note that the “No” and “NA” answers affect the attack model differently. The additive contributions from the SNV j to the final LRT statistics, according to the equation, for answers “Yes”, “No”, and “NA” are $log\frac{1-D_n^j}{1-\delta D_{n-1}^j}$, $log\frac{D_n^j}{\delta D_{n-1}^j}$, and 0, respectively.

In the previous section, we only considered the former case (i.e., the context of the iDASH challenge). Here we consider the latter case, which is likely to be more practical for real-world deployment. The latter case is more complicated than the former one; however, by showing how to solve a slightly different problem, we aim to illustrate how a game theoretic approach could be applied in a variety of scenarios.

To represent the protection mechanism more formally, let us say $s = \{s_1, \cdots, s_m\}$ corresponds to a strategy for all SNVs. Specifically, $s_j = 1$ if, and only if, the strategy for SNV j is disclosing its original response; $s_j = 0$ if, and only if, the strategy for SNV j is flipping to its opposite response; and $s_j = 0.5$ if, and only if, the strategy for SNV j is masking. The LRT statistic, as a function of d_i, s and x, can be represented as:

$$\Lambda(d_i, y(s,x)) = \sum\nolimits_{j\in J} d_{ij}\Lambda_j\left(y_j(s_j,x_j)\right), \forall i \in I \tag{4.14}$$

$$\Lambda_j\left(y_j(s_j,x_j)\right) = y_j(s_j,x_j)\, log\frac{1-D_n^j}{1-\delta D_{n-1}^j} + y_j(s_j, 1-x_j)\, log\frac{D_n^j}{\delta D_{n-1}^j}, \forall i \in I, j \in J, \tag{4.15}$$

$$y_j(s_j,x_j) = (x_j + s_j - 1)(2s_j - 1), \forall j \in J, \tag{4.16}$$

where $y = \{y_1, \cdots, y_m\}$ corresponds to the set of Beacon responses after applying the protection layer for all SNVs in the set J.

4.2.3.3.3 Stackelberg game model

With the added protection layer, the attacker's payoff function and best strategy are unchanged. As such, the attacker's best strategy, according to Equation 4.13, as a function of the Beacon responses after applying the protection layer, is as follows:

$$a_i{}^*(y) = \begin{cases} 1, & \Lambda(d_i, y) > \theta_i, \\ 0, & \Lambda(d_i, y) \leq \theta_i, \end{cases} \quad \forall i \in I. \tag{4.17}$$

If we use a Stackelberg game model similar to the one mentioned in the last chapter, then the defender's simplified payoff function is as follows:

$$Y(g, a) = w_g \cdot \sum\nolimits_{j \in J} g_j - w_a \cdot \sum\nolimits_{i \in I} a_i \tag{4.18}$$

where J is a set of SNVs and I is a set of targeted individuals. w_g corresponds to the worth of the correct presence information of each SNV available to the defender, w_a corresponds to the worth of each protected individual to the defender. Specifically, g_j is a binary variable determined by the defender's strategy: $g_j = 1$ if, and only if, SNV j is truthfully disclosed; and a_i is a binary variable representing the attacker's strategy: $a_i = 1$ if, and only if, target i is attacked. Clearly, for each SNV j, g_j is a function of s_j: $g_j = 1$ if, and only if, $s_j = 1$.

$$g_j = (2s_j - 1)s_j, \forall j \in J \tag{4.19}$$

Note that this is just an example of a payoff function. For simplicity, we assume each SNV has the same amount of value to the defender, and each target has the same amount of value to the attacker. We also assume the flipping protection mechanism and masking protection mechanism have the same influence on the defender's utility. In this case, the defender makes the decision according to the monetary incentive. The payoff function could be further relaxed into a weighted sum of the utility and privacy risk measures, $U(g)$ and $R(a)$, as functions of the defender's strategy and the attacker's strategy respectively:

$$U(g) = \frac{1}{m} \sum\nolimits_{j \in J} g_j \tag{4.20}$$

$$R(a) = \frac{1}{n} \sum\nolimits_{i \in I} a_i \tag{4.21}$$

$$Y(g, a) = W_g \cdot U(g) - W_a \cdot R(a) \tag{4.22}$$

where W_g is the weight of the utility measure, and W_a is the weight of the privacy risk measure. It will be the same as the previous payoff function if $W_g = w_g m$ and $W_a = w_a n$.

In summary, the optimization problem can be written as:

$$s^* = \operatorname*{argmax}_s W_g \cdot U(g(s)) - W_a \cdot R(a^*)$$

$$s.t. \quad a_i{}^*(s,x) = \begin{cases} 1, & \Lambda(d_i, y(s,x)) > \theta_i, \\ 0, & \Lambda(d_i, y(s,x)) \le \theta_i, \end{cases} \quad \forall i \in I \tag{4.23}$$
$$\theta_i = \log(c) - \log(b \cdot p_i), \forall i \in I$$

A typical payoff to the defender is dependent upon both the defender's and the attacker's strategies. It also depends upon both the utility and privacy measures of the defender's strategy to be effective. But it does not have to be a linear combination of the utility and privacy measures. In the iDASH Challenge, the effectiveness of a strategy was defined as the number of answers that can be correctly served before a certain portion ($r = 0.6$) of targeted individuals' presence is successfully detected. In this case, the optimization problem can be written as:

$$s^* = \underset{s}{\operatorname{argmax}}\, j^*$$
$$s.t. \quad \frac{1}{n}\sum\nolimits_{i \in I} a_{ij}{}^*(s,x) \le r, \forall j \le j^*$$
$$a_{ij}{}^*(s,x) = \begin{cases} 1, & \displaystyle\sum_{j=1}^{k} d_{ij}\Lambda_j\left(y_j(s_j, x_j)\right) > \theta_i, \\ 0, & \displaystyle\sum_{j=1}^{k} d_{ij}\Lambda_j\left(y_j(s_j, x_j)\right) \le \theta_i, \end{cases} \quad \forall i \in I, k \in J \tag{4.24}$$
$$\Lambda_j\left(y_j(s_j, x_j)\right) = y_j(s_j, x_j)\, log \frac{1 - D_n^j}{1 - \delta D_{n-1}^j} + y_j(s_j, 1 - x_j)\, log \frac{D_n^j}{\delta D_{n-1}^j}, \forall i \in I, j \in J$$
$$\theta_i = \log(c) - \log(b \cdot p_i), \forall i \in I$$

These are not optimization problems with direct analytical solutions. However, it is straightforward to solve these using empirical methods. Basically, to do so, one can initiate a search from a strategy and iteratively uncover better strategies from the neighborhood of the current strategy until no better strategy can be found. Moreover, there are some principles one can exploit to accelerate the search process. For instance, since the attacker does not make the decision according to the false positive rate, the measurement on SNVs for discriminative power, based on the reference population, as described in Equation 4.11 is no longer useful. To find a good starting point, we can flip or mask the top k SNVs with the greatest difference in detection power.

We define the detection power of SNV j as the additional LRT statistic contributed by SNV j. Since we assume the flipping and masking mechanisms have the same influence on the defender's utility, the mechanism with lower detection power will be preferred for SNV j.

$$\Lambda_j\left(y_j(0, x_j)\right) = y_j(0, x_j)\, log \frac{1 - D_n^j}{1 - \delta D_{n-1}^j} + y_j(0, 1 - x_j)\, log \frac{D_n^j}{\delta D_{n-1}^j}, \forall j \in J \tag{4.25}$$

$$\Lambda_j\left(y_j(0.5, x_j)\right) = y_j(0.5, x_j)\, log\frac{1 - D_n^j}{1 - \delta D_{n-1}^j} + y_j(0.5{,}1 - x_j)\, log\frac{D_n^j}{\delta D_{n-1}^j}, \forall j \in J \tag{4.26}$$

Given the original response x_j, the preferred mechanism s_j is determined (i.e., $s_j = 0$ or $s_j = 1$). In the case where SNV j has at least one alternative allele in the beacon (i.e., $x_j = 1$):

$$\Lambda_j\left(y_j(0{,}1)\right) = log\frac{D_n^j}{\delta D_{n-1}^j}, \forall j \in J \tag{4.27}$$

$$\Lambda_j\left(y_j(0.5{,}1)\right) = 0, \forall j \in J \tag{4.28}$$

The flipping mechanism is preferred if, and only if, $log\, D_n^j < log\, \delta D_{n-1}^j$ (or $f_j > 1 - \sqrt{\delta}$).

In the case where SNV j has no alternative alleles in the beacon (i.e., $x_j = 0$):

$$\Lambda_j\left(y_j(0{,}0)\right) = log\frac{1 - D_n^j}{1 - \delta D_{n-1}^j}, \forall j \in J \tag{4.29}$$

$$\Lambda_j\left(y_j(0.5{,}0)\right) = 0, \forall j \in J \tag{4.30}$$

Here, the flipping mechanism is preferred if, and only if, $log\left(1 - D_n^j\right) < log\left(1 - \delta D_{n-1}^j\right)$ (or $f_j < 1 - \sqrt{\delta}$). When no protection mechanism is applied, we have:

$$\Lambda_j\left(y_j(1{,}1)\right) = log\frac{1 - D_n^j}{1 - \delta D_{n-1}^j}, \forall j \in J \tag{4.31}$$

$$\Lambda_j\left(y_j(1{,}0)\right) = log\frac{D_n^j}{\delta D_{n-1}^j}, \forall j \in J \tag{4.32}$$

We define the detection power of SNV j as:

$$\mathrm{A}_j\left(y_j(s_j, x_j)\right) = \sum\nolimits_{i \in I} d_{ij} L_j\left(y_j(s_j, x_j)\right), \forall j \in J$$

We define the difference of detection power as the maximal difference of LRT statistics between a truthful answer and an untruthful answer for each SNV:

$$D_j(x_j) = \max\left(\mathrm{A}_j\left(y_j(0, x_j)\right) - \mathrm{A}_j\left(y_j(1, x_j)\right), \mathrm{A}_j\left(y_j(0, x_j)\right) - \mathrm{A}_j\left(y_j(0.5, x_j)\right)\right), \forall j \in J \tag{4.33}$$

In the case of $x_j = 1$ and $f_j \leq 1 - \sqrt{\delta}$, this can be simplified to be:

$$D_j = log\frac{1 - D_n^j}{1 - \delta D_{n-1}^j}, \quad \forall j \in J \tag{4.34}$$

This function decreases monotonically when f_j, the AAF of SNV j in the population, increases. As a result, SNVs with several lowest AAFs are masked in the starting search point. The defender does not need to prepare strategies for the case of $x_j = 0$ because the targeted individuals are not in the dataset when the answer is "No", and by flipping or masking this response, the privacy measure will remain constant while the utility will decrease. The defender does not need to prepare strategies for the case of $f_j > 1 - \sqrt{\delta}$, because when δ is a very small number, this case will not exist.

4.2.3.3.4 Stochastic game model

Here, I improve the game model by modeling the uncertainty of adversary's behavior in the SB Attack and time factors explicitly using a stochastic game model with Markov decision process (MDP), which is more precise and complex.

An MDP can model sequential decision making where there is a reward associated with taking an action at each state, while the outcome of the action can be random [285, 286]. MDP has been used in robotics, automated control, economics, and manufacturing. For example, Letchford and Vorobeychik [287] use MDP to model the adversary's optimal planning in security problems. Xia et al. [288] first apply MDP in the privacy-preserving data publication setting, particularly, in the scenario of record-linkage attack. In their model, the adversary makes a decision whether or not to contact an individual that is linked to a record in the de-identified dataset at each step.

A stochastic game is a dynamic game with probabilistic transitions played in a sequence of stages [289]. A stochastic game can be regarded as an MDP with multiple decision-makers [290]. To the best of my knowledge, no one has published any research on using stochastic games to model both data publisher and adversary's behavior in privacy-preserving data sharing settings, especially for cases of sharing genomic data. In [201], Wang et al. model the strategic and dynamic competition between a smartphone user and a malicious adversary as a zero-sum stochastic game. The user controls the location data granularity while the adversary selects which sensing data as the source. In [203], Shokri et al. formulate the location privacy problem as Stackelberg Bayesian games with the consideration of a user's service quality and adversary's cost. However, these location privacy problems are quite different from the privacy-preserving data sharing discussed in this chapter.

I use a two-player stochastic game to model the data protection system. The state of the system changes whenever the adversary's inference or re-identification is successful. I adapt a Markov model to capture the transitions between states. The data publisher is assumed to know the Markov chain of the adversary, and he is able to add some masks or noise to the dataset to be published. My goal is to find the optimal protection strategy for the data publisher based on the game theoretic analysis. Both players in the game are assumed to be rational and driven by monetary incentives for simplicity.

A two-player stochastic game G consists of six-tuple. $\langle S, A^1, A^2, r^1, r^2, P \rangle$, in which S is the discrete state space, A^k is the action space of player k ($k = 1{,}2$). $r^k: S \times A^1 \times A^2 \to \mathbb{R}$ is the stage payoff function for player k. $P: S \times A^1 \times A^2 \to \Delta(S)$ is the transition probability map, where $\Delta(S)$ is the set of probability distributions over S. The game G is played in a sequence of stages, where each player k receives a stage payoff $r^k(s, a^1, a^2)$ based on players' actions $a^k \in A^k$ and current state $s \in S$. Each player k attempts to maximize its expected sum of discounted payoffs.

At each time t, the player can take action. The attack results observed at time t, are denoted by S^t. $S^t = 1$ means the attack is successful, $S^t = 0$ otherwise.

The actions of the data publisher at time t are defined as $a_p^t = \{a_{p,1}^t, a_{p,2}^t, \dots, a_{p,M}^t\}$ in which $a_{p,m}^t \in \{0,0.5,1\}$ is the data publisher's action for attribute m at time t. $a_{p,m}^t = 0$ represents flipping the answer to its opposite response, $a_{p,m}^t = 0.5$ represents masking the answer, and $a_{p,m}^t = 1$ represents publishing original response. $m = 1, \cdots, M$ represents each attribute, in which M is the number of attributes.

The inference actions of the adversary at time t are defined as $a_a^t = \{a_{a,1}^t, a_{a,2}^t, \dots, a_{a,N}^t\}$ in which $a_{a,n}^t \in [0,1]$ is the adversary's inference action for target n at time t. $a_{a,m}^t = 1$ represents the target is inferred to be in the dataset/cohort (or have a sensitive attribute), and $a_{a,n}^t = 0$ represents the target is not inferred to be in the dataset/cohort (or have a sensitive attribute. Notice that this action space is for a membership-inference attack (or an attribute-disclosure attack). In a re-identification attack (i.e., an identity-disclosure attack), the action space will include all the identities in the linkable dataset with identities: $a_{a,n}^t \in \{1, \cdots, N'\}$. N' is the number of identities in the linkable dataset. $n = 1, \cdots, N$ represents each attribute, in which N is the number of targets. In addition, the query action of the adversary at time t is $a_q^t \in \{0, \cdots, M\}$, and the intrusion action of the adversary at time t is $a_i^t \in \{0, \cdots, N\}$. An intrusion action means executing the attack (e.g., contacting the target) and learning the attack result (i.e., successful or not), which brings cost and benefit to the adversary and changes the system's state.

The reward function of the data publisher is dependent upon the quality of the published data and the privacy loss which are functions of players' actions and system states, which can be written as

$$r_p(s^t, a_p^t, a_a^t) = U(a_p^t) - L(s^t) \tag{4.35}$$

where $U(a_p^t)$ is the utility of the published data for the data publisher which depends on the quality of the data, and $L(s^t)$ is the penalty to the publisher because of the privacy loss.

The reward function of the adversary is dependent upon the privacy loss and cost for each action, which can be written as:

$$r_a(s^t, a_p^t, a_a^t) = L(s^t) - C(a_a^t) \tag{4.36}$$

where $L(s^t)$ is what the adversary gains from a successful attack, and $C(a_a^t)$ is the cost for each action which includes the expected cost of penalty if the adversary's attack is detected.

Therefore, the data publisher's payoff is the expected sum of discounted stage payoffs as follows:

$$U_p = E\left(\sum_{t=0}^{\infty} \gamma^t r_p(s^t, a_p^t, a_a^t)\right) \tag{4.37}$$

where γ is the discount factor. We first look for stationary policies. The publisher's strategy is denoted by $\pi_p: S \rightarrow \Delta(A_p)$, and the adversary's strategy is denoted by $\pi_a: S \rightarrow \Delta(A_a)$, where S is the state space for S^t, $\Delta(A_p)$ and $\Delta(A_a)$ are the probability distributions over the publisher's action space A_p and the adversary's action space A_a, respectively.

Here, the initial state is defined to be the state at time $t = 0$, denoted by S^0. Given policies π_p, π_a and a state $s \in S$, the publisher's utility can be written as

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E\left(r_p\left(s^t, a_p^t, a_a^t\right) | \pi_p, \pi_a, S^0 = s\right) \tag{4.38}$$

Denote the actions a_p^t, a_a^t determined by policies π_p, π_a to be a_p^{π}, a_a^{π}, respectively. Then we can have

$$V^{\pi}(s) = r_p\left(s, a_p^{\pi}, a_a^{\pi}\right) + \gamma \sum_{s' \neq s} Pr\left(s' | s, a_p^{\pi}, a_a^{\pi}\right) V^{\pi}(s') \tag{4.39}$$

Both publisher and adversary follow their optimal policies $\{\pi_p^*, \pi_a^*\}$ that maximize their own utilities, called optimal strategy profile.

We define the Nash equilibrium in a stochastic game as the optimal strategy profile $\pi^* = \{\pi_p^*, \pi_a^*\}$, such that for all state $s \in S$:

$$V^{\pi^*}(s) \geq V^{\{\pi_p, \pi_a^*\}}(s), \forall \pi_p \tag{4.40}$$

and

$$V^{\pi^*}(s) \geq V^{\{\pi_p^*, \pi_a\}}(s), \forall \pi_a \tag{4.41}$$

The optimal strategy profile can be derived via existing reinforcement learning algorithms such as minimax-Q learning [291] or using two-level linear programming [288].

It is anticipated that using a game theoretic approach will lead to a better payoff, which represents a better balance between the data utility and privacy risk. In the method that implicitly uses the game theoretic approach, the privacy risk was overestimated by not considering the attacker's costs, which include the cost of access and the cost of penalty. However, introducing the cost of penalty to the model needs two prerequisites. First, it requires each user to sign a data use agreement before querying a Beacon service. Second, it requires that the malicious users can be detected as attempting to re-identify records and could be pursued as violators of a contract and penalized for liquidated damages. Neither of these two perquisites is an easy task. As pointed out by Craig et al. [170], the prior probability that a targeted individual is actually in the pool is likely to be much smaller than 50%, as was the case for the iDASH challenge. By not considering the prior probability, the privacy risk was overestimated.

4.3 Results

In this section, we introduce the evaluation measures, the experimental setup, and the results of the empirical analysis.

4.3.1 Evaluation Measures

In addition to the measure of effectiveness provided by the organizers of the iDASH Challenge, we propose several alternatives to assess the effectiveness of each method in a more comprehensive manner.

4.3.1.1 Utility

We measure utility according to the proportion of queries responded to truthfully (U).

4.3.1.2 Privacy

We measure privacy according to two different criteria, which we refer to as P_1 and P_2. First, we measure P_1 as a binary variable that is set to 1 if, and only if, a certain portion of the targeted individuals were never detected in the beacon, and 0 otherwise. We select 60% as the threshold because this is the definition used in the iDASH Challenge. It should be noted that our method generalizes to any threshold, but the results we present are limited to this parameterization. Second, we measure P_2 as the expected false-negative rate when the number of SNVs to be queried is uncertain. We assume that the total number of SNVs about which the attacker has already queried, when they stop, is a random integer number uniformly distributed in the range [0, m].

4.3.1.3 Effectiveness

The effectiveness, which considers both utility and privacy, is measured according to two criteria. First, we measure E_1 as the proportion (in terms of all SNVs) of queries that are responded to truthfully before 60% of the individuals are successfully detected. Second, we measure E_2 as the proportion of truthful answers plus the expected false-negative rate.

4.3.1.4 Computational Efficiency

Computational efficiency is an important factor to be considered for the deployment of solutions in a working system of beacons. As such, we also present the running time for each method. To measure the running time, we use a machine with Intel Core i7 quad-core 3.00GHz CPU and 8 GB memory.

4.3.2 Experimental Design

To evaluate the effectiveness of the methods, we created a pool based on the first 400,000 SNVs in Chromosome 10. The pool is composed of 250 individuals randomly selected from the 2,504 individuals in Phase 3 of the 1000 Genomes Project [230]. The reference includes 250 individuals randomly selected from the remaining individuals in Phase 3 of the project. Note that the higher the association between the allele frequencies in the pool and the allele frequencies estimated by the attacker (as demonstrated by a sensitivity analysis), the higher the detection power. In our experiments, we assume a scenario where the allele frequency estimates available to the attacker are the allele frequencies in the entire population of 1000 Genomes. All of the other parameters are set, as shown in Table 4.1. The maximal allowable false positive rate (α) is set to 0.05, and the sequencing error rate (δ or the mismatch rate) is set to 0.000001, according to Shringarpure and Bustamante [86]. The number of sampled query sequences is set to 10, and the number of examined neighbors is set to 2 for simplicity. The percentage of flipped answers is set to 5, but this setting will be discussed further in the section of sensitivity analysis. The noise level in the Random Flip method is set to 0.75, and the maximal allowable power in the Query Budget method is set to 0.9, according to Raisaro et al. [159].

Table 4.1 Parameter settings for the experiments.

Parameter	**Notation**	**Setting**
Size of pool	n	250
Number of SNVs	m	400,000
Maximal allowable false positive rate	α	0.05
Sequencing error rate	δ	0.000001
Number of sampled query sequences	q	5
Number of examined neighbors	l	2
Percentage of flipped answers	k	5
Noise level in the Random Flip method	ε	0.75
Maximal allowable power in the query budget method	β	0.9

4.3.3 Findings

We compare all methods using the average results across ten randomly generated query sequences. Figure 4.2 shows how the detection power and the proportion of flipped answers (i.e., lies to the attacker) every 1000 queries change as a function of the number of queried SNVs for one of the query sequences.

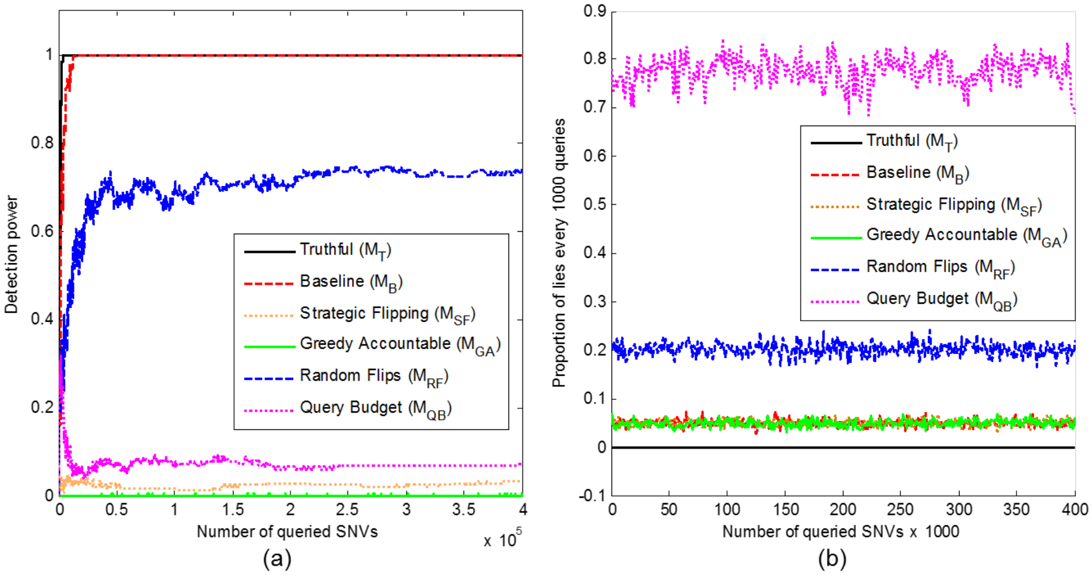


Figure 4.2 How the number of SNVs queried influences the (a) detection power and (b) proportion of lies.

By inspecting Figure 4.2(a) and Figure 4.2(b), it can be seen that the detection power does not increase monotonically when the defender is flipping answers. From Figure 4.2(a), it can be seen that the three methods with the lowest (on average) detection power are M_{GA}, M_{SF}, and M_{QB}. Notably, none of these methods exceed the threshold of 60%. From Figure 4.2(b), it can be seen that the three methods with the fewest (on average) induced lies are M_{GA}, M_B, and M_T. Considering the intersection of these results, it appears that M_{RF} and M_{GA} are likely the best options.

Inspecting the result of only one query sequence provides some intuition into the trends of the utility and privacy measures, but it may be biased by a single run of the experiment. Table 4.2 summarizes the mean and +/-1 standard deviation for each of the performance measures across 10 query sequences.

Table 4.2 The performance of the genomic data protection methods across 10 runs.

Method	Statistic	Utility	Privacy		Effectiveness		Runtime (seconds)
		U	P_1	P_2	E_1	E_2	
M_T	Avg. Std.	1 ±0.0000	0 ±0.0000	0.0025 ±0.0003	0.0026 ±0.0004	1.0025 ±0.0003	51.3525 ±0.9354
M_B	Avg. Std.	0.95 ±0.0000	0 ±0.0000	0.0072 ±0.0021	0.0054 ±0.0017	0.9572 ±0.0021	51.584 ±0.5016
M_{SF}	Avg. Std.	0.95 ±0.0000	1 ±0.0000	0.9729 ±0.0042	0.95 ±0.0000	1.9229 ±0.0042	76.4638 ±0.9588
M_{GA}	Avg. Std.	0.9512 ±0.0013	1 ±0.0000	1 ±0.0000	0.9512 ±0.0013	1.9511 ±0.0013	64.7131 ±0.8459
M_{RF}	Avg. Std.	0.7983 ±0.0004	0 ±0.0000	0.299 ±0.0107	0.0536 ±0.0387	1.0973 ±0.0107	52.4259 ±0.9768
M_{QB}	Avg. Std.	0.2234 ±0.0011	1 ±0.0000	0.9312 ±0.0057	0.2234 ±0.0011	1.1547 ±0.0053	49.029 ±3.6415

Table 4.2 reveals several notable findings. First, according to the evaluation measures defined in the iDASH Challenge (E_1), our proposed method (M_{SF}) is the second best. However, since the best method (M_{GA}) assumes the users are accountable – which does not exist in the current system – it cannot be regarded as a practical solution for the iDASH Challenge.

With respect to effectiveness, we find that the two measures (E_1 and E_2) are in complete agreement regarding the rank order of the best methods. With respect to privacy, the second measure (P_2) does a better job of distinguishing between the methods than the first measure.

With respect to running time, we find that the solution we proposed (M_{SF}) exhibited the longest running time – about 56% longer than the quickest method. However, we assume that such a difference in efficiency may not be critical because the defender's strategy for each beacon is determined once the beacon is published rather than calculated on-the-fly.

While there is no perfect way to combine utility and privacy measures, we can compare the methods directly in these dimensions. In Figure 4.3, for each method, utility (U) is shown by the x-axis, privacy (P_2) is shown by the y-axis, and effectiveness (E_1) is shown in the text. It can be seen that, after dismissing the two impractical methods (M_T and M_{GA}), the method we proposed for the iDASH Challenge dominates all other solutions.

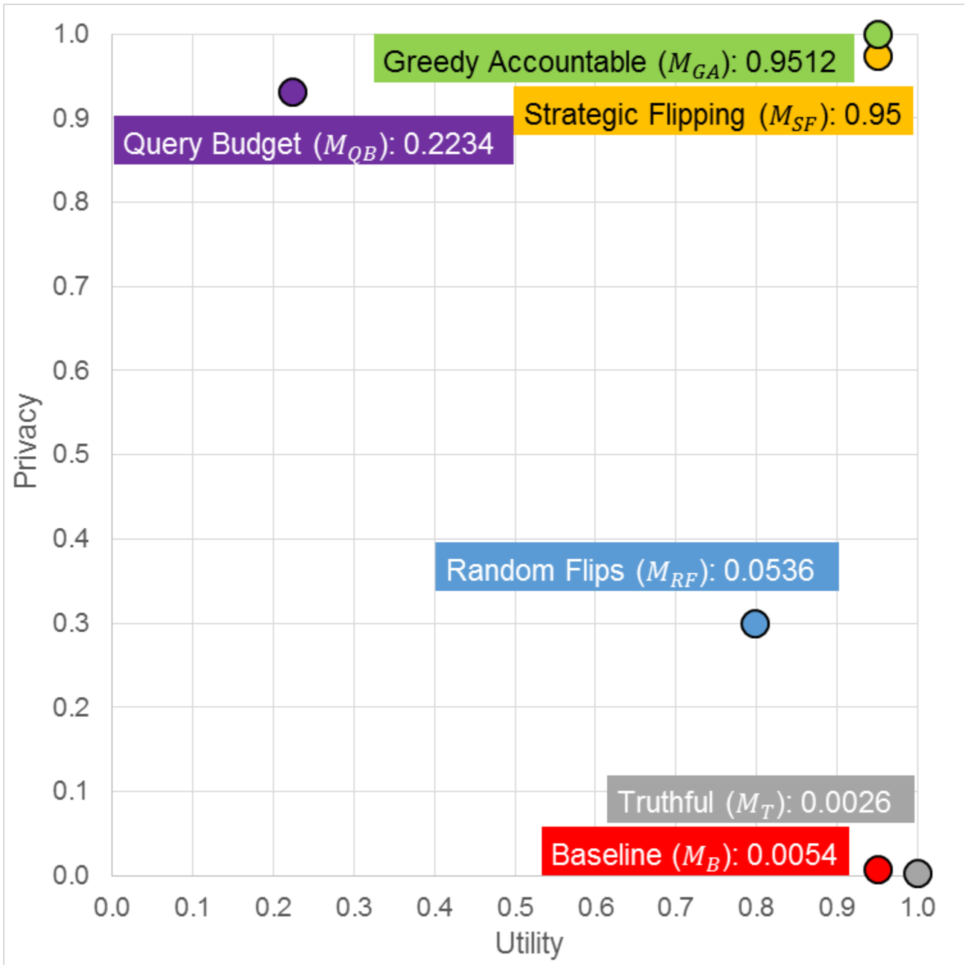


Figure 4.3 A comparison of the genomic data protection methods with respect to utility and privacy.

4.3.4 Sensitivity Analysis

To gain a deeper appreciation for the stability of the results of the iDASH Challenge, we assessed the performance of the proposed methods when certain key parameters are varied.

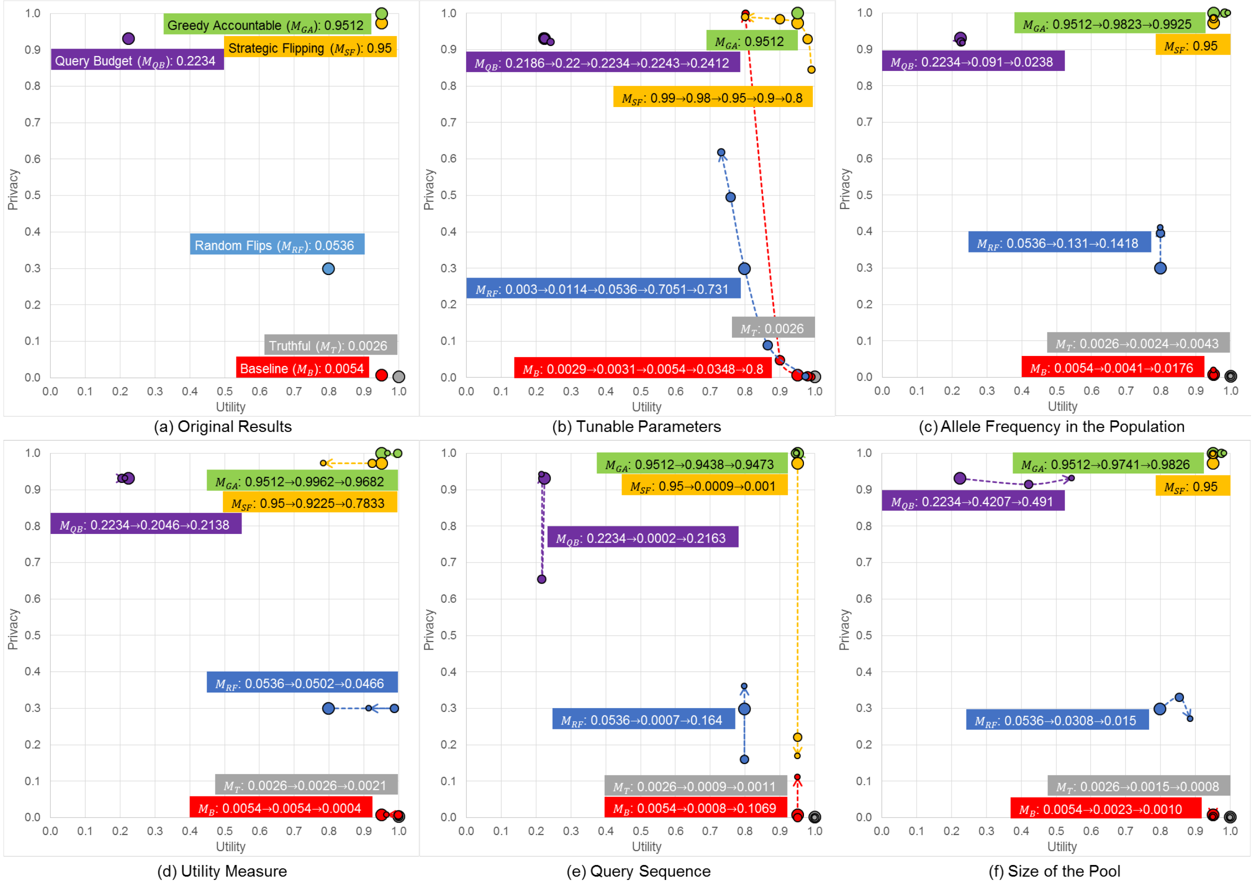
4.3.4.1 Tunable Parameters

All of the protection methods, except M_T and M_{GA}, include a tunable parameter. Here, we systematically investigate how changes to the value of this parameter influence their performance. For brevity in presentation, we designed four cases: 2 with values smaller and 2 with values larger than the value applied in the scenario investigated above. The specific values for the sensitivity analysis are detailed in Table 4.3. Only the most representative values are chosen for each parameter. The value of each parameter in Case 3 (Mid) is the default value as we used in the above experiment that simulated the iDASH Challenge. The values for each parameter in Case 1 (Low) and 5 (High) are the smallest and largest values, respectively, while Cases 2 (Low-Mid) and 4 (Mid-High) provide gradations between these cases to provide a complete view.

Table 4.3 Parameterizations for the sensitivity analysis of the genomic data protection methods.

Method	**Parameter**	**Case 1 (Low)**	**Case 2 (Low-Mid)**	**Case 3 (Mid)**	**Case 4 (Mid-High)**	**Case 5 (High)**
M_B	Percentage of flipped answers (k)	1	2	5	10	20
M_{SF}						
M_{RF}	Noise level (ε)	0.1	0.5	0.75	0.9	1
M_{QB}	Maximal allowable power (β)	0.1	0.5	0.9	0.95	1

Figure 4.4(b) displays the results of the sensitivity analysis of tunable parameters with respect to the utility, privacy, and effectiveness measures of different methods. The series of numbers near the series of circles represent the effectiveness (E_1) of the methods in different cases (from the Low Case to the High Case). There are several notable findings from this analysis to highlight. First, it should be recognized that utility is negatively correlated with the parameter in methods M_B, M_{SF} and M_{RF} and positively correlated with the parameter in the method M_{QB}. Second, privacy (P_2) is positively correlated with the parameter in M_B, M_{SF} and M_{RF} and negatively correlated with the parameter in M_{QB}. Third, effectiveness (E_1) is positively correlated with the parameter in M_B, M_{RF} and M_{QB} and negatively correlated with the parameter in the M_{SF}. The effectiveness of our proposed M_{SF} method is larger than all of the alternative methods, including M_{GA}, when the value for the k parameter is smaller than five. Most importantly, we found that the method we proposed for the iDASH Challenge dominates all of the alternative methods – except M_T and M_{GA} when the value for the k parameter is two.



(a) Original Results (b) Tunable Parameters (c) Allele Frequency in the Population

(d) Utility Measure (e) Query Sequence (f) Size of the Pool

Figure 4.4 Influence of the key parameters and factors on the utility and privacy measures.

(**a**) Original Results. (**b**) Tunable parameters. (**c**) Attacker's knowledge about allele frequencies in the population. (**d**) The utility measure. (**e**) Query sequence. (**f**) Size of the Pool.

4.3.4.2 Allele Frequency in the Population

In the iDASH Challenge, it was assumed that the attacker's estimate of the allele frequencies is similar to those in the 1000 Genomes population of around 2500 individuals. However, in the real world, the attacker is likely to have a stronger capability. For instance, the attacker may know the allele frequencies are from a smaller population.

Thus, we investigated how an enhancement of the attacker's capability influences the performance of the methods. Specifically, we assess the performance of the protection methods when the attacker has a more accurate estimate of the allele frequencies by gaining access to a smaller population of only 500 individuals. We further examine the scenario where the attacker relies on allele frequencies that are exactly the same as the pool behind the beacon.

Figure 4.4(c) displays the results of the sensitivity analysis of allele frequency in the population with respect to the utility, privacy, and effectiveness measures of different methods. The series of numbers near the series of circles (from larger to smaller circles) represent the effectiveness (E_1) of the methods in different cases (from larger to smaller populations). There are several important take-away messages from this analysis. First, both the effectiveness of protection (in terms of both E_1 and E_2) increases for M_B, M_{SF}, M_{GA}, M_{RF} and decreases for M_{QB} when the attacker's allele frequency estimate becomes more accurate. Second, the solution we proposed for the iDASH Challenge remains the best method for the current application (noting that M_{GA} still remains the best overall).

4.3.4.3 Utility Measure

In the iDASH Challenge, we assumed the utility for each SNV was equivalent. Here, we assume the utility is measured as a weighted sum of the truthful answers. The weight for each SNV can be (1) the absolute difference between the allele frequency in the pool and population or (2) a worst-case scenario for our proposed method, where the utility is equal to the absolute differential discriminative power the defender used for each SNV.

The results of this analysis are shown in Figure 4.4(d). The series of numbers near the series of circles (from larger to smaller circles) represent the effectiveness (E_1) of methods in different cases (from the original setting to the Low Case, and from the Low Case to the Low-Mid Case). In these scenarios, we find that our proposed M_{SF} for the iDASH Challenge performs substantially worse than it did in the challenge because the utility is measured differently. Nonetheless, even in such an extreme case, our method is never dominated by others, except M_{GA}, and dominates M_{QB}. However, it should be recognized when method A fails to dominate method B, it does not imply that method B dominates method A.

4.3.4.4 Query Sequence

Different sequences of the queries have the potential to yield different detection results. As such, we need to consider how well the attacker can perform if he chooses the sequence with the highest possible detection power. Let us consider two scenarios: (1) the attacker always queries the most discriminative set of SNVs first; (2) the attacker always queries the rarest SNVs first.

The performance of the protection methods with respect to these two scenarios are depicted in Figure 4.4(e). The series of numbers near the series of circles (from larger to smaller circles) represent effectiveness (E_1) of the methods in different cases (from the original setting to case 1, and from case 1 to case 2). As anticipated, it can be seen that the defender tends to lose privacy no matter what protection method is invoked in the first scenario. This is because the SNVs queried first have very strong discriminative power. However, in the second scenario, the defender loses privacy only when our proposed method is invoked and gains privacy when other methods are invoked.

These results primarily stem from three reasons. First, the SNVs with high differential discriminative power are not the same as the SNVs with high discriminative power or the SNVs with low alternative allele frequency. As a result, in the face of these two query sequences, the defender does not flip any SNVs until the very end of the query sequence. This leads to high detection power quickly. Second, a flipping strategy works best in the scenario where all SNVs that need to be flipped are also queried first. In other words, our proposed method works best when the top k percent SNVs with the highest differential discriminative power are queried first. By contrast, the baseline method works best when the top k percent SNVs with the lowest alternative allele frequency are queried first. Third, Raisaro et al. [159] assumed that the rarest SNVs are queried first. Thus, the Random Flips method (M_{RF}) and the Query Budget method (M_{QB}) perform well in the scenario where this assumption holds true. Still, the solution we proposed for the iDASH Challenge is not dominated by any of the other methods except for the Greedy Accountable method (M_{GA}) and dominates the baseline method (M_B). Still, if the attacker's query sequence is fixed and known by the defender, the Greedy Accountable method (M_{GA}) becomes practical and the defender will end up with nearly perfect scores.

4.3.4.5 Size of the Pool

In the iDASH Challenge, we used a dataset where there were 250 individuals in the pool behind the beacon. Here, we consider scenarios where there are fewer individuals in the beacon. Specifically, we assess the performance of the methods when there are only 1) 100 individuals and 2) 50 individuals in the pool. The results are shown in Figure 4.4(f). The series of numbers near the series of circles (from larger to smaller circles) represent the effectiveness (E_1) of the protection methods in different cases (from larger to smaller sized pools).

It can be seen that the effectiveness of protection is positively correlated with the pool size in methods M_T, M_B and M_{RF}. And, once again, the method we submitted to the iDASH Challenge remains the best practical method in terms of all evaluation measures.

4.4 Discussion

The findings illustrate that the ASBA attack against Beacon services can be sufficiently mitigated through a strategy that intelligently prioritizes which genomic variants to provide truthful answers about.

4.4.1 Limitations

There are three primary aspects of our protection method that should be addressed before it is instituted in practice: 1) the deterministic approach to flips, 2) an estimation of the attacker's target set, and 3) a grounded approach to select *k* (i.e., the percentage of SNVs to flip). It is also a limitation that the effectiveness of proposed game theoretic models for protecting Beacon services has not been evaluated using experiments.

First, our solution invokes a deterministic approach to selecting which SNVs to flip. This is potentially problematic because if the attacker was able to ascertain some of the allele frequencies in the pool behind the beacon, then they could mimic the strategy of the defender. In other words, the attacker would be able to determine about which SNVs the defender would choose to lie about. As a consequence, each query response for such SNVs could then be flipped back to the correct answer about the underlying pool, thus rescinding all of the protection. Therefore, in the event that there is a concern about such exposure, our model could incorporate a randomization component, where the answers provided to the adversary are non-deterministic. If such a feature were to be incorporated, it is critical to minimize the level of randomization to achieve the desired level of security.

Second, in the iDASH Challenge, the target set (i.e., the set of genomic records for presence/absence testing) was provided to the competition teams. However, in the real world, there may be multiple attackers, each of which may harbor a different target set. In such a scenario, the computation of the discriminative power for each SNV in the pool should be dependent on the underlying population of the beacon instead of a particular target set. In other words, when the defender is uncertain about the number of attackers and their targets, the entire population from which the beacon is sampled should be used as the target set in our model. However, the mismatch in target sets may affect the performance of our method.

Finally, the parameter k in our method determines the starting point of the search for local optimal strategy. A well-specified value of k increases the probability that the local optimal strategy is also globally optimal. The best choice for k is dependent upon a number of factors, including 1) the size of the pool, 2) the number of SNVs, 3) the maximal allowable false-positive rate, 4) the specific data in the pool, and 5) the target set. In practice, when the defender needs to determine *k*, he could simulate an attacker with an estimated maximal allowable false-positive rate (which is often set to 5%), as well as a target set, and then select the best choice empirically. For example, according to the results in Figure 6.3, the best choice of k is five in terms of maximizing the effectiveness (E_2).

4.4.2 Conclusions

This chapter introduced a technical solution for mitigating the Shringapure and Bustamante (SB) attack towards the Beacon services of the Global Alliance for Genomics and Health. This solution was specifically tailored to address an augmented version of this attack, as posed in the 2016 iDASH Challenge. This solution was the winning entry, and in this chapter, we provided a formalization of the iDASH Challenge problem, a general design of the protection model, and an empirical evaluation of our solution to demonstrate its potential for protecting privacy with minimal influence on the utility of the system within a practical computational runtime. Based on the solution, I proposed two game theoretic models that have the potentials to be applied to this data protection problem with anticipation of a better balance between the data utility and privacy.

We further showed, via an extensive experimental evaluation, that our proposed method outperforms all posited baseline and state of the art methods (that were applicable to real-world scenarios) regardless of how key parameters that drive the attack (e.g., the effectiveness measure, the number of records behind the beacon, and the attacker's estimate of allele frequency) vary. In most scenarios, the advantages of our method over other alternative methods are substantial. Still, it should be recognized that our method is limited by the design of the iDASH Challenge scenario (e.g., a strategy space limited to changing query answers) and the evaluation protocol (e.g., adversarial knowledge of minor allele frequencies).

Chapter 5

MULTI-STAGE RE-IDENTIFICATION GAME

Person-specific biomedical data is now collected on a large scale in a wide range of settings. Many believe that sharing such data beyond their initial point of collection is crucial to maximizing their societal value. However, data sharing efforts are often limited by privacy concerns, particularly over the identifiability of the individuals to whom the data corresponds. Seemingly anonymous records can be re-identified by linkage attacks.

Formal risk assessments can inform decisions about whether, and how, to share data, given the possibility of a high-profile attack. However, most assessment frameworks to date have been limited mainly because of their simplistic adversarial models. For instance, they only consider adversaries with access to a single resource. As has been shown in recent re-identification attacks, an adversary can access multiple data repositories associated with a targeted individual and combine them in a stage-wise manner to enhance the chance of success.

In this chapter, we introduce a game theoretic model to analyze a data-sharing problem in the face of a two-stage attack. Solving the game enables the discovery of the optimal data sharing strategy by balancing a privacy-utility tradeoff. Through experiments on large-scale genomic data, we demonstrate that maximal utility can be achieved if sharing partial data is allowed, in which case most data can be shared with little privacy sacrificed. Through extensive sensitivity analyses on essential parameters, we provide insights into risk mitigation directions for corresponding stakeholders.

5.1 Introduction

Person-specific biomedical data is now collected on a large scale in a wide range of settings. For instance, in the clinical realm, personal information is routinely stored in electronic health records. The biomedical research domain now supports studies that collect data on a diverse array of participants [1]. And, most recently, the commercial setting has led to a number of ventures where data is collected, such as direct-to-consumer genetic testing (DTC-GT) companies that collect data from various consumers and build repositories that now cover over 10% of the US population [2]. Many believe that sharing such data beyond their initial point of collection is crucial to maximizing their societal value. However, data sharing efforts are often limited by privacy concerns, particularly over the identifiability of the individuals to whom the data corresponds [4].

Genomic data, which is shared in various settings in the US, provides a clear illustration of the threat and concern. Linking such data to explicit identifiers (i.e., re-identification) poses a threat to one's anonymity. While data managers remove explicit identifiers (e.g., personal names and phone numbers) to adhere to de-identification guidance [5, 6, 292], numerous demonstration attacks have shown that data, and particularly genomic records, can be re-identified through a variety of means [7, 8, 87, 293]. Although individuals are incentivized to share data [61, 10], they usually lack the ability to identify and assess privacy risks properly in order to make the right sharing decisions.

It is important to recognize that not all re-identification attacks are equally easy to execute and that an oversimplified attack model can lead to an inaccurate measure of risk. Moreover, this inaccuracy is not biased in any particular direction, such that risk may be underestimated in some cases, but overestimated in others. Initially, attacks were based on a single stage [72, 73, 75], where the adversary linked two datasets – one de-identified and one identified – using attributes shared by these datasets (e.g., residual demographics or DNA sequences). More recently, attacks have evolved into multi-stage forms [31, 34, 294, 35, 78], where each stage reveals another piece of information about a targeted individual. In this chapter, we introduce the first approach to assess and strategically mitigate risks by explicitly modeling and quantifying the privacy-utility tradeoff for data subjects in the face of multi-stage attacks. In doing so, we bridge the gap between more complex models of attack and informed data sharing decisions.

For illustration, we rely on the well-known two-stage attack model of Gymrek et al. [31], which, to re-identify genomic data, combines surname inference with direct linkage. The attack specifically targeted 10 participants in the Center for Study of Human Polymorphisms (CEPH) family collection, whose genomes were sequenced as part of the 1000 Genomes Project [230] by performing surname inference through public genetic genealogy databases made accessible by Ysearch and the Sorenson Molecular Genealogy Foundation (SMGF), the latter subsequently purchased by Ancestry.com. In response, in consultation with the local institution for the CEPH study, the National Institutes of Health moved certain demographics about the participants in the corresponding repository into an access-controlled database [76]. Ysearch and SMGF databases are no longer accessible to the public [295], but it is not unreasonable to assume that similar databases may be made publicly accessible in the future.

Various approaches for preventing biomedical data re-identification have been developed from regulatory [296, 297, 298] and technological perspectives [109, 167, 111, 113, 114]. However, most of these approaches focus on worst-case scenarios, such that their impacts on data utility and privacy risk in practice are unclear. For example, the adversary considered in these approaches always attacks without considering the attacking costs in the real world [7, 8], which may overestimate the privacy risk. In addition, parameters in technical protection models (e.g., k-anonymity [45] or differential privacy [48]) are usually set without, or before, measuring their impacts in specific use cases [125, 126], which may either sacrifice too much data utility or provide insufficient protection. To address this problem, risk assessment and mitigation based on game theoretic models have been introduced [236, 224, 237]. In this work, assuming an adversary acts rationally with limited resources, we show that a game theoretic model can reveal the optimal sharing strategy to data subjects. Most importantly, we conducted experiments involving protection against a multi-stage attack using either real-world datasets or large-scale simulated datasets. Our results demonstrate that the game theoretic model can efficiently assess and effectively mitigate privacy risks. The fine-grained sharing strategy recommended by our model virtually eliminates the re-identification risk and maximizes the data utility.

5.2 Methods

5.2.1 System Model

As shown in Figure 5.1, we investigate the problem of privacy-preserving data sharing from the perspective of a system that includes a data subject, three databases, and an adversary. Our model's goal is to help a data subject, given a multi-stage attack model, decide whether and how to share person-specific data with the targeted database (i.e., the database in the middle). In our model, the data subject wants to share data for various reasons but has privacy concerns. In addition, the targeted database only releases de-identified (i.e., without personal identifiers) datasets to the public or a third party. The dataset recipient could be a malicious adversary who wants to re-identify the targeted data subject's record from the dataset and has limited incentives, abilities, and resources.

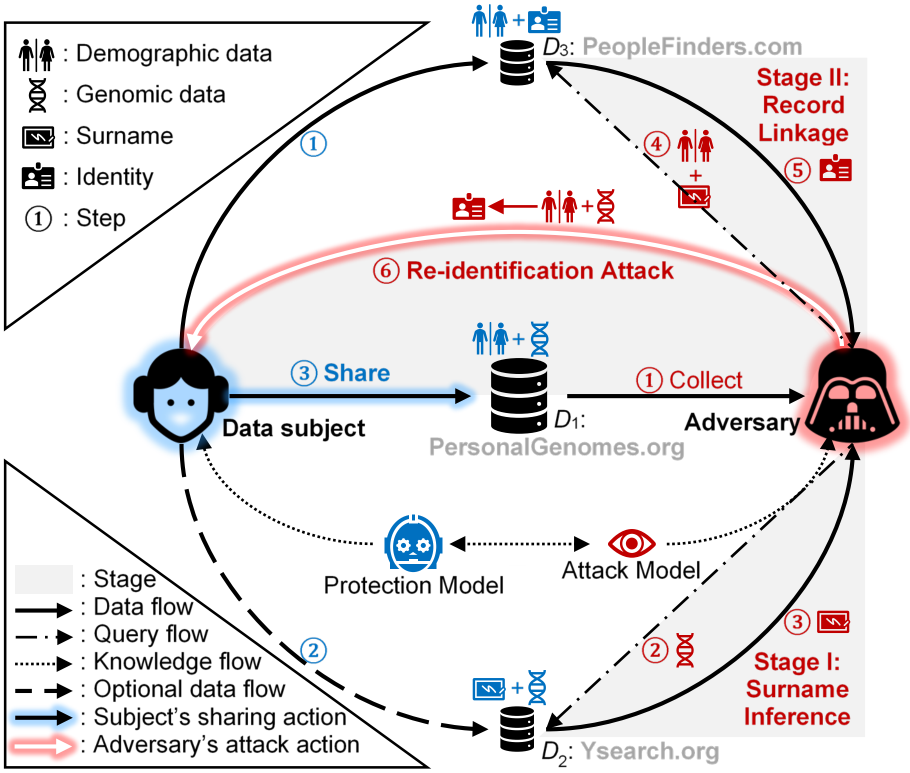


Figure 5.1 A system-wide perspective of a multi-stage re-identification attack and its protection.

Person-specific data records of a subject are accessible to an adversary through three databases: a targeted genomic database (D_1), a genetic genealogy database (D_2), and a public registration database (D_3). The adversary re-identifies a genomic record by inferring surnames in Stage I and linking it to a public record in Stage II. The data subject selects a sharing strategy based on a game model only when sharing data in D_1.

In this system, data flow from the data subject to the adversary through three channels. Databases in the first channel are public registration databases such as a voter registration list [71], a public record search engine (e.g., Intelius[14] and PeopleFinders[15]) or a social media service [299] (e.g., Facebook[16] and LinkedIn[17]). These data are usually associated with real identities. Let us assume that some personal identifiers and demographic attributes of the data subject are accessible to the adversary through one of these public databases.

Databases in the second channel are not likely to be targeted by the adversary because of their well-established data sharing policies. For example, they only share anonymous granular data with a limited set of trusted third parties, and they only share summarized or anonymized individual-level data (i.e., with

[14] Intelius. https://www.intelius.com/.
[15] PeopleFinders. https://www.peoplefinders.com/.
[16] Facebook. https://www.facebook.com/.
[17] LinkedIn. https://www.linkedin.com/.

no demographic information or with rigorous risk mitigation) to the public. Databases in this channel are usually controlled by large non-profit organizations (e.g., the Global Alliance for Genomics and Health[18]), government-sponsored research programs (e.g., All of Us Research Program[19] [1] and UK Biobank[20] [300]) and large direct-to-consumer genetic testing (DTC-GT) companies (e.g., 23andMe[21] and AncestryDNA[22]). Although data subjects associated with data in this channel are not necessarily targeted by the adversary, the adversary can use their data to infer additional attributes of targeted data subjects, as illustrated in our attack model.

The adversary targets databases in the third channel because either their data sharing policies are vulnerable to privacy attacks, or their datasets are openly accessible, making them less trustworthy. Databases in this channel include recreational genomic databases (e.g., GEDmatch[23], Ysearch[24]) and open-access databases (e.g., OpenSNP[25] [61] and Personal Genome Project[26] [60]). Compared with the second channel, the third channel is more likely to allow a data subject to choose which part of a record can be shared. In contrast to the first channel, the third channel allows a data subject to share de-identified genomic and phenotypic data. Genomic data shared in this channel are vulnerable to various attacks such as re-identification [75, 31] and data extraction attacks [104]. Because the re-identification attack is more typical than other attacks, we specifically investigate the problem of optimal data sharing in this channel against a multi-stage re-identification attack.

5.2.2 Attack Model

The attack model considered in this system has multiple stages. Let us use a two-stage attack as an example. Before an attack, the adversary receives a piece of information (x_1) about a target (i.e., a targeted data subject) from a dataset (D_1). In the first stage, the adversary infers a piece of information (x_2) about the target according to another dataset (D_2) and information $x_1$$\mathrm{x}_1$. In the second stage, a piece of information (x_3) about the target is learned by querying another dataset (D_3) with x_1 and x_2. Alternatively, the adversary may query dataset D_3 with x_1 only (without inferring x_2), which reduces the number of

[18] Global Alliance for Genomics and Health. https://www.ga4gh.org/.
[19] All of Us Research Program. https://allofus.nih.gov/.
[20] UK Biobank. https://www.ukbiobank.ac.uk/.
[21] 23andMe. https://www.23andme.com/.
[22] AncestryDNA. https://www.ancestry.com/dna/.
[23] GEDmatch. https://www.gedmatch.com/.
[24] Ysearch. http://www.ysearch.org/.
[25] OpenSNP. https://opensnp.org/.
[26] Personal Genome Project. https://www.personalgenomes.org/.

attack stages to one. Still, if additional information gets inferred in the first stage, the prediction in the two-stage attack tends to be more accurate than the one in the one-stage attack. In general, the attack can have more than two stages, in which each stage infers a piece of new information, based on an additional dataset, which can be used in subsequent stages. Generally, the prediction in a multi-stage attack tends to be more accurate than predictions in the attack's variations with fewer stages.

Figure 5.2 illustrates the two-stage attack executed based on Gymrek et al.'s work [31], which we refer to as the Gymrek attack. This attack blends surname inference with record linkage. More formally, in this attack, D_1 is a dataset with genomic and demographic attributes as information x_1; D_2 is a genetic genealogy dataset with genomic attributes as the first part of information x_1 and with surnames as information x_2; and D_3 is an identified dataset with demographic attributes as the second part of information x_1, with surnames as information x_2, and with first names as information x_3. More specifically, in their experiments, Gymrek et al. used extracted short tandem repeats on Y-chromosome (Y-STRs) from Ybase[27], the 1000Genomes project, and the National Center for Biotechnology Information archives as genomic dataset D_1, with demographic attributes (namely, year of birth and state of residence). They also used records from Ysearch and Sorenson Molecular Genealogy Foundation (SMGF) as genetic genealogy dataset D_2, with attributes including Y-STRs and surnames. Finally, they used the record search engine, PeopleFinders, based on mined records from voter and driver registries, as identified dataset D_3, with attributes including age, state of residence, first name, and surname. We use the same surname inference and record linkage methods as Gymrek et al. used and further assume that an adversary randomly attacks one matched identified record in the final re-identification stage.

[27] Ybase. http://www.ybase.org/, archived in 2010.

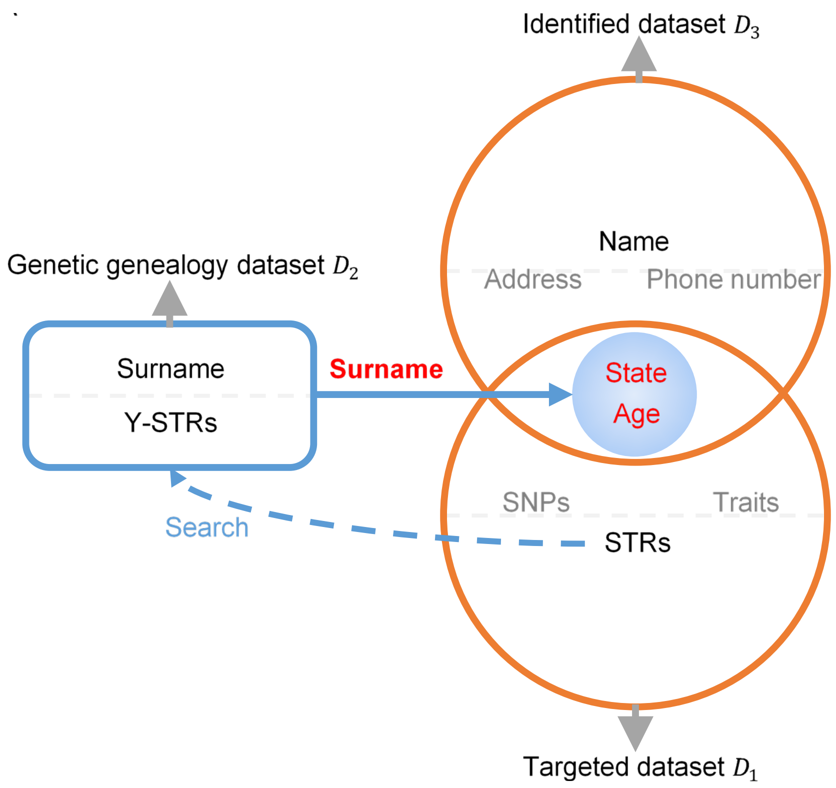


Figure 5.2 Illustration of the Gymrek attack model.

In this attack model, attributes exist in both the targeted genomic dataset and the identified dataset include state of residence and age (or birth year). Genomic attributes such as SNP and Y-STR markers and phenotypic attributes such as eye colors and hair colors only exist in the targeted genomic dataset. Full name and contact information only exist in the identified dataset. Surnames could be inferred from the Y-chromosome of a targeted DNA sequence by searching a reference dataset, which enhances the record linkage attack. SNP stands for single-nucleotide polymorphism, and STR stands for short tandem repeat.

As in the Gymrek attack model, we assume that the database holding dataset D_1 releases the entire dataset on an individual level and that databases holding datasets D_2 and D_3 provide query services. Note that the attack can be executed either before or after data is shared through the first two channels. The inferred surname in the first stage will generally be correct if a record corresponding to the data subject is in dataset D_2 when the attack happens. The attack will likely fail if the data subject's record is not in the dataset D_3 when the attack happens. Thus, for simplicity, in our experiments, we assume that only dataset D_3 (instead of dataset D_2) includes a record corresponding to the data subject when the attack happens.

5.2.3 Protection Model

Table 5.1 Notation, with description and setting, used throughout this work.

Notation	Description	Range	Setting
	Parameters in the experiment based on simulated datasets		
D_1	The targeted genomic database (or dataset)	/	/
D_2	The genetic genealogy database (or dataset) used for surname inference	/	/
D_3	The public identified database (or dataset) used for record linkage	/	/
M_1	Number of demographic attributes	$\mathbb{N}$	2
M_2	Number of genomic attributes	$\mathbb{N}$	12
γ	Proportion of missing genomic data in dataset D_2	[0,1]	0.3
θ	The threshold for confidence score	[0,1]	0.5
n	Number of records in the targeted genomic dataset D_1	$\mathbb{N}$	1,000
n_2	Number of records in the genetic genealogy dataset D_2	$\mathbb{N}$	20,000
n_3	Number of records in the identified dataset D_3	$\mathbb{N}$	20,000
L	A data subject's loss from being re-identified	$\mathbb{R}^+$	\$150
B	The benefit of sharing all data in a data record	$\mathbb{R}^+$	\$100
C	The adversary's cost of an attack (cost to execute an attack)	$\mathbb{R}^+$	\$10
N_i	Number of iterations for each experiment	$\mathbb{N}$	100
	Parameters in the Gymrek attack		
N_e	Effective male population size	$\mathbb{N}$	10,000
T	Number of generations for the patrilineal surname system	$\mathbb{N}$	200
N_c	Number of pre-selected candidates	$\mathbb{N}$	10
P_t	Proportion of tolerable unmatched markers to maximal matched ones	[0,1]	0.2
θ_s	Jaro-Winkler string distance threshold	[0,1]	1
θ_m	Lower bound on the number of available markers	$\mathbb{N}$	17
	Parameters in the population simulation environment		
N_s	Number of surnames	$\mathbb{N}$	1,000
S_{mp}	Size of the simulated male population	$\mathbb{N}$	90,064
N_{sp}	Number of subpopulations	$\mathbb{N}$	3
P_m	Migration proportion	[0,1]	0.1
T_g	Number of generations	$\mathbb{N}$	10
S_{sp}	Size of each subpopulation in the first generation	$\mathbb{N}$	20,000
	Variables ordered by their chronological introduction		
m	The number of attributes in the data subject's record	$\mathbb{N}$	/
$\boldsymbol{s}$	A vector of binary elements representing an action of the data subject	$\mathbb{B}^m$	/
s_j	An element in $\boldsymbol{s}$ indicating whether j^{th} attribute is shared ($j \in [1, m]$)	$\mathbb{B}$	/
a	A binary variable representing an action of the adversary	$\mathbb{B}$	/
$b(\boldsymbol{s})$	The benefit of sharing data given the data subject's strategy	$\mathbb{R}^+$	/
B_j	The benefit for sharing the j^{th} genomic attribute in a data record	$\mathbb{R}^+$	/
$p(\boldsymbol{s})$	The probability of an attack's success given the subject's strategy	[0,1]	/
$v_d(\boldsymbol{s}, a)$	The data subject's payoff given both players' actions	$\mathbb{R}$	/
$v_a(\boldsymbol{s}, a)$	The adversary's payoff given both players' actions	$\mathbb{R}$	/
$\hat{p}(\boldsymbol{s})$	The adversary's estimated probability of an attack's success given $\boldsymbol{s}$	[0,1]	/
$\hat{v}_a(\boldsymbol{s}, a)$	The adversary's estimated payoff given both players' actions	$\mathbb{R}$	/
$\phi(\boldsymbol{s})$	The set of the adversary's best actions given the subject's strategy $\boldsymbol{s}$	$\{\mathbb{B}\}$	/
$\boldsymbol{s}^*$	The data subject's best strategy	$\mathbb{B}^m$	/
$p_1(\boldsymbol{s})$	The probability of Stage I's success	[0,1]	/
$p_2(\boldsymbol{s})$	The probability of Stage II's success given that Stage I succeeds	[0,1]	/
$a'(\boldsymbol{s})$	A binary variable representing whether Stage I is not omitted	$\mathbb{B}$	/

$p'(\boldsymbol{s})$	The probability of Stage II's success given that Stage I is omitted	[0,1]	/
$\hat{p}_1(\boldsymbol{s})$	The adversary's estimation on the probability of Stage I's success	[0,1]	/
$\boldsymbol{s}_i^*$	The i^{th} data subject's best strategy in dataset D_1	$\mathbb{B}^m$	/
a_i^*	The adversary's best response for the i^{th} data subject's best strategy	$\mathbb{B}$	/
$\bar{V}$	The average payoff of n data subjects whose records are in dataset D_1	$\mathbb{R}$	/
V_i	The i^{th} data subject's optimal payoff in dataset D_1	$\mathbb{R}$	/
$\bar{U}$	The average data utility of n subjects whose records are in dataset D_1	[0,1]	/
U_i	The i^{th} subject's data utility in dataset D_1 given the best sharing strategy	[0,1]	/
$\bar{P}$	The average privacy of n data subjects whose records are in dataset D_1	[0,1]	/
P_i	The i^{th} subject's privacy in D_1 given players' best actions and the record	[0,1]	/
σ_V	The standard deviation of n data subjects' payoffs	$\mathbb{R}$	/
σ_U	The standard deviation of n data subjects' data utility metrics	[0,1]	/
σ_P	The standard deviation of n data subjects' privacy metrics	[0,1]	/
x_1	Demographic and genomic attributes in dataset D_1	/	/
x_2	The surname attribute in dataset D_2	/	/
x_3	Identity attributes (e.g., name) in dataset D_3	/	/
v_d	The subject's payoff if all data is shared and the adversary attacks	$\mathbb{R}$	/
p	The probability of an attack's success with all data being shared	[0,1]	/
s	The data subject's strategy in the opt-in game	$\mathbb{B}$	/
s^*	The data subject's best strategy in the opt-in game	$\mathbb{B}$	/
p_1	The probability of Stage I's success in the opt-in game	[0,1]	/
p_2	Probability of Stage II's success given Stage I succeeds in opt-in game	[0,1]	/
r	The correctness of the inferred surname	[0,1]	/
k	Number of matched identified records in the linkage upon quasi-identifiers and the inferred surname	$\mathbb{N}$	/
p'	probability of Stage II's success given Stage I is omitted in opt-in game	[0,1]	/
k'	Number of matched identified records in linkage upon quasi-identifiers	$\mathbb{N}$	/
v_a	The adversary's payoff if all data is shared and the adversary attacks	$\mathbb{R}$	/
$\widehat{v_a}$	The adversary's estimated payoff with all data shared and the attack	$\mathbb{R}$	/
a^*	The adversary's best action in the opt-in game given the shared data	$\mathbb{B}$	/
$\hat{p}$	Adversary's estimated prob. Of an attack's success with all data shared	[0,1]	/
$\hat{p}_1$	The adversary's estimation on the probability of Stage I's success	[0,1]	/
$\hat{r}$	The adversary's estimated correctness of the inferred surname	[0,1]	/
a'	The adversary's optimal decision on whether not to omit Stage I	$\mathbb{B}$	/
$\phi(s)$	The set of the adversary's best responses to the data subject's strategy s	$\{\mathbb{B}\}$	/
$\boldsymbol{s}'$	A "child" of the strategy $\boldsymbol{s}$ in a lattice representation	$\mathbb{B}^m$	/
U	The data subject's data utility given the best sharing strategy	[0,1]	/
$\boldsymbol{S}_t$	The data subject's t^{th} sharing strategy ($t \in [1, 2^m]$)	$\mathbb{B}^m$	/
$\boldsymbol{A}_l$	The adversary's l^{th} attacking strategy ($l \in [1, 2^{2^m}]$)	$\mathbb{B}^{2^m}$	/

To mitigate the re-identification risk, Gymrek et al. recommended protecting the data through data masking (i.e., hiding some data) and access control. However, these strategies' performances have not been empirically evaluated or strategically analyzed in the context of a multi-stage attack. Without a proper protection strategy, responses to these attacks in practice could be shrinkage of the user base or closure of the entire database, which could substantially harm the data utility and hamper scientific progress.

In our framework, we determine the optimal protection strategy for a data subject to share personal data, assuming they are rational and driven by incentives. To do this, we consider three typical scenarios. These scenarios are slightly different in complexity, with the first the simplest and the last the most complex. We further assume that all databases' sharing strategies are fixed and known to all parties in the system in each scenario.

The notation used throughout this chapter is summarized in Table 5.1.

5.2.3.1 Always-attack Scenario

In the *always-attack* scenario, we consider the worst case in which the adversary acts as if he or she has unlimited resources and incentives (i.e., always attacks), and the data subject makes a binary decision on whether to share personal data with the targeted database. Given the benefit of sharing data (B) (i.e., sharing information x_1), the loss from being re-identified (L) by the adversary, and the probability of an attack's success (p), the data subject's payoff equals the benefit minus the expected cost:

$$v_d = B - Lp. \tag{5.1}$$

For a data subject, the sharing decision (s) can be represented as 1 (i.e., to share) or 0 (i.e., not to share). For a rational data subject, the optimal sharing decision (s^*) will be 0 if and only if the expected payoff is smaller than 0, as shown in Equation 5.2:

$$s^* = \begin{cases} 1, & B \geq Lp, \\ 0, & B < Lp. \end{cases} \tag{5.2}$$

The parameters in the model can be set according to specific cases. Parameter B's setting depends on the monetary incentive for sharing data or the value of a complimentary service that requires person-specific data. For example, in the case of sharing data with a free relative finder service, it could be set to the price that 23andMe, a DTC-GT company, asks for its relative finder service, valued at $99. In other cases, it could also be set according to the price of shared data in an online marketplace (e.g., blockchain-based platforms [154]), the "data dividends" paid by companies that use shared data (e.g., the EU's TRUSTS project [301]), or compensation for participation in a research program (e.g., the All of Us Research Program [1]).

Parameter p's calculation depends on the simulation of an attack, which requires knowledge of the attack model and datasets used in the attack. If the attack model is published in a high-impact journal and datasets used in the attack are public, the attack's simulation is possible. For the Gymrek attack, we assume that an adversary randomly attacks one matched identified record when there are more than one matched identified records in the final re-identification stage. Thus, p is calculated as the reciprocal of the number of matched identified records if the inferred surname is correct. Note that p = 0 if an incorrect surname is inferred. As a result, in a two-stage attack, the probability of an attack's success can be represented and calculated as:

$$p = p_1 p_2 = r/k, \tag{5.3}$$

in which p_1 is the probability of the first stage's success, and p_2 is the probability of the second stage's success given that the first stage succeeds. Additionally, we have $p_1 = r$ in which r is the correctness of the inferred surname, and we have $p_2 = 1 / k$ in which k is the number of matched identified records in the linkage upon quasi-identifiers (i.e., common attributes in two linkable datasets) and the inferred surname. In the situation that no surname is inferred, we have:

$$p = p' = 1/k', \tag{5.4}$$

in which p' is the probability of the second stage's success, given the first stage is omitted. Additionally, k' is the number of matched identified records available for linkage solely upon their quasi-identifiers. This situation happens when the confidence score is below a threshold (i.e., no surname is inferred) or the number of matched identified records is zero (i.e., the adversary realizes that the inferred surname is incorrect). Note that parameters like p_2 and p' are defined according to the Marketer re-identification risk model [163]. Because it is assumed that a record corresponding to each targeted data subject is in the identified dataset when the attack happens, we always have $k' \geq k \geq 1$ and $p' \leq p_2 \leq 1$, if the inferred surname is correct.

Parameter L, the loss that an attack brings to a data subject, could be set according to the outcome of a successful attack or the data subject's valuation on privacy. For example, it could be set to the increment of insurance payment (e.g., $150 per year) if the adversary is a life insurance company, and a pathological gene is identified in the target's genome. While the settings of parameters like L may be uncertain, we did extensive sensitivity analysis on these parameters.

5.2.3.2 Opt-in Game Scenario

In the *opt-in game* scenario, we assume that both the adversary and the data subject make rational decisions. In other words, we assume that the adversary will attack if and only if the benefit of the attack outweighs or equals the cost of the attack (although it is still arguable whether the adversary will attack if the benefit equals the cost). In this scenario, game theory can be naturally harnessed to determine the data subject's best strategy.

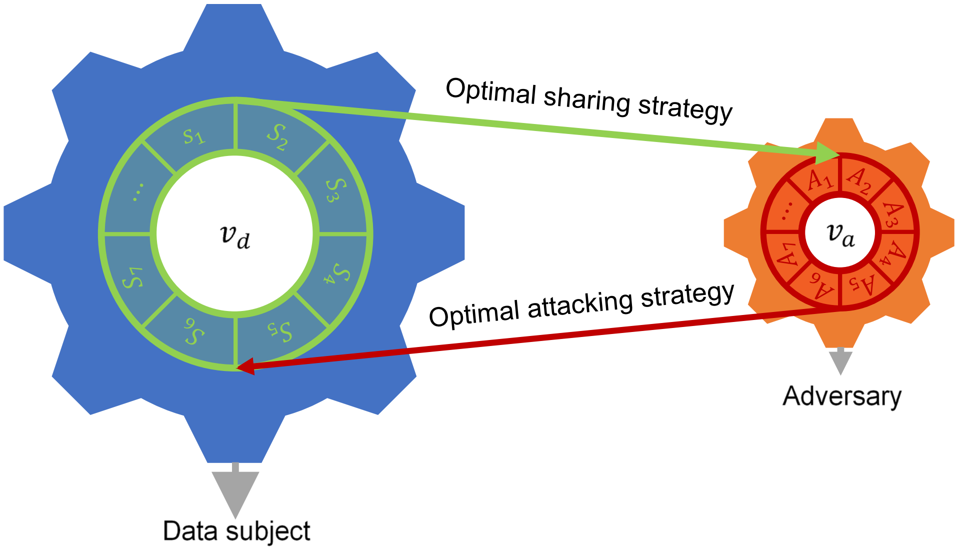


Figure 5.3 Illustration of the two-player game model.

Given the adversary's attacking strategy, the data subject chooses the optimal sharing strategy that maximizes the data subject's payoff. Given the data subject's strategy, the adversary chooses the optimal attacking strategy that maximizes the adversary's payoff. The game is solved when the whole system is in a Nash equilibrium, which means that both players can find no better strategy. In the figure, v_d and v_a are payoffs (or utilities) for the data subject and the adversary, respectively. Notably, $\boldsymbol{S}_t$ represents the data subject's t^{th} sharing strategy, and the A_l represents the adversary's l^{th} attacking strategy.

A typical two-player game model is illustrated in Figure 5.3. Specifically, a two-player Stackelberg (i.e., leader-follower) model is applied to this scenario. In this model, the data subject is the leader, and the adversary is the follower. The data subject's action is to share or not, represented by a binary variable (s). The adversary's action is to attack or not, represented by a binary variable (a). Assuming the expected gain of the adversary is the same as the expected loss of the data subject (Lp), given the expected benefit of a successful attack to the adversary ($L\hat{p}$) and the cost of an attack I, the payoff to the adversary is:

$$v_a = Lp - C, \tag{5.5}$$

and the adversary's estimated payoff is:

$$\widehat{v_a} = L\hat{p} - C, \tag{5.6}$$

in which $\hat{p}$ is the adversary's estimated probability of an attack's success. Since the adversary does not know the ground truth of the inferred information, the adversary must compute the probability of an attack's success by estimating the correctness of the inference in the first stage. For an adversary, given the shared data, the attack decision (a) can be represented by 1 (i.e., to attack) or 0 (i.e., not to attack). For a rational adversary, given the shared data, the optimal attack decision (a^*) will be one if and only if the expected payoff is larger than 0, as shown in Equation 5.7:

$$a^* = \begin{cases} 1, & L\hat{p} > C, \\ 0, & L\hat{p} \le C. \end{cases} \tag{5.7}$$

In the case of the Gymrek attack, $\hat{p}$ can be represented and calculated as:

$$\hat{p} = \hat{p}_1 p_2 = \hat{r}/k, \tag{5.8}$$

in which $\hat{p}_1$ is the adversary's estimation of the probability of the first stage's success, and p_2 is the probability of the second stage's success given that the first stage succeeds. In addition, we have $\hat{p}_1 = \hat{r}$ in which $\hat{r}$ is the estimated correctness of the inferred surname, which could be calculated as the confidence score as defined in the Gymrek attack model. In the situation that no surname is inferred, no estimation is required:

$$\hat{p} = p = p' = 1/k'. \tag{5.9}$$

This situation happens and only happens when $1/k' \ge \hat{r}/k$ (i.e., $p' \ge \hat{p}_1 p_2$). In other words, we assume that the adversary will only choose to infer a surname if and only if the inference brings the adversary a higher estimated probability of an attack's success. The adversary's optimal decision on whether to infer surname (i.e., whether not to omit) in the first stage (a') can be represented by 1 (i.e., to infer) or 0 (i.e., to omit) and can be calculated as in Equation 5.10:

$$a' = \begin{cases} 1, & \hat{p}_1 p_2 > p', \\ 0, & \hat{p}_1 p_2 \le p'. \end{cases} \tag{5.10}$$

The cost of an attack I can combine the cost for accessing data, the computing cost, and the cost of expected liquidated damage penalty. For simplicity, we assume that the penalty cost and the computing cost are both negligible. Thus, the parameter C could be set to \$10, which is the cost for accessing a subject's report from PeopleFinders or Intelius. The report contains the subject's personal information such as phone numbers, home addresses, family members, and property details. Given C = \$10 and L = \$150, a data subject will be attacked if $k \le 1/\hat{p} < L/C = 15$ (according to Equations 5.7 and 5.8) or $k' = 1/\hat{p} < L/C = 15$ (according to Equations 5.7 and 5.9).

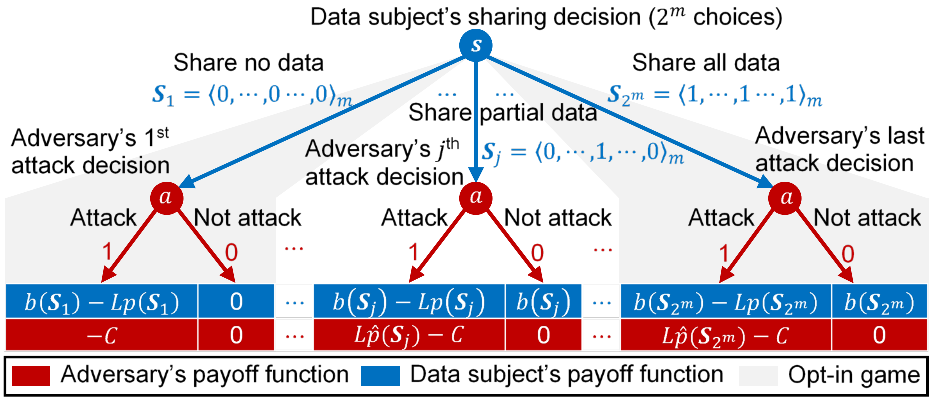

Data subject's sharing decision (2^m choices)
s
Share no data
$S_1 = \langle 0,\cdots,0\cdots,0\rangle_m$
Share partial data
$S_j = \langle 0,\cdots,1,\cdots,0\rangle_m$
Share all data
$S_{2^m} = \langle 1,\cdots,1\cdots,1\rangle_m$
Adversary's 1st attack decision
Adversary's jth attack decision
Adversary's last attack decision
a
Attack
Not attack
1
0
$b(S_1) - Lp(S_1)$
0
$-C$
0
$b(S_j) - Lp(S_j)$
$b(S_j)$
$L\hat{p}(S_j) - C$
0
$b(S_{2^m}) - Lp(S_{2^m})$
$b(S_{2^m})$
$L\hat{p}(S_{2^m}) - C$
0
Adversary's payoff function
Data subject's payoff function
Opt-in game

Figure 5.4 A masking game represented in the extensive form.

In a masking game, the data subject moves first, and the adversary moves next. Each terminal node is associated with both players' payoffs. $\boldsymbol{Sj}$ is an m-dimensional binary vector representing the j^{th} concrete action of the data subject. More denotation details are in Table 5.1. The opt-in game is a special variation of the masking game in which the data subject only has those two strategies.

The Stackelberg game model in the opt-in game scenario can be represented in the extensive form (shaded in gray), as shown in Figure 5.4. The solution to the Stackelberg game (i.e., both player's optimal decisions) can be represented as an optimization problem, as shown in Equation 5.11:

$$\begin{aligned}(s^*, a^*) &= \underset{s,a\in\phi(s)}{\operatorname{argmax}}(Bs - Lpsa), \\ \phi(s) &= \left\{a \,\middle|\, \underset{a}{\operatorname{argmax}}(L\hat{p}sa - Ca)\right\},\end{aligned} \tag{5.11}$$

in which $\phi(s)$ is the set of the adversary's best responses to the data subject's strategy s. Given certain conditions, the optimal solution can be represented in the explicit form, as shown in Equation 5.12:

$$\begin{cases} s^* = 1, a^*(s^*) = 1, & L\hat{p} \geq C, Lp \leq B, \\ s^* = 0, a^*(s^*) = 0, & L\hat{p} \geq C, Lp > B, \\ s^* = 1, a^*(s^*) = 0, & \quad L\hat{p} < C. \end{cases} \tag{5.12}$$

According to Equation 5.12, the data subject should share data if there is no attack. This situation happens when the cost of an attack is high enough, the loss from being re-identified is small enough, or the estimated probability of an attack's success is small enough (i.e., the number of matched identified records is large enough). Even if there is an attack, a rational data subject will still choose to share data if the benefit of sharing is large enough. Given B = \$100 and L = \$150, a data subject should share no data if $k \leq 1/p < L/B = 1.5$ (according to Equations 5.3 and 5.12) and $k' = 1/p < L / B = 1.5$ (according to Equations 5.4 and 5.12) (i.e., if the record that needs to be shared can be uniquely identified).

Note that, when an adversary does not incur any cost (i.e., C = 0), the opt-in game scenario becomes the always-attack scenario. For this reason, the always-attack scenario, a special case of the opt-in game, is not considered as an independent scenario in our experiments.

5.2.3.3 Masking Game Scenario

The *masking game* is almost the same as the opt-in game except that, we consider a data subject with more control over data sharing in the masking game scenario. Instead of a binary decision on whether or not to share the whole data record, the data subject is afforded a larger action space. When uploading data to GEDmatch, OpenSNP, or Personal Genome Project, the choices a data subject can invoke to modify data include i) uploading only a portion of it or ii) uploading fake data. However, data quality affects service quality. Thus, data subjects are incentivized to share more, as well as authentic, data. Here, we assume that the data subject will upload only authentic data. For simplicity, we further assume that the only action the data subject can take is masking. In other words, the data subject's action determines whether to mask or share each attribute in his or her record. It is represented by an m-dimensional vector of binary elements: $\boldsymbol{s} = \langle s_1, \cdots, s_j, \cdots, s_m \rangle \in \mathbb{B}^m$, in which m is the number of attributes in the record, with $s_j = 0$ if j^{th} attribute is masked or $s_j = 1$ if j^{th} attribute is shared (i.e., not masked). Each strategy of the data subject corresponds to one action. As a result, the size of the strategy space for the data subject increases exponentially as a function of the number of attributes. For illustration, let us assume that only fixed strategies are considered. For a record with m attributes, the number of sharing strategies is 2^m, the same as the number of the adversary's decision nodes. Given a sharing strategy (or shared data), the adversary's action (a) is either 1 (i.e., to attack) or 0 (i.e., not to attack). A strategy of the adversary specifies the adversary's action at each of his or her decision nodes.

The benefit of sharing increases monotonically as the amount of shared data increases. It can be regarded as a function of the data subject's sharing strategy, as shown in Equation 5.13:

$$b(\boldsymbol{s}) = \sum_{j=1}^{m} B_j s_j, \tag{5.13}$$

where B_j is the benefit of sharing the j^{th} attribute, set according to the j^{th} attribute's information entropy in either the genetic genealogy dataset D_2 or the identified dataset D_3.

The other variables in this scenario that need to be considered include the probability of an attack's success, $p(\boldsymbol{s})$, and adversary's estimated probability of an attack's success, $\hat{p}(\boldsymbol{s})$, which can be regarded as functions of the sharing strategy. It is expected that the attack will become less successful as the amount of shared data becomes smaller.

The Stackelberg game model in the masking game scenario can be represented in the extensive form, as shown in Figure 5.4. The payoff functions for both players, given a fixed sharing strategy $\boldsymbol{s}$, are calculated as follows. Given the shared data, if an adversary attacks a data subject, the data subject's payoff is $b(\boldsymbol{s}) - Lp(\boldsymbol{s})$, and the adversary's estimated payoff is $L\hat{p}(\boldsymbol{s}) - C$. Alternatively, if the adversary does not attack, the data subject's payoff is $b(\boldsymbol{s})$, and the adversary's payoff is zero.

The data subject's best sharing strategy ($\boldsymbol{s}^*$) and the adversary's corresponding best response (a^*), as the solution to the Stackelberg game, is shown in Equation 5.14:

$$\begin{gathered}(\boldsymbol{s}^*, a^*) = \underset{\boldsymbol{s}, a \in \phi(\boldsymbol{s})}{\operatorname{argmax}} (b(\boldsymbol{s}) - Lp(\boldsymbol{s})a)\,, \\ \phi(\boldsymbol{s}) = \left\{ a \,\middle|\, \underset{a}{\operatorname{argmax}} (L\hat{p}(\boldsymbol{s})a - Ca) \right\},\end{gathered} \tag{5.14}$$

in which $\phi(\boldsymbol{s})$ is the set of best responses of the adversary to the data subject's strategy $\boldsymbol{s}$. We refer to this type of Stackelberg games with a multi-stage re-identification attack as the Multi-Stage Re-Identification

Game (MSRIG). The solution to the MSRIG (either the opt-in game or the masking game) can be found by calculating the payoffs of all data subject's strategies, as shown in Figure 5.4 using a backward induction algorithm, which exhaustively searches the space. To break a tie, we assume that the data subject prefers a strategy that shares more data and that the adversary prefers not to attack.

The critical component of the calculation is to obtain an adversary's estimated probability of an attack's success, $\hat{p}(\boldsymbol{s})$, by simulating the process of a multi-stage attack, whose running time is dependent upon the specific attack in consideration. This process may be time-consuming because a state-of-the-art attack model is typically very complex.

Note that the data subject's strategy space will become even larger if strategies such as generalization (i.e., replacing a value with the name of a group it is in) [236] and noise addition (i.e., modifying a value by adding it with a random number) [127] are allowed. For example, if the value for age can be generalized to age groups according to a hierarchy of four levels: 35 → [30-39] → [18-39] → * (represents an age group that includes all ages), the strategy space will double in size. However, investigating these scenarios is out of the scope of this work.

5.2.3.4 Other Scenarios

To demonstrate our game theoretic protection models' advantages, we designed two additional variations of the masking game and four baseline scenarios for comparison. In each variation of the masking game, an additional constraint is added to the scenario. In the first variation, we assume that the data subject to choose a strategy so that a rational adversary will never attempt an attack. Whereas in the second variation, we constrain the genetic genealogy dataset does not exist such that the attack in the masking game has only one stage. In each baseline scenario, a data subject chooses a sharing strategy according to a different model. In the first two of them, the data subject always chooses a fixed sharing strategy. Whereas in the next two of them, the data subject randomly selects a sharing strategy. These scenarios are explained below in further detail with the introduction of their purposes.

5.2.3.4.1 Variations of the masking game scenario

In the *no-attack masking game*, we assume that the data subject is not allowed to choose any strategy that will make a rational adversary driven by economic incentives attack. To implement this game variation, we recorded the optimal strategy within the set of searched strategies, which lead to no attack. The purpose of this scenario is to examine the maximal benefit that the data subject can get while ensuring full protection of his or her data privacy.

In the *one-stage masking game*, we assume the genetic genealogy dataset does not exist in the game model. In other words, the re-identification attack only has one stage (i.e., direct linkage without inference). To implement this game variation, we forced the adversary in the game use no surname in the linkage stage. The purpose of this scenario is to test whether and how the first stage (i.e., the inference stage) of the two-stage re-identification attack is essential for the attack.

5.2.3.4.2 Baseline scenarios

In the *no-protection* scenario, the data subject always shares the entire data record. The purpose of this scenario is to test the maximal power of the two-stage re-identification attack.

In the *demographics-only* scenario, the data subject always shares only demographic attributes in the data record. The purpose of this scenario is to show the maximal power of the one-stage re-identification attack (i.e., without the inference stage).

In the *random opt-in* scenario, the data subject randomly decides to opt-in to share the entire record with a probability of 0.05. This probability is set according to the participation rate of GEDmatch, an online DNA comparison service provider, from customers of two major DTC-GT companies (i.e., 23andMe and AncestryDNA). GEDmatch has about 1.2 million users, and 23andMe and AncestryDNA has about 25 million customers in total by the end of 2019. This scenario aims to simulate how people are sharing targeted datasets in the real world.

In the *random masking* scenario, the data subject randomly decides to share each attribute in the record with a probability of 0.8. The expected number of shared attributes for each data subject in the last baseline scenario is the same as the optimal one in the masking game scenario. The last baseline scenario's purpose is to show the lower bound on the effectiveness of sharing partial data.

5.2.4 Search Approaches to Solve the Masking Game

Here, we introduce approaches to search for the optimal strategy in our game theoretic models. It is computationally challenging to search for the optimal strategy in the masking game because the strategy space expands exponentially as the number of attributes *m* increases. The backward induction algorithm, a brute force algorithm, could work well when m is relatively small. However, when m is relatively large, pruning techniques should be used to reduce the search space. Furthermore, heuristics could be used to find a locally optimal strategy quickly.

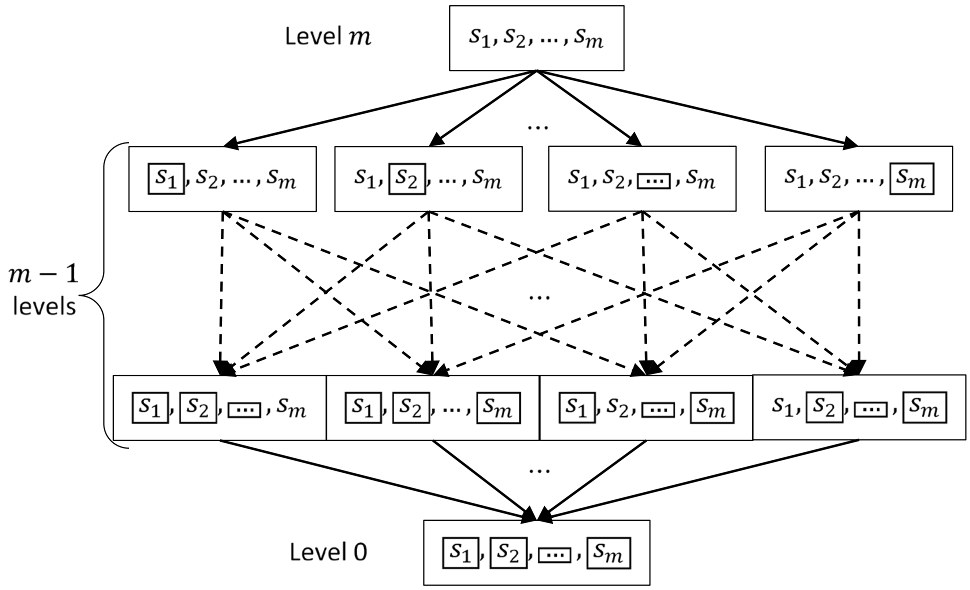

Level m
$s_1, s_2, \dots, s_m$
$\boxed{s_1}, s_2, \dots, s_m$
$s_1, \boxed{s_2}, \dots, s_m$
$s_1, s_2, \boxed{\cdots}, s_m$
$s_1, s_2, \dots, \boxed{s_m}$
$m-1$ levels
$\boxed{s_1}, \boxed{s_2}, \boxed{\cdots}, s_m$
$\boxed{s_1}, \boxed{s_2}, \dots, \boxed{s_m}$
$\boxed{s_1}, s_2, \boxed{\cdots}, \boxed{s_m}$
$s_1, \boxed{s_2}, \boxed{\cdots}, \boxed{s_m}$
Level 0
$\boxed{s_1}, \boxed{s_2}, \boxed{\cdots}, \boxed{s_m}$

Figure 5.5 The lattice representation of the data subject's strategy space for the masking game.

The node at the top represents the strategy of sharing all attributes, while the node at the bottom represents the strategy of sharing nothing. The level in which a node is located represents the number of shared attributes in the strategy the node represents, and it equals the number of "children" the node has. Each arrow points from a "parent" to a "child". Notably, *m* is the number of attributes in the data subject's record, and s_j represents the data subject's sharing decision for the j^{th} attribute.

The strategy space for a data subject in a masking game could be represented in a lattice structure, as shown in Figure 5.5. There are 2^m strategies to share a record with m attributes. That is to say, the lattice that represents the strategy space of sharing a record with 20 attributes has more than one million nodes. In the lattice, each strategy node represents a simulation of a complete two-stage attack. In the Gymrek attack, the adversary queries the genetic genealogy database of hundreds of thousands of records in the inference stage before querying the identified database of millions of records in the linkage stage. According to Gymrek et al., hours of manual investigation is required for one successful attack in practice. Even performing a demonstration attack or a simulated attack is time-consuming.

To accelerate the computation, we can search only part of the strategy space to accelerate the search process instead of search exhaustively. Furthermore, we equipped our program with several algorithms and heuristics and saved intermediate results in memory to prevent repeated calculations.

The game-solving system was implemented in Python 3.8.5 using several machine learning libraries (e.g., Numpy 1.19.1, Scikit-learn 0.23.2, Pandas 1.1.3, Matplotlib 3.3.1, Seaborn 0.11.0, and SciPy 1.5.2) managed by Anaconda3. All source code and all datasets in our experiments are accessible from https://hiplab.mc.vanderbilt.edu/projects/msrigame/.

In our experiments based on large-scale simulated datasets, by using a brute-force algorithm, searching for the best strategy for each data subject in each run of masking game took, on average, 11.5607 seconds on a machine with a six-core 64-bit central processing unit (CPU) clocked at 4.19 GHz, and a 32 GB random-access memory (RAM) clocked at 2,400 MHz.

5.2.4.1 Greedy Algorithm

As illustrated by Figure 5.5, the "child" of a sharing strategy $\boldsymbol{s}$ is defined as a strategy that shares one less attribute than the strategy $\boldsymbol{s}$ shares. As a result, the strategy at the bottom – sharing nothing – has no "children". A greedy algorithm searches the strategy space from the top. It keeps searching the "children" of the current best strategy, calculating their payoffs, and continues with a "child" that has the highest payoff. When several "children" have the same highest payoff, a "child" with the highest privacy measure will be selected to continue the search. The search stops at the bottom of the lattice. This algorithm can only find a locally optimal strategy.

In our experiments, by using the greedy algorithm, we reduced the average running time for each data subject in each run of the masking game by 98.33% to 0.1932 seconds. The globally optimal strategies for most data subjects are found. The data subjects' average payoff using the greedy algorithm was almost the same as the one using the brute-force algorithm, with a negligible difference of 2.63%.

5.2.4.2 Pruning Technique

Note that both the benefit of sharing $b(\boldsymbol{s})$ and the adversary's estimated probability of successful attack $\hat{p}(\boldsymbol{s})$ are functions of the sharing strategy ($\boldsymbol{s}$) and increase monotonically along with the increasing amount of shared data. In other words, if the strategy $\boldsymbol{s}'$ is a "child" of the strategy $\boldsymbol{s}$, then $b(\boldsymbol{s}') \leq b(\boldsymbol{s})$ and $\hat{p}(\boldsymbol{s}') \leq \hat{p}(\boldsymbol{s})$. Thus, if the payoff of the strategy $\boldsymbol{s}$ is lower than the current highest payoff, and if the adversary's corresponding best action is $\phi(\boldsymbol{s}) = \{0\}$ (i.e., the attacker decides not to attack), according to Figure 5.5, we can infer that payoffs of all "offspring" of the strategy $\boldsymbol{s}$ are also lower than the current highest payoff. To explain our pruning process in more detail, without loss of generality, let us make the following assumptions:

1) The sharing strategy $\boldsymbol{s}$ is sharing the first 12 attributes from a total of 16 attributes;

2) One of its "children" $\boldsymbol{s}'$ is sharing the first 11 attributes;

3) One of its "grandchildren" $\boldsymbol{s}''$ is sharing the first 10 attributes;

4) The current highest payoff is $\boldsymbol{v}^* = \$80$;

5) $b(\boldsymbol{s}) = \$75$;

6) The benefit of sharing each attribute is \$6.25; And

7) $\hat{p}(\boldsymbol{s}) = 1$ (i.e., there is only one matched identified record for the targeted data subject).

To apply the pruning, with the condition that $\phi(\boldsymbol{s}) = \{0\}$ and $v_d\big(\boldsymbol{s}, \phi(\boldsymbol{s})\big) < \boldsymbol{v}^*$, we need to show that $v_d\big(\boldsymbol{s}', \phi(\boldsymbol{s}')\big) < \boldsymbol{v}^*$ and $v_d\big(\boldsymbol{s}'', \phi(\boldsymbol{s}'')\big) < \boldsymbol{v}^*$. The demonstration process is as follows:

1) Because the "child" $\boldsymbol{s}'$ shares one less attribute than the strategy $\boldsymbol{s}$ shares, there can be more than one matched identified record for the targeted data subject, and thus $\hat{p}(\boldsymbol{s}') < \hat{p}(\boldsymbol{s})$;

2) Because the "grandchild" $\boldsymbol{s}''$ shares two fewer attributes, there can be even more matched identified records for the targeted data subject, and thus $\hat{p}(\boldsymbol{s}'') < \hat{p}(\boldsymbol{s}')$;

3) Because $\phi(\boldsymbol{s}) = \{0\}$, according to the definition of $\phi(\boldsymbol{s}) = \left\{a \,\middle|\, \underset{a}{\operatorname{argmax}}(L\hat{p}(\boldsymbol{s})a - Ca)\right\}$, we can infer $L\hat{p}(\boldsymbol{s}) - C \leq 0$;

4) Because $\hat{p}(\boldsymbol{s}'') < \hat{p}(\boldsymbol{s}') < \hat{p}(\boldsymbol{s})$, we have $L\hat{p}(\boldsymbol{s}'') - C \leq L\hat{p}(\boldsymbol{s}') - C \leq L\hat{p}(\boldsymbol{s}) - C \leq 0$, thus $\phi(\boldsymbol{s}'') = \phi(\boldsymbol{s}') = \phi(\boldsymbol{s}) = \{0\}$;

5) Because $\boldsymbol{s}'$ shares one less attribute, we have $b(\boldsymbol{s}') = \$68.75$;

6) Because $\boldsymbol{s}''$ shares two fewer attributes, we have $b(\boldsymbol{s}'') = \$62.5$;

7) Because $\phi(\boldsymbol{s}'') = \phi(\boldsymbol{s}') = \phi(\boldsymbol{s}) = \{0\}$, we have $v_d\big(\boldsymbol{s}'', \phi(\boldsymbol{s}'')\big) = v_d(\boldsymbol{s}'', 0) = b(\boldsymbol{s}'')$, $v_d\big(\boldsymbol{s}', \phi(\boldsymbol{s}')\big) = v_d(\boldsymbol{s}', 0) = b(\boldsymbol{s}')$, and $v_d\big(\boldsymbol{s}, \phi(\boldsymbol{s})\big) = v_d(\boldsymbol{s}, 0) = b(\boldsymbol{s})$; And

8) Because $b(\boldsymbol{s}'') < b(\boldsymbol{s}') < b(\boldsymbol{s})$, we can safely conclude that: $v_d\big(\boldsymbol{s}'', \phi(\boldsymbol{s}'')\big) < v_d\big(\boldsymbol{s}', \phi(\boldsymbol{s}')\big) < v_d\big(\boldsymbol{s}, \phi(\boldsymbol{s})\big) < \boldsymbol{v}^*$.

Thus, these "offspring" of the strategy $\boldsymbol{s}$ can be pruned from the search space. This pruning technique could be applied to both the brute force algorithm and the greedy algorithm to accelerate the search. When applied to the brute force algorithm, the globally optimal strategy could be found more quickly.

In our experiments, by applying this pruning technique, we reduced the average running time of the greedy algorithm by 41.72% to 0.1126 seconds, and we reduced the average running time of the brute force algorithm by 35.96% to 7.4029 seconds. We solved the masking games in the sensitivity analysis using the greedy algorithm with pruning to accelerate the computation.

5.3 Results

5.3.1 Experimental Design

To demonstrate our model and evaluate our methods' effectiveness, we conducted two sets of experiments based on genomic datasets. In the first set of experiments, we used real datasets composed of short tandem repeats on the Y-chromosome (Y-STRs) derived from Craig Venter's genomic record and the Ysearch dataset with 156,761 records and 100 Y-STRs, as they were used by Gymrek et al. [31]. To protect the privacy of the corresponding subjects and enable replications of our investigation, we sanitized the original datasets (i.e., modified for privacy protection) without affecting the demonstration (see **Appendix B.1** for the data sanitization details). In the second set of experiments, to evaluate the effectiveness of our methods in a larger and more controllable environment and to facilitate replications of our investigation without privacy concerns, we simulated a genetic genealogical population of 600,000 individuals (see **Appendix B.2** for details about the data preparation process), from which multiple datasets were sampled. To further evaluate our methods' effectiveness under various circumstances, we conducted a sensitivity analysis for eight parameters and three experimental settings. The default values for parameters for the experiments are provided in Table 5.1.

To measure the effectiveness, we calculated the average payoff for a pool of n data subjects, whose records may be shared in genomic dataset D_1, as: $\bar{V} = \sum_{i=1}^{n} V_i$, where V_i represents the i^{th} data subject's optimal payoff. We further calculated the average data utility ($\bar{U} = \sum_{i=1}^{n} U_i$) and the average privacy ($\bar{P} = \sum_{i=1}^{n} P_i$) of those data subjects to show how the game model can balance these two factors. In other words, we can calculate effectiveness measures (namely, the average payoff, the average data utility, and the average privacy) after obtaining the optimal payoff, the corresponding data utility, and the corresponding privacy for each data subject, given the best strategy ($\boldsymbol{s}_i^*$), the data record, and parameter settings. For simplicity, we assumed those data subjects use the same parameter settings. More specifically, the i^{th} data subject's data utility (U_i) is defined as the benefit of sharing divided by the benefit of sharing all data as follows: $U_i = b(\boldsymbol{s}_i^*)/B$. In addition, the i^{th} data subject's privacy (P_i) is defined as one minus the privacy risk (i.e., the probability to be successfully attacked) as: $P_i = 1 - p(\boldsymbol{s}_i^*)a_i^*$, in which $a_i^* \in \phi(\boldsymbol{s}_i^*)$ is the adversary's best response. Notably, the i^{th} data subject's optimal payoff (V_i) can be represented as a linear combination of the corresponding data utility and privacy: $V_i = v_d(\boldsymbol{s}_i^*, a_i^*) = BU_i - L(1 - P_i) = BU_i + LP_i - L$. As the primary measure of effectiveness, the average payoff of those data subjects is positively correlated with the metrics for utility and privacy, while these two metrics are negatively correlated with each other.

Our simulated population was generated with 20 attributes, including ID, surname, birth year, U.S. state of residence, and 16 genomic attributes. Based on the simulated population, we ran experiments to find the best sharing strategy for each subject in a targeted dataset in four game scenarios and four baseline scenarios. In each run of the experiments, from the simulated population, we randomly selected 1,000 records for targeted genomic dataset D_1, 20,000 records for identified dataset D_3, and 20,000 records for genetic genealogy dataset D_2. Based on these datasets, we compared eight scenarios: 1) no-protection, 2) demographics-only, 3) random opt-in, 4) random masking, 5) opt-in game, 6) masking game, 7) no-attack masking game, and 8) one-stage masking game. In all scenarios, the adversary aims to re-identify all records in dataset D_1, but he or she makes rational decisions according to their estimated payoff. Details about the datasets used in the experiments are summarized in Table 5.2.

Table 5.2 A summary of datasets used in experiments.

	Name of experiment set	**Large-scale effectiveness evaluation**	**Real-world demonstration**
Targeted genomic dataset (*D*₁)	**Name of dataset**	Simulated genomic datasets	Craig Venter's data
	Attributes	Birth year, state, 16 STRs	Age, state, 50 STRs
	Number of records	1,000	1
Genetic genealogy dataset (*D*₂)	**Name of dataset**	Simulated genetic genealogy datasets	Ysearch
	Attributes	Surname, 16 STRs	Surname, 50 STRs
	Number of records	20,000	58,218
Identified dataset (*D*₃)	**Name of dataset**	Simulated demographic datasets	PeopleFinders
	Attributes	ID, name, birth year, state	Name, age, state
	Number of records	20,000	About 250 million

STR stands for short tandem repeat.

5.3.2 Experiments based on a Large-scale Simulated Population

We ran the experiment 100 times using the backward induction algorithm with pruning and depicted the results in Figure 5.6 and Figure 5.7. Figure 5.6 displays a violin plot of the distributions of the data subjects' average payoffs in all eight scenarios, and Figure 5.7 displays a scatter plot of the data subjects' average privacy and utility in each scenario.

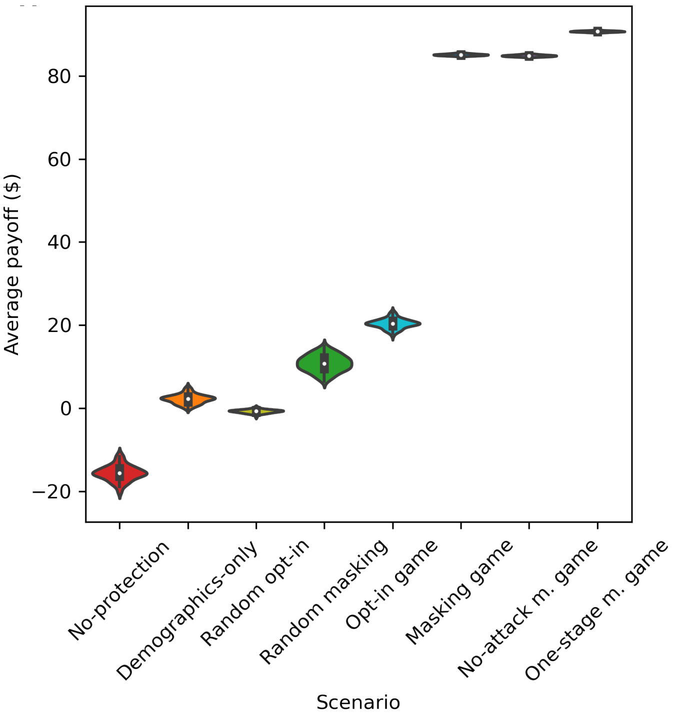


Figure 5.6 Violin plot of eight distributions of the data subjects' average payoffs in eight scenarios against a multi-stage re-identification attack targeting a dataset of 1000 subjects.

Each distribution corresponds to one scenario. The violin plot, depicted using the Seaborn package, combines boxplot and kernel density estimate for showing the distribution of data subjects' payoffs in each scenario. A Gaussian kernel is used with default parameter settings.

Several observations are worth highlighting. First, in Figure 5.6, it can be seen that the data subjects' average payoff is lowest in the no-protection scenario and highest in the masking game. Second, the data subjects' average payoff is improved substantially in the masking game, compared to that in the opt-in game. This observation illustrates one of the essential advantages of providing some degree of personalized choice in the data sharing process. Third, the masking game works better when the adversary uses fewer data resources and, thus, keeps fewer stages in the attack. Finally, a universal strategy, whether it is sharing all data or sharing demographics only, or a randomized strategy, brings a negative or negligible average payoff to the data subjects.

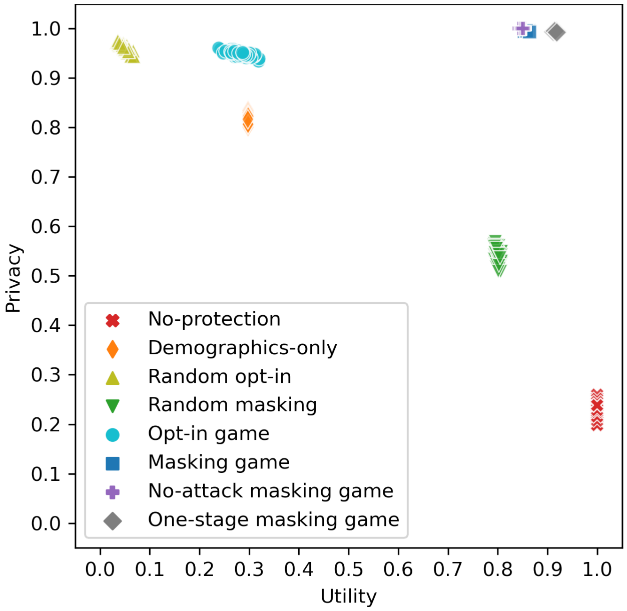


Figure 5.7 Scatter plot of data subjects' average privacy metrics and average utility metrics in eight scenarios against a multi-stage re-identification attack targeting a dataset of 1000 subjects. Each mark corresponds to one scenario and one run.

In Figure 5.7, the results representing the opt-in game are all in the plot's upper-left corner, which implies that the data subjects' strategies in this scenario tend to achieve high privacy but low utility. By contrast, the results representing the no-protection scenario are all in the plot's lower-right corner, which implies that the data subjects' strategies in this scenario tend to achieve high utility but low privacy. Only the results representing the masking game (and two of its variations) are in the plot's upper-right corner, where the data subjects' strategies achieve relatively high utility and high privacy at the same time. Notably, the data subject's strategies in the no-attack masking game guarantee full privacy protection with a substantial amount of shared data. In addition, a higher level of data utility is achieved when the attack has fewer stages. Nevertheless, the data subjects' strategies in rest scenarios are all worse than those in game scenarios. Specifically, compared to what the masking game does, the random opt-in scenario brings a data subject a similar privacy level but a much lower level of data utility. Additionally, the random masking strategy brings a similar level of data utility but a lower privacy level. Likewise, compared to what the opt-in game does, the demographics-only scenario brings a data subject a similar utility level but a lower privacy level.

Table 5.3 Effectiveness measures of protection scenarios against a multi-stage re-identification attack targeted the dataset in the first run of the experiments.

Notation	Description	Scenario							
		Baseline				Game			
		1	2	3	4	5	6	7	8
$\bar{V}$	Average payoff of data subjects	-$13.31	-$1.92	-$0.62	$14.19	$21.64	$85.22	$84.94	$90.63
σ_V	Standard deviation of data subjects' payoffs	$57.65	$33.38	$11.96	$60.72	$35.85	$7.23	$7.22	$5.78
$\bar{U}$	Average data utility of subjects	1.0	0.2980	0.043	0.7946	0.301	0.8599	0.8494	0.9153
σ_U	Standard deviation of subjects' data utility	0.0	0.0	0.2029	0.1161	0.4587	0.0746	0.0722	0.0623
$\bar{P}$	Average privacy of data subjects	0.2446	0.8141	0.9672	0.5649	0.9436	0.9948	1.0	0.9940
σ_P	Standard deviation of data subjects' privacy	0.3844	0.2225	0.1730	0.4359	0.1271	0.0194	0.0	0.0209

Scenarios include 1) no-protection scenario, 2) demographics-only scenario, 3) random opt-in scenario, 4) random masking scenario, 5) opt-in game scenario, and 6) masking game scenario, 7) no-attack masking game, and 8) one-stage masking game. We used background color to represent the scale of effectiveness measures. In each row of average payoffs, the highest value is shaded in blue, the lowest value is shaded in red, and the values in between the highest and the lowest are shaded in colors between blue and red. In each row of standard deviations, the highest value is shaded in red, the lowest value is shaded in blue, and the values in between the highest and the lowest are shaded in colors between red and blue.

The evaluation results in the first run of the experiments are reported in Table 5.3 with more statistics. It can be seen that the data subjects get a higher average payoff and higher average privacy when they make decisions based on game models. In addition, the data subjects' average payoff in the masking game is three times greater than the one in the opt-in game. The masking game also brings higher average utility and higher average privacy than the opt-in game does to data subjects. More specifically, about 30% of data is shared, and about 6% of data subjects are expected to be re-identified in the opt-in game. In contrast, about 86% of data is shared, and fewer than 1% of data subjects are expected to be re-identified in the masking game.

Compared to all baseline scenarios and the opt-in game, the masking game brings higher average payoff and average privacy with lower standard deviations to data subjects, which indicates that no matter how the data records are different from each other, the masking game guarantees greater protection for almost every data subject. By contrast, although the opt-in game generates a positive average payoff, a few data subjects' payoffs are still negative because the corresponding standard deviation is higher than the average. All baseline scenarios bring data subjects a negative or relatively low average payoff. Among all scenarios, the random masking scenario brings the highest standard deviation to data subjects in terms of payoff and privacy, while the opt-in game leads to the most diverse data utility among subjects.

Table 5.3 further illustrates the additional stage's contribution to the two-stage re-identification attack in terms of the chance of success. Note that in the demographics-only scenario, the adversary can only perform a one-stage attack. It can be seen that, in the no-protection scenario, about 76% of data subjects are expected to be re-identified, whereas, in the demographics-only scenario, about 19% of data subjects are expected to be re-identified. In other words, the additional stage raised the success rate of the attack by a factor of 4. Nevertheless, such a high re-identification rate implies that most targeted data subjects are uniquely identifiable from the identified dataset D_3. The additional stage's contribution is further demonstrated by comparing two masking games with a different number of attack stages. It can be observed that the game theoretic protection against the one-stage attack brings better averages and standard deviations with respect to both payoff and utility. Notably, the additional stage makes data subjects sacrifice about 6% more data utility to secure a clearly lower privacy risk, which corresponds to about 13% fewer re-identified data subjects.

Figure 5.8 shows the best strategies for the first 700 data subjects in the first run of the experiments in the masking scenarios. In the masking game, only a small portion of data is masked for most data subjects. Notably, two variations of the masking game bring either full privacy protection or better average payoff to those data subjects. In contrast, in the random masking game, the data subjects' average privacy risk is substantially higher, although the utility loss is almost the same as those in the game scenarios.

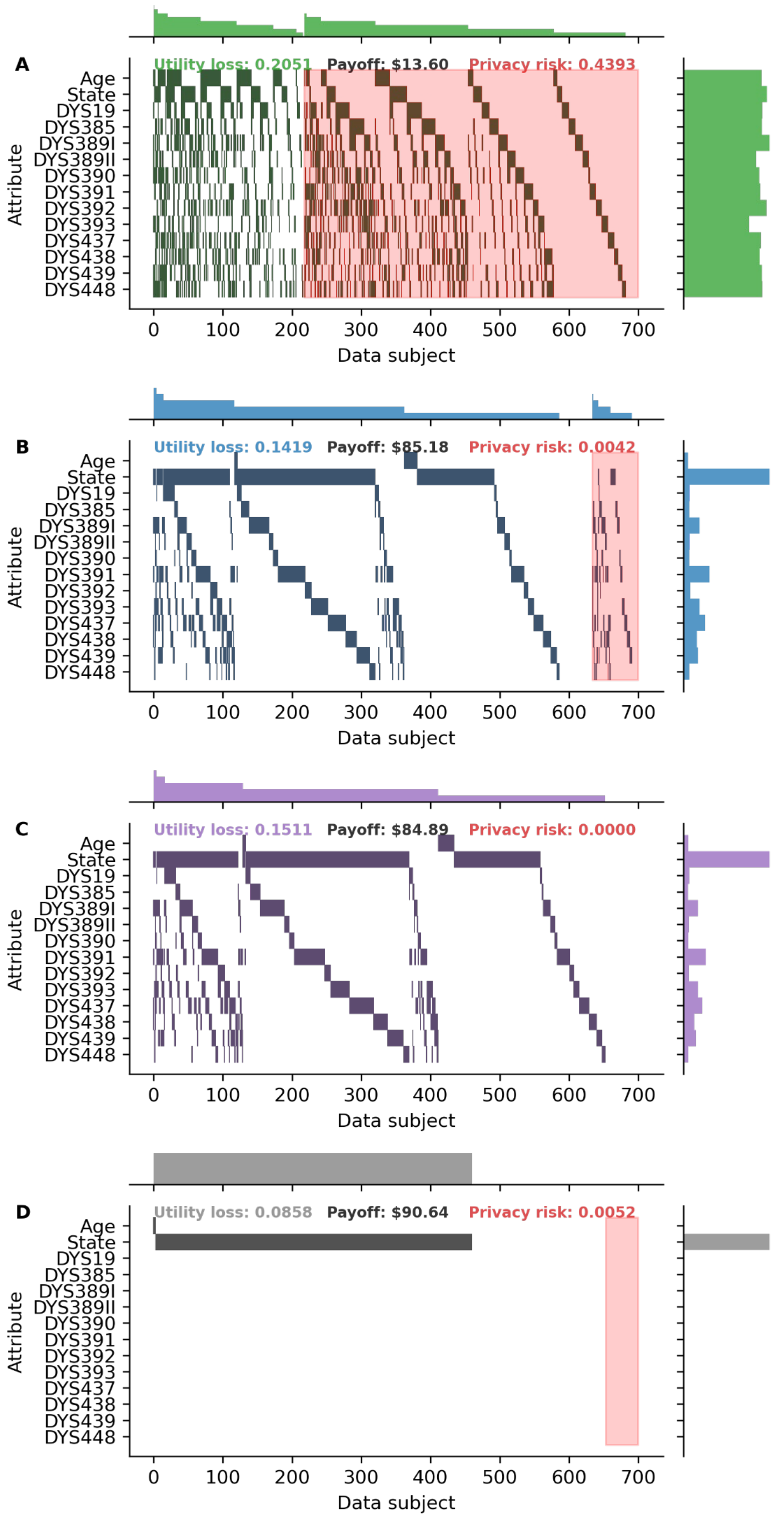


Figure 5.8 Best strategies for a group of the first 700 data subjects in the first run of experiments in the random masking scenario and three masking game scenarios.

(**A**) Random masking scenario. (**B**) Masking game. (**C**) No-attack masking game. (**D**) One-stage masking game. Each non-white block indicates that a data subject masks a specific attribute. Each row represents an attribute, and each column represents a data subject. The distribution for data subjects (and attributes) are summarized in the histogram on the top (on the right) with the number of bins equals to the number of data subjects (attributes). Data subjects are split into two groups: those on the left will not be attacked. Data subjects who will be attacked are on the right and covered by a red region. Each column (or data subject) is sorted within each group by the number of masked attributes in descending order. Columns (or data subjects) within each group are sorted by the number of masked attributes in descending order. The order of rows (or attributes) is the same as the order of attributes in the dataset. For each scenario, the average payoff, average utility loss, and average privacy risk are presented at the upper-center location, the upper-left corner, and the upper-right corner, respectively. For each data subject, the utility loss is defined as one minus the data utility measure, and the privacy risk is defined as one minus the privacy measure.

5.3.3 Sensitivity Analysis on Parameters and Settings Using Simulated Datasets

To test the model's sensitivity to eight parameters and three experimental settings, we compared effectiveness measures in eight scenarios with one parameter or setting changed using a targeted dataset of 200 data subjects in 20 runs of experiments. Sensitivity analysis of more parameters and settings could be potentially conducted in the same way.

5.3.3.1 Sensitivity Analysis on Parameters Using Simulated Datasets Regarding Payoff

The sensitivity analysis results on eight parameters regarding the data subjects' average payoff are shown in the upper body of Figure 5.9. Generally, regardless of how the targeted parameter varies, the data subjects' average payoff in the masking game is much higher than their average payoff in other scenarios (except two variations of the masking game). Specifically, the one-stage variation brings more average payoff, and the no-attack variation guarantees full privacy protection.

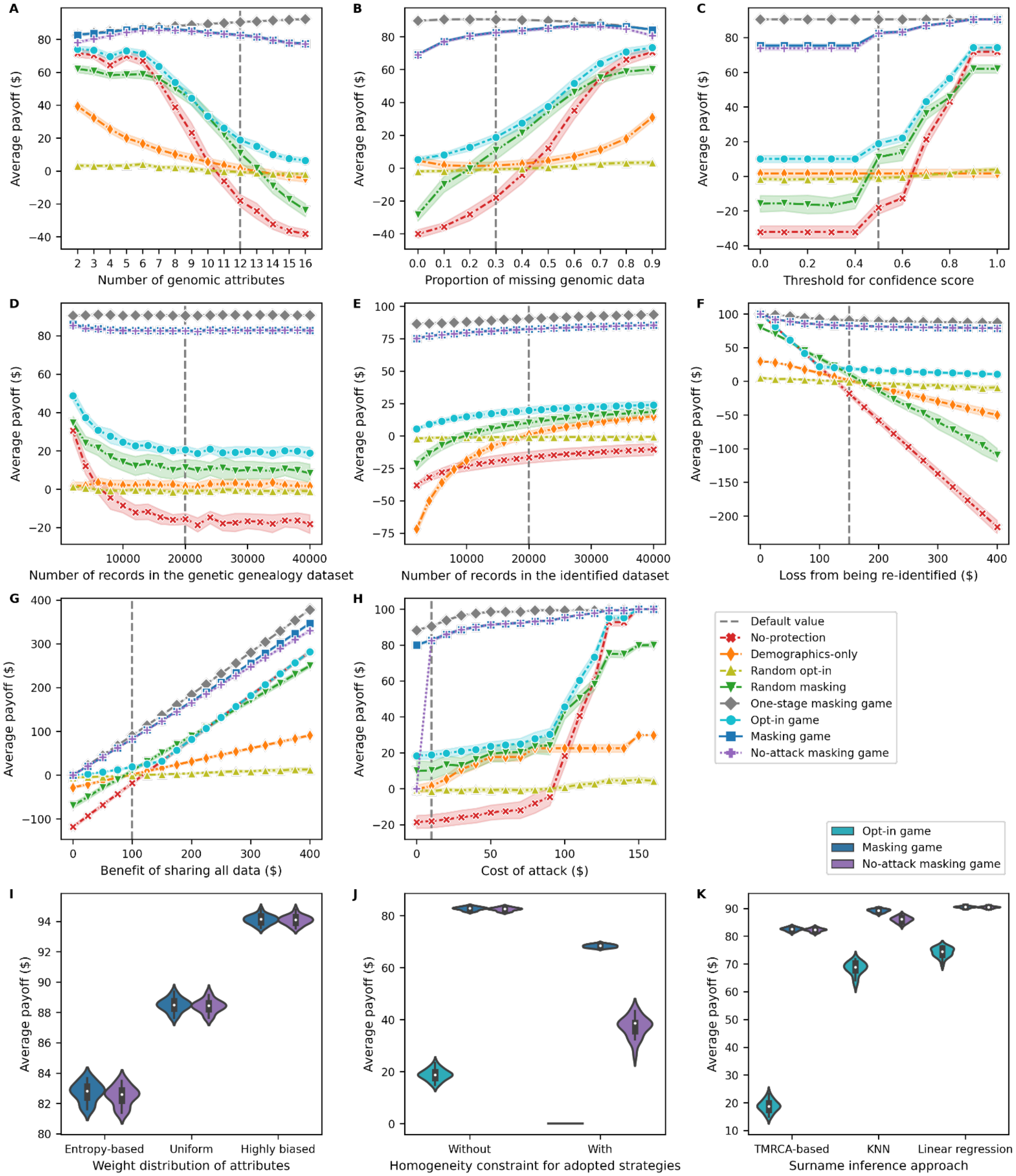


Figure 5.9 Sensitivity of the data subjects' average payoff to parameters and settings.

(**A**) Line plot of payoff's sensitivity to the number of genomic attributes. (**B**) Line plot of payoff's sensitivity to the proportion of missing genomic data. (**C**) Line plot of payoff's sensitivity to the threshold for confidence score. (**D**) Line plot of payoff's sensitivity to the number of records in the genetic genealogy dataset. I Line plot of payoff's sensitivity to the number of records in the identified dataset. (**F**) Line plot of payoff's sensitivity to the loss from being re-identified. (**G**) Line plot of payoff's sensitivity to the benefit of sharing all data. (**H**) Line plot of payoff's sensitivity to the cost of an attack. (**I**) Violin plot of payoff distribution's sensitivity to the strategy adoption setting. (**J**) Violin plot of payoff distribution's sensitivity to the surname inference approach. (**K**) Violin plot of payoff distribution's sensitivity to the weight distribution of attributes. Each line plot, depicted using the Seaborn package, shows the data subjects' average payoffs, with error bars representing standard deviations, in a scenario. The violin plot, depicted using the Seaborn package, combines boxplot and kernel density estimate for showing the distribution of data subjects' payoffs in each scenario. A Gaussian kernel is used with default parameter settings. TMRCA stands for time to most recent common ancestor. KNN means the k-nearest neighbors algorithm.

5.3.3.1.1 Number of genomic attributes

Figure 5.9A shows how the data subjects' average payoff changes as the number of genomic attributes on the Y-chromosome varies from 2 to 16. To ensure comparability, we fix the benefit of sharing all data B in this experiment and thus make the benefits of sharing each attribute B_j $(j = 1, \cdots, m)$ change according to the examined number of genomic attributes m. First, we can observe in Figure 5.9A that there is little change in the data subjects' average payoff in the masking game (less than 5% of the maximal change as in the no-protection scenario), which indicates that the data subjects' average payoff is not sensitive to the number of genomic attributes in the masking game. Second, the data subjects' average payoff decreases substantially as the number of genomic attributes increases in all baseline scenarios and the opt-in game. This trend occurs because the accuracy of the surname-inference attack increases as more genomic data is shared (for all scenarios except the demographics-only scenario and the one-stage masking game). For the demographics-only scenario, the underlying reason is that the utility of the shared demographic data decreases as more genomic data is shared. Third, the difference between the opt-in game and the no-protection scenario in terms of the data subjects' average payoff decreases as the number of genomic attributes increases. This difference is because the data subject has a greater chance to opt-out when more genomic attributes are available in the opt-in game.

5.3.3.1.2 Proportion of missing genomic data

The genetic genealogy dataset D_2 may have missing values (or missing data), especially in real-world cases. For example, 26% values for genomic attributes are missing values in the Ysearch dataset (after filtering out records with too few Y-STR markers). Note that all data subjects' surnames are included in the dataset. We denoted by γ the proportion of missing genomic data in dataset D_2.

Figure 5.9B shows how the data subjects' average payoff changes as the proportion of missing genomic data in dataset D_2 varies from 0 to 0.9 (in increments of 0.1). It can be seen that there is little change in the data subjects' average payoff in the masking game (less than 14% of the maximal change as in the no-protection scenario). Moreover, the data subjects' average payoff increases as the missing proportion increases in all scenarios save the one-stage masking game. This trend occurs because data missing from dataset D_2 can be considered as the result of a randomized masking strategy, which reduces the privacy risk but does not affect data utility.

5.3.3.1.3 Threshold for the confidence score

In the Gymrek attack, a confidence score is used to measure the adversary's confidence that an attack's surname-inference stage will be successful. A threshold for confidence score is required for an adversary to make the attack decision.

Figure 5.9C shows how the data subjects' average payoff changes when the confidence score threshold varies from 0 to 1 (in increments of 0.1). It can be seen that there is little change in the data subjects' average payoff in the masking game (less than 15% of the maximal change as in the no-protection scenario). In addition, the data subjects' average payoff increases as the threshold increases in all scenarios save the demographics-only scenario and the one-stage masking game. This trend occurs because the adversary is unlikely to infer a surname with a high threshold for confidence score, making the attack less successful.

5.3.3.1.4 Number of records in the genetic genealogy dataset

Figure 5.9D shows how the data subjects' average payoff changes as the number of records in the genetic genealogy dataset D_2 is varied from 2,000 to 40,000 (in increments of 2,000). It can be seen that there is little change in the data subjects' average payoff in the masking game (less than 7% of the maximal change as in the no-protection scenario). It should also be recognized that the data subjects' average payoff decreases as the number of records in the genetic genealogy dataset increases in all scenarios save the demographics-only scenario, the random opt-in scenario, and the one-stage masking game. This trend occurs because the adversary is more likely to find someone sharing similar genomes with the targeted data subject in a larger genetic genealogy dataset, making the attack more likely to be successful.

5.3.3.1.5 Number of records in the identified dataset

We assume the adversary in the re-identification attack links the targeted data subject to only one matched identified record, following the marketer re-identification risk model [163]. As a result, a larger identified dataset increases the expected number of matched identified records and, thus, reduces the probability of the attack's success for each targeted data subject.

Figure 5.9E shows how the data subjects' average payoff changes as the number of records in identified dataset D_3 is varied from 2,000 to 40,000 (in increments of 2,000). It can be seen that there is little change in the data subjects' average payoff in the masking game (less than 12% of the maximal change as in the demographics-only scenario). In addition, the data subjects' average payoff increases as the number of records in the identified dataset increases in all scenarios save the random opt-in scenario. This trend occurs because a larger identified dataset reduces the likelihood of the attack's success. Thus, when a data subject can choose to share data, they will be more likely to do so if the identified dataset has more records than not.

5.3.3.1.6 Loss from being re-identified

Figure 5.9F shows how the data subjects' average payoff changes when the financial loss associated with being re-identified varies from 0 to $400 (in increments of $25). Note that we assume that the adversary's gain from re-identification is always equal to the data subject's loss from being re-identified. It can be seen that, in all baseline scenarios, the data subjects' average payoff decreases linearly as the loss from re-identification increases. Further, in the opt-in game, the data subjects' average payoff keeps decreasing linearly until the loss reaches $100, which is equal to the benefit of sharing all data. This trend manifests for two reasons. First, in this game, the data subject always chooses to share data until their loss surpasses the benefit of sharing data. Second, an adversary is not able to attack the data subjects whose data is not shared. Note that, in the masking game (and two of its variations), the data subjects' average payoff decreases as the loss from being re-identified increases. This trend occurs because the loss indicates that the impact of a successful attack on a data subject is negative. However, the curve is relatively flat (the range of change in the masking game is less than 7% of the maximal change as in the no-protection scenario), which implies that the protection strategy works well.

5.3.3.1.7 Benefit of sharing all data

Figure 5.9G shows how the data subjects' average payoff changes when the benefit of sharing all data varies from 0 to $400 (in increments of $25). It can be seen that in all scenarios (except the opt-in game), the data subjects' average payoff linearly increases as the benefit of sharing all data increases. This trend occurs because the data subjects' average payoff is a positively correlated linear function of the benefit of sharing all data, which is unavoidable. In addition, in the opt-in game, the data subjects' average payoff starts to linearly increase after the benefit of sharing all data surpasses $150, which is equal to the loss from being re-identified. This trend occurs because a data subject tends to share data if their benefit surpasses the loss from sharing data.

5.3.3.1.8 Cost of an attack

Figure 5.9H shows how the data subjects' average payoff changes when the adversary's cost to execute an attack varies from 0 to $160 (in increments of $10). It can be seen that, in all scenarios, the data subjects' average payoff increases as the cost of an attack increases. This trend implies that, in addition to data masking, raising an adversary's cost (e.g., penalizing the adversary for privacy breach) can effectively deter adversaries' attacks. However, the curve is relatively flat (the range of change in the masking game is less than 17% of the maximal change as in the no-protection scenario), which implies that the protection strategy works well.

5.3.3.2 Sensitivity Analysis on Settings Using Simulated Datasets Regarding Payoff

The sensitivity analysis results on the three experimental settings (assumptions) regarding the data subjects' average payoff are shown at the bottom of Figure 5.9. Generally, regardless of how the targeted experimental setting (assumption) varies, the data subjects' average payoff in the masking game is much higher than their average payoff in other scenarios (except the no-attack masking game).

5.3.3.2.1 Weight distribution of attributes

In the masking game, each attribute in the targeted dataset has a certain weight in calculating the data utility. In the current setting, an attribute's weight is proportional to its information entropy in either the genetic genealogy dataset or the identified dataset. In this experiment, we examine two alternative settings of weight distribution to investigate the degree to which the weight distribution affects a data subject's optimal strategy and payoff. In the more balanced setting, every attribute has the same weight, while in the more biased setting, the weight of the first two genomic attributes is ten times the weight of all other attributes.

Figure 5.9I displays a violin plot that shows how the weight distribution of attributes changes the data subjects' average payoff in the masking game. It can be seen that, in the masking game (and its no-attack variation), the data subjects' average payoff will change if the weights of attributes are changed. Specifically, the further the attributes' weights are set away from their information entropies, the higher the average payoff the data subjects can obtain. This trend occurs because the attributes in a dataset that contain more information than others can be used to uniquely identify a record more easily. In other words, if any of these attributes are assigned with a weight that is smaller than its entropy, it is more likely to be masked without reducing too much utility but reducing the privacy risk significantly, and thus raises the data subjects' average payoff.

5.3.3.2.2 Homogeneity constraint for adopted strategies

In certain situations, a group of data subjects is required to adopt the same strategy. These situations happen when an agent chooses the same sharing strategy for a group of data subjects, or when those data subjects in a group make decisions interdependently.

Figure 5.9J displays a violin plot that shows how the homogeneity of adopted strategies influences the data subjects' average payoff in three games. It can be seen that, in all games, the data subjects' average payoff will decrease if those data subjects in a dataset are required to adopt the same strategy. In other words, the data subjects need to trade their average payoffs for the homogeneity of adopted strategies. By comparing the masking game with the opt-in game, it can be seen that the data subjects' average payoff with the homogeneity constraint is still relatively high in the masking game, but it is always as low as zero in the opt-in game. This difference may be caused by the different flexibilities of these two games.

5.3.3.2.3 Surname inference approach

The adversary's approach to infer surnames affects the accuracy of the inference and, thus, affects the data subjects' payoffs. In the Gymrek attack, a surname is inferred by finding the nearest neighbor with the Time to Most Recent Common Ancestor as the distance measure. However, the surname associated with a genomic record could be inferred using off-the-shelf machine learning approaches as well, and the accuracy of the inference might be acceptable. In the context of machine learning, a genealogy dataset can be treated as a training set in which surnames are labels. After applying two machine learning approaches (namely, k-nearest neighbors (KNN) and linear regression) to the surname inference, we tested our protection methods against the two-stage attack in which a surname is inferred using one of these approaches.

Figure 5.9K displays a violin plot that shows how the data subjects' average payoff changes according to the surname inference approach. First, we calibrated the parameters in these two machine learning models. Afterward, we set the confidence score to 1 for all compared approaches to ensure comparability. It can be seen that, for each game, attacks based on machine learning approaches result in higher average payoff for data subjects compared with the original surname inference approach, in which the KNN approach works better for the adversary (i.e., brings lower average payoff to the data subjects). This is because these off-the-shelf machine learning approaches are not customized to the surname inference problem, and the KNN approach has fewer parameters and thus is easier to be calibrated.

5.3.3.3 Sensitivity Analysis on Parameters Using Simulated Datasets Regarding Privacy and Utility

The sensitivity analysis results on eight parameters regarding the data subjects' average privacy and average utility are shown in Figure 5.10 and Figure 5.11, respectively. Generally, in all experiments, the data subjects' average privacy in the masking game is almost always higher than their average privacy in the other scenarios except two variations of the masking game. In addition, the data subjects' average utility in the masking game is only lower than their average utility in the no-protection scenario and the one-stage masking game.

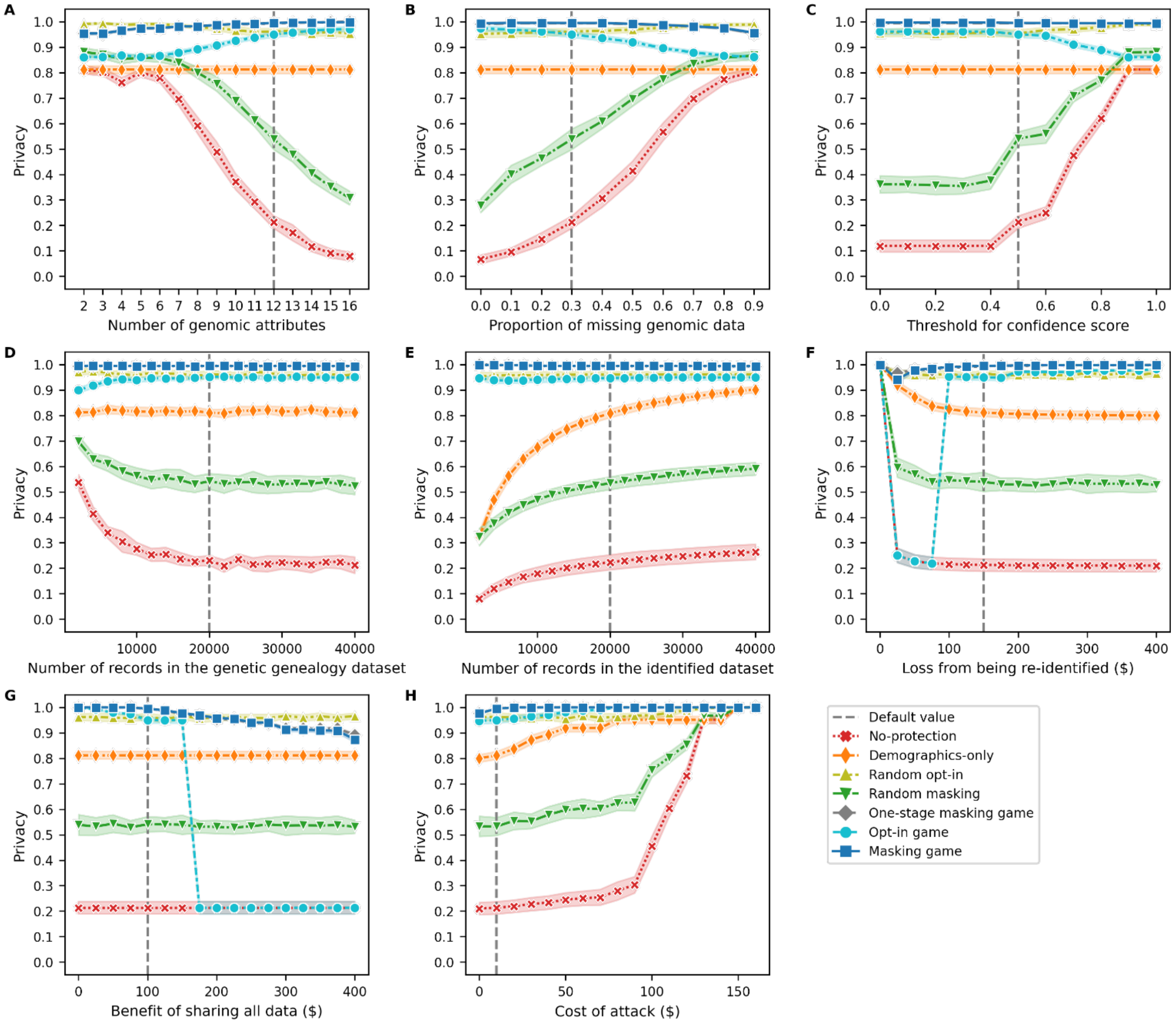


Figure 5.10 Sensitivity of the data subjects' average privacy to parameters in seven scenarios.

(**A**) Line plot of payoff's sensitivity to the number of genomic attributes. (**B**) Line plot of payoff's sensitivity to the proportion of missing genomic data. (**C**) Line plot of payoff's sensitivity to the threshold for confidence score. (**D**) Line plot of payoff's sensitivity to the number of records in the genetic genealogy dataset. I Line plot of payoff's sensitivity to the number of records in the identified dataset. (**F**) Line plot of payoff's sensitivity to the loss from being re-identified. (**G**) Line plot of payoff's sensitivity to the benefit of sharing all data. (**H**) Line plot of payoff's sensitivity to the cost of an attack. Each line plot, depicted using the Seaborn package, shows the data subjects' average payoffs, with error bars representing standard deviations, in a scenario.

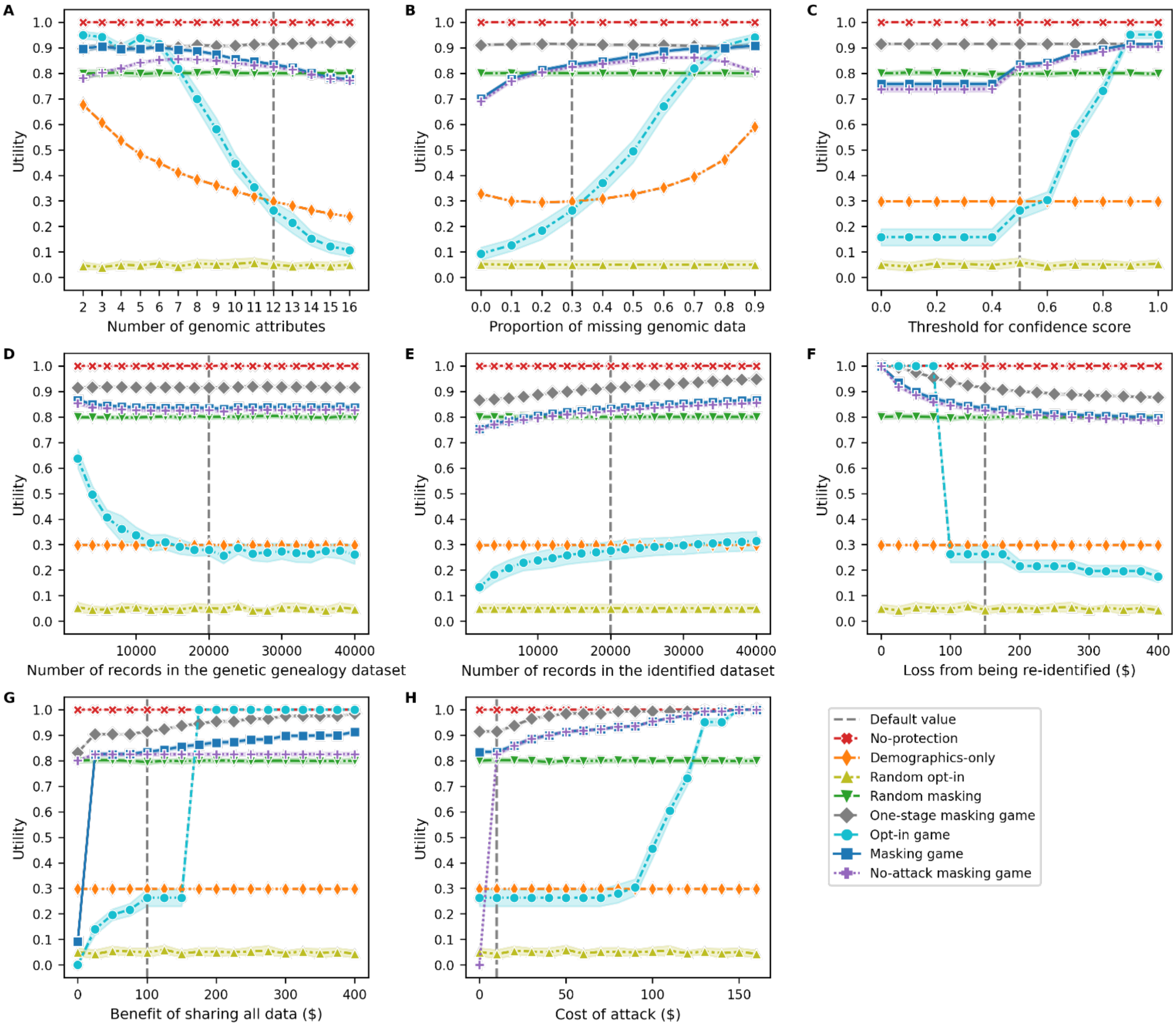


Figure 5.11 Sensitivity of the data subjects' average data utility to parameters in eight scenarios.

(**A**) Line plot of payoff's sensitivity to the number of genomic attributes. (**B**) Line plot of payoff's sensitivity to the proportion of missing genomic data. (**C**) Line plot of payoff's sensitivity to the threshold for confidence score. (**D**) Line plot of payoff's sensitivity to the number of records in the genetic genealogy dataset. I Line plot of payoff's sensitivity to the number of records in the identified dataset. (**F**) Line plot of payoff's sensitivity to the loss from being re-identified. (**G**) Line plot of payoff's sensitivity to the benefit of sharing all data. (**H**) Line plot of payoff's sensitivity to the cost of an attack. Each line plot, depicted using the Seaborn package, shows the data subjects' average payoffs, with error bars representing standard deviations, in a scenario.

5.3.3.3.1 Number of genomic attributes

In Figure 5.10A and Figure 5.11A, we can see that changes in data subjects' average privacy and average utility in the masking game are mild. Notably, privacy increases (and utility decreases) monotonically as the number of genomic attributes in the opt-in game increases. In addition, as the number of genomic attributes increases, privacy increases (and utility decreases) in the demographics-only scenario. Moreover, as the number of genomic attributes increases, the utility remains unchanged (and privacy decreases) in the random opt-in and random masking scenarios. Finally, in the no-protection scenario, the utility does not change (and privacy decreases drastically) as the number of genomic attributes increases. It is because the accuracy of a surname-inference attack increases with more genomic data being shared.

5.3.3.3.2 Proportion of missing genomic data

In Figure 5.10B, we can see that the data subjects' average privacy in the masking game remains almost unchanged when the proportion of missing data changes. In contrast, in the no-protection scenario and the random masking scenario, as the proportion of missing data increases, the data subjects' average privacy increases monotonically because the surname inference becomes less accurate in this situation. However, in the opt-in game, the data subjects' average privacy decreases as the proportion of missing data increases.

At the same time, in Figure 5.11B, we can see that the data subjects' average utility increases monotonically as the proportion of missing data increases in both opt-in and masking games. This trend implies that more data should be shared if the adversary has access to less external data.

5.3.3.3.3 Threshold for the confidence score

In Figure 5.10C, we can see that the data subjects' average privacy in the masking game remains unchanged when the threshold for confidence score changes. In contrast, in the no-protection scenario and the random masking scenario, as the threshold for confidence score increases, the data subjects' average privacy increases monotonically. It is because fewer surnames can be inferred in this situation. However, in the opt-in game, the data subjects' average privacy decreases as the threshold for confidence score increases.

At the same time, in Figure 5.11C, we can see that the data subjects' average utility increases monotonically as the threshold for confidence score increases in both opt-in and masking games. This trend implies that more data should be shared if the adversary is more conservative on attacks.

5.3.3.3.4 Number of records in the genetic genealogy dataset

In Figure 5.10D, we can see that the data subjects' average privacy in the masking game remains unchanged when the number of records in the genetic genealogy dataset changes. In contrast, in the no-protection scenario and the random masking scenario, as the number of records in the genetic genealogy dataset increases, the data subjects' average privacy decreases monotonically. It is because more surnames can be inferred correctly. However, in the opt-in game, the data subjects' average privacy increases as the number of records in the genetic genealogy dataset increases.

At the same time, in Figure 5.11D, we can see that the data subjects' average utility decreases monotonically as the number of records in the genetic genealogy dataset increases in both opt-in and masking games. This trend implies that fewer data should be shared if the adversary has access to a larger genetic genealogy dataset.

5.3.3.3.5 Number of records in the identified dataset

In Figure 5.10E, we can see that the data subjects' average privacy in all games remains almost unchanged when the number of records in the identified dataset changes. In contrast, in the no-protection scenario, the demographics-only scenario, and the random masking scenario, as the number of records in the identified dataset increases, the data subjects' average privacy increases monotonically because more identities can be linked to each target.

At the same time, in Figure 5.11E, we can see that the data subjects' average utility increases monotonically as the number of records in the identified dataset increases in all games. This trend implies that more data should be shared if the adversary has access to a larger identified dataset.

5.3.3.3.6 Loss from being re-identified

From Figure 5.10F and Figure 5.11F, we can see that the data subjects' average privacy, in all baseline scenarios instead of four game scenarios, remains almost unchanged when the loss is greater than $25. It is because the cost of an attack is too low ($10) to let the loss from being re-identified affect the effectiveness of the attack if the loss is greater than the cost. In contrast, in the opt-in game, although utility does not increase (and privacy does not decrease) as the loss increases when the loss is greater than $25, both utility and privacy change dramatically when the loss reaches $100, which is equal to the benefit of sharing all data. Similarly, in the masking game (and its no-attack variation), privacy increases (and utility decreases) substantially until the loss is greater than $100.

5.3.3.3.7 Benefit of sharing all data

From Figure 5.10G and Figure 5.11G, we can see that the data subjects' average privacy and utility, in all baseline scenarios instead of four game scenarios, almost remain the same because the effectiveness of attack is not related to the benefit of sharing all data. In contrast, in the opt-in game, although privacy does not increase (and utility does not decrease) as the benefit increases, utility and privacy drastically change when benefit surpasses $150, which is equal to the loss from being re-identified. However, in the masking game (and its one-stage variation), privacy slowly decreases (and utility slowly increases) when the benefit of sharing all data is greater than $25.

5.3.3.3.8 Cost of an attack

From Figure 5.10H and Figure 5.11H, we can see that, in all scenarios, privacy increases (and utility does not decrease) as the cost of an attack increases. In addition, they reach the maximal value when the cost of an attack reaches $150, which is equal to the maximal gain of the adversary from re-identification (i.e., the data subject's loss from being re-identified). This is because the increase in the cost of an attack disincentivizes an adversary to attack and does not negatively impact the data utility. This suggests that alternative deterrents (e.g., penalties in data use agreements), if invoked, can increase the cost of an attack and so mitigate the privacy risk as well.

5.3.4 Experiments based on Craig Venter's data and the Ysearch dataset

In this set of experiments, we used a case study to illustrate how our model could be applied to real-world datasets. Specifically, we used Craig Venter's demographic attributes (in terms of birth year, state of residence, and gender) and 50 Y-STRs profiled from his genome sequence as one of the data subjects' records in dataset D_1. In addition, we used the Ysearch dataset as the genetic genealogy dataset D_2. After filtering out records with too few targeted Y-STRs, we got a dataset of 58,218 records and 50 Y-STRs with a missing proportion of approximately 26%. Finally, we used Intelius.com and the 2010 US Census as sources of dataset D_3.

A query of Craig Venter's demographic attributes and the surname on Intelius.com returned with two records, one of which is Craig Venter. According to the 2010 US Census, 157,681 people were estimated to share the same values on the three demographic attributes with Craig Venter in the US in 2018. The number indicates that, without the correctly inferred surname, the re-identification would hardly be successful. For simplicity, we masked only genomic attributes instead of all attributes in this demonstrational case study. Parameters in both sets of experiments were set in the same way. For example, the data utility corresponding to each genomic attribute in this set of experiments was set according to its information entropy in the Ysearch dataset. Settings for all other parameters are reported in Table 5.1. We ran the experiment using the greedy algorithm with pruning and presented the results in Table 5.4. In addition to showing those effectiveness measures of recommended strategies from eight scenarios, it shows effectiveness measures of all searched strategies in the masking game (with the best strategy being marked out).

Table 5.4 Results for the case study based on Craig Venter's data and the Ysearch dataset.

Searched suboptimal strategy	# of shared Y-STRs	Inferred surname	Confidence score	Utility	Privacy	Benefit	Loss	Payoff	Scenario recommends
0	50	Venter	0.6688	1	0.5	$100	$75	$25	1
1	49	Venter	0.6688	0.9967	0.5	$99.67	$75	$24.67	/
2	48	Venter	0.6688	0.9933	0.5	$99.33	$75	$24.33	/
3	47	Venter	0.6688	0.9897	0.5	$98.97	$75	$23.97	/
4	46	Venter	0.6688	0.9862	0.5	$98.62	$75	$23.62	/
5	45	Venter	0.6688	0.9825	0.5	$98.25	$75	$23.25	/
6	44	Venter	0.6688	0.9786	0.5	$97.86	$75	$22.86	/
7	43	Venter	0.6435	0.9746	0.5	$97.46	$75	$22.46	/
8	42	Venter	0.5484	0.9706	0.5	$97.06	$75	$22.06	/
9	41	Venter	0.5481	0.9661	0.5	$96.61	$75	$21.61	/
10*	40	Karlsson	0.4662	0.9593	1	$95.93	$0	$95.93	6, 7
11	39	Karlsson	0.4661	0.9544	1	$95.44	$0	$95.44	/
/	0	/	/	0.44	0.00	$44	$0.00	$44.00	2
/	0	/	/	0	1	$0	$0	$0	3, 5
/	38	Venter	0.6402	0.8716	0.5	$87.16	$75	$12.16	4
/	50	/	/	1	0.00	$100.00	$0.00	$100.00	8

The best strategy in the masking game, found using the greedy algorithm with pruning, is marked by a star (*). Scenarios include 1) no-protection scenario, 2) demographics-only scenario, 3) random opt-in scenario, 4) random masking scenario, 5) opt-in game scenario, and 6) masking game scenario, 7) no-attack masking game, and 8) one-stage masking game. Only genomic attributes could be masked in this case study. Y-STR stands for Y-chromosome short tandem repeat.

From Table 5.4, it can be seen that the best strategy in the masking game is significantly superior to not only other searched strategies in the game but also all recommended strategies from all baseline scenarios and the opt-in game in terms of the resulting data subject's payoff. The underlying reason is that the first strategy during the search process with a confidence score below the threshold of 0.5 has an incorrectly inferred surname, which leads to an unsuccessful re-identification and hence successful protection.

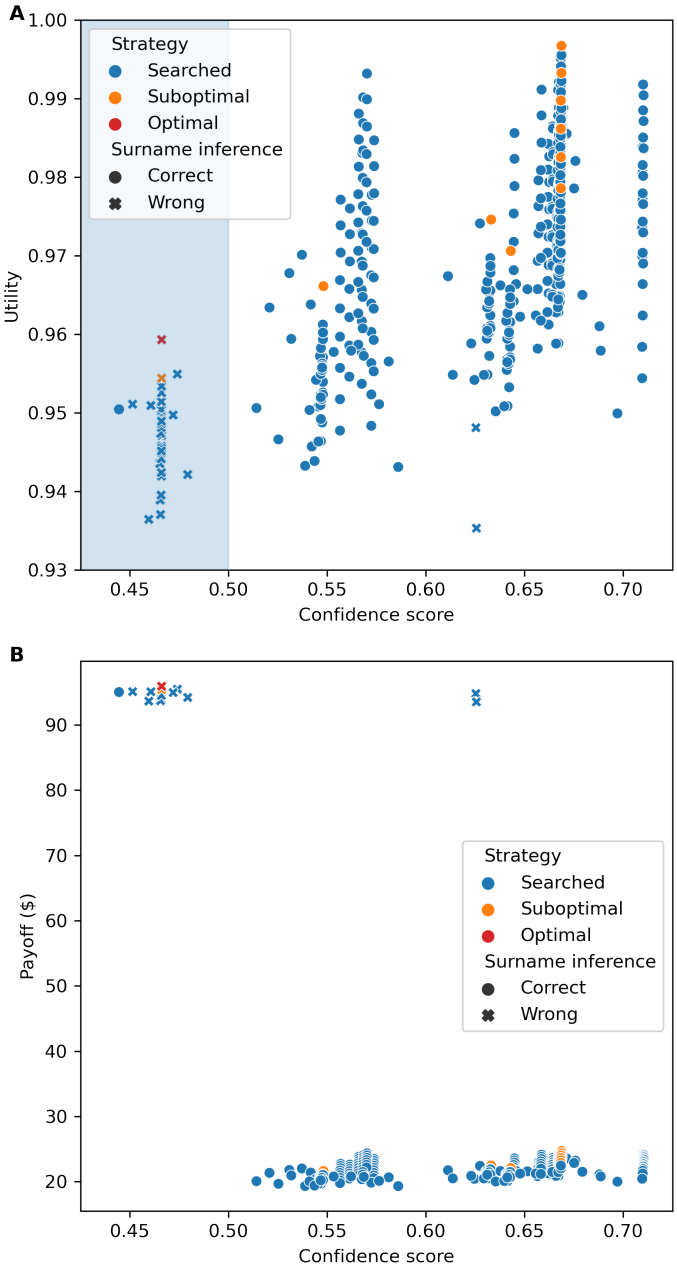


Figure 5.12 Scatter plots of measures associated with all searched strategies in the masking game in the case study based on Craig Venter's data and the Ysearch dataset.

(**A**) Craig Venter's data utility and confidence score associated with each strategy. (**B**) Craig Venter's payoff and confidence score associated with each strategy. The strategy with the highest utility in the region shaded in blue is the optimal strategy. If there is no strategy in that region, the optimal strategy is the strategy that shares all data (i.e., no-protection). A circle mark indicates that the inferred surname is correct, and a cross mark indicates that the inferred surname is incorrect, or no surname is inferred.

Here, we analyze the relationships between three evaluation metrics (namely, utility, payoff, and inference correctness) and confidence score in the case study based on Craig Venter's data and the

Ysearch dataset. Figure 5.12 displays two scatter plots of those measures associated with all searched strategies using the greedy algorithm with pruning in the case study. In each subplot, the horizontal axis indicates the adversary's confidence score (ranging from 0 to 1) for the inferred surname resulting from one of the data subject's strategies. The vertical axis in Figure 5.12A indicates the data subject's data utility resulting from a strategy. At the same time, the vertical axis in Figure 5.12B indicates the data subject's payoff resulting from a strategy. From Figure 5.12A, we can find that the best strategy must be in the region (shaded in blue) enclosed by two sets of parallel lines, which can be represented by the following equations:

$$U \geq 1 - \left(\frac{1}{k} - \frac{1}{k'}\right)\frac{L}{B} = 1 - \left(\frac{1}{2} - \frac{1}{157{,}681}\right)\frac{\$150}{\$100} = 0.25 \tag{5.15}$$

$$0 \leq U \leq 1 \tag{5.16}$$

$$\hat{r} \leq \begin{cases} max\left(\frac{kC}{L}, \theta\right) = max\left(\frac{2 \times \$10}{\$150}, 0.5\right) = 0.5, & r = 1 \\ max\left(\frac{k'C}{L}, \theta\right) = max\left(\frac{157{,}681 \times \$10}{\$150}, \theta\right) = 10{,}512.07, & r = 0 \end{cases} \tag{5.17}$$

$$0 \leq \hat{r} \leq 1 \tag{5.18}$$

Here, U is the utility of the shared data. K is the number of the query results of Craig Venter's demographic attributes with the surname. K' is the number of the query results of Craig Venter's demographic attributes without the surname. L is the loss from being re-identified. B is the benefit of sharing all data. C is the cost of an attack. $\hat{r}$ is the confidence score as the estimated correctness of the inferred surname. r is the correctness of the inferred surname. θ is the threshold for confidence score.

The strategy in this region with the highest utility is the optimal strategy. The strategy on the bottom line of this region has the same payoff as the no-protection strategy. If no strategies are in this region, then the no-protection strategy is optimal. A more straightforward way to find the best strategy is through Figure 5.12B, in which the best strategy is the strategy with the highest payoff.

From Figure 5.12B, we can find that the success of the surname inference plays a considerable role in choosing the optimal strategy. However, this is not always the case. For a common surname, the surname inference tends to be always successful. However, the difference between a successful and unsuccessful attack regarding the data subject's loss is relatively small (i.e., close to zero) for a common surname. For a rare surname, the difference between a successful and unsuccessful attack in terms of the loss to the data subject is relatively large (i.e., close to one). However, the surname inference will hardly be successful in this situation because it is highly likely that the surname does not exist in the referenced dataset. For the surname, Venter, which is a relatively rare surname (shared by 653 people in the US according to the 2010 US Census), the surname inference will be successful for some targets, and the difference between a successful attack and an unsuccessful attack in terms of the loss to the data subject substantially affects the payoff, which makes this two-stage re-identification attack quite successful. In the surname inference stage, only the success rate matters, whereas in the record linkage stage, only the number of matched identified records matters. Thus, in the two-stage attack, which integrates the surname inference and the record linkage, both success rate and group size matter, and the rareness of a surname will simultaneously affect these two variables.

5.4 Discussion

The methodology described in this study enables data subjects to make informed data sharing decisions in the face of complex state-of-the-art re-identification models. It enables people to answer questions such as, "Should I opt in to the health data sharing program?" and "Which portion of my genomic data should I share with the health program?" Moreover, the methodology is sufficiently flexible to enable data subjects to make decisions in platforms where sharing partial or modified data is allowed.

Our illustration of this methodology in the context of a known multi-stage attack on genomic data led to several notable findings. First, although an additional stage can substantially increase the accuracy of the re-identification when there is no protection, it makes the attack more vulnerable to our game theoretic protection because the adversary could be tricked into inferring wrong intermediate information, and thus mitigates the privacy risk. Second, most people (acting rationally) would choose to opt out of a data repository if partial data sharing is not permitted. By contrast, most people would share most of their data if sharing partial data is permitted. This finding is intriguing because it suggests that providing data subjects with options could encourage a greater degree of data sharing while preventing re-identification. Third, our extensive sensitivity analysis illustrates how parameters of our model influence behavior differently. For example, to effectively promote data sharing in general, policymakers could increase penalties for detected privacy breaches and rewards for data sharing. In addition, the analysis shows what direction a data subject should care more about for each parameter. Specifically, considering the sensitivity in the masking game, a data subject should take extra care when the benefit of sharing all data and the cost of an attack are low. For the same reason, over-estimating these two parameters may be worse than over-estimating others. More importantly, the analysis demonstrates the robustness of the methodology that, while an adversary can push most parameters such as damages of attacks and sizes of datasets to a risky point for data subjects in poorly protected scenarios, the highly effective and stable protection provided by the masking game are almost immune to these exacerbations.

Limitations exist in our current model, which provide directions for our future works. First, we only considered one adversary and did experiments with two-stage attacks. In the future, we will consider game models with multiple adversaries and conduct experiments with attacks that have more than two stages. For instance, in each stage, the adversary could infer a set of attributes using an external dataset or launch a linkage attack, as illustrated by attacks with far more than two stages [288, 294]. It is a challenging task because the chain of attack could expand over time. For example, a previously safe and trusted database may start to be targeted and attacked by adversaries if it loosens its access policies or if its vulnerabilities are discovered. Second, we used a simplified decision-making model in which either each data subject makes the sharing decision independently or all data subjects pick the same protection strategy. In the future, we will consider interactions among related data subjects (e.g., family members), and interactions between a data subject and databases. Third, the adversary modeled in our current game model is myopic without fully reasoning about all the uncertainties. To be more realistic, more complex game models, such as the Bayesian game, can be employed in case of incomplete information or imperfect information. Fourth, our game theoretic model could be applied to other multi-stage privacy attacks, such as membership inference and genome reconstruction attacks [79, 101, 102]. Because several works have shown correlations between SNP and STR markers [98, 107] and the Gymrek attack infers surnames from datasets with only Y-STR markers, it is worth examining the effectiveness of our protection model against multi-stage attacks involving datasets with SNP markers. Last, to deal with the high-computation cost of solving complex game models, especially in the face of multi-stage attacks, a type of Game-as-a-Service architecture could be deployed in the cloud to provide distributed game modeling and computing as a service to data subjects or their agents.

Chapter 6

CONCLUSION

6.1. Discussion

Table 6.1 Comparison of four games for privacy-preserving data sharing.

Task	1	2	3	4
Game	One-stage Re-identification	Beacon Services	Membership Inference	Multi-stage Re-identification
Game Model	Two-player Stackelberg games			
Released Data Type	Demographic data	Genomic statistics	Genomic statistics	Genomic and demographic data
Released Data Granularity	Individual SNPs with demographics	Allele presences	Allele frequencies	Individual Y-STRs with surname and demographics
Dataset	Adult (US Census)	1000 Genomes Phase 3	eMERGE SPHINX, 1000 Genomes Phase 3	Simulated dataset, Ysearch, Craig Venter's sequence
Dataset Size	Age, race, gender, and ZIP codes for 32,561 records	400,000 SNVs for 500 individuals	51,826 SNPs for 10,698 individuals	Surname and 50 Y-STRs for 58,218 records
Privacy Threat	Identity disclosure	Attribute (membership) disclosure	Attribute (membership) disclosure	Identity disclosure and attribute disclosure
Attack Method	Record linkage	Likelihood ratio test	Likelihood ratio test	Record linkage and surname inference
Number of Stages	One-stage	One-stage	One-stage	Multi-stage
Attack Introducer(s)	Sweeney	Shringapure and Bustamante	Sankararaman et al.	Gymrek et al.
Adversary's Background Knowledge	Voter registration lists	Targets' SNVs with identity and allele frequencies in the population	Targets' SNPs with identity and allele frequencies in the population	Public record registry, genetic genealogy dataset
Protection Operation(s)	Generalization	Flipping and masking	Masking	Masking
Game Experiment	Yes	No	Yes	Yes
Algorithm	Lattice-based algorithm	N/A	Genetic Algorithm	Lattice-based greedy algorithm with pruning
Open Source Tool	No	No	Yes	Yes

Abbreviations: SNP: Single Nucleotide Polymorphism. SNV: Single Nucleotide Variant. Y-STR stands for Y-chromosome short tandem repeat. eMERGE: Electronic Medical Records and Genomics. SPHINX: Sequence and Phenotype Integration Exchange. CEPH: Centre d'Etude du Polymorphism Humain.

Although all game models I built for privacy preserving data sharing are two-player Stackelberg games, they are different in many aspects. First, the membership inference game and the Beacon services game target genomic data (SNPs or SNVs), the one-stage re-identification game targets demographic data, and the multi-stage re-identification game targets both genomic data (Y-STRs) and demographic data. Second, the membership inference game and the Beacon services game release summary-level statistics, while both re-identification games release individual-level data. Third, the datasets used in experiments for the membership inference game have more attributes (or higher dimensionality) and fewer records (or smaller scale) than those for both re-identification games. Fourth, the membership inference game and the Beacon services game counteract membership inference attacks, while the one-stage and multi-stage re-identification games counteract re-identification attacks. Fifth, the multi-stage re-identification game considers a multi-stage attack while others only consider a one-stage attack. Sixth, the membership inference game and the multi-stage re-identification game use masking operation, the one-stage re-identification game uses generalization operation, and the Beacon services game uses both flipping and masking operations. They are also different in other aspect. More details regarding the differences among these four games are summarized in Table 6.1.

6.2. Contributions

The methodology developed throughout my dissertation research enables data sharers to make informed data sharing decisions in the face of complex state-of-the-art privacy attacks. These results are remarkable because they demonstrate that blending economic, legal, and technical approaches together though a game theoretic lens can dramatically improve our ability to strike the right balance between data utility and privacy risk in privacy-preserving data sharing.

I demonstrated that game theoretic approaches have the potential to revolutionize how policies are designed to protect data privacy. It should be recognized that it is fundamentally a problem of risk management that data will be compromised only if adversaries are sufficiently incentivized to do so. Strategic adversarial models lead to more rigorous and practical risk assessments and risk mitigations.

To enable pragmatic discussion about privacy concerns, I introduced a novel formalization of the several high-profile attack problems as Stackelberg games between a data sharer and an adversary. I translated the risk and the utility of the released dataset into the data sharer's benefit and cost and proposed several methods for computing the optimal data sharing policy for the sharer in a massive strategy space. I illustrated my models using real case studies and conducted experiments using real health and genomic data.

The most prominent finding is that sharers can choose strategies that allow for sharing a significant amount of data, which is much higher than the baselines, with a high payoff almost as much as the optimal solution, while ensuring that it is never beneficial for the data recipient to attempt re-identification, thus ensuring no attack and zero risk within the context of our modeling framework.

I further demonstrated that fine-grained sharing policies based on game theory can achieve high payoffs for a data sharer but bring computational challenges, especially for privacy-preserving sharing of high-dimensional data. To overcome the computational challenges, I implemented Artificial Intelligence search algorithms and heuristics to accelerate the search for the optimal solution. The first algorithm is a genetic algorithm which is perfectly compatible with genomic data sharing problems because all control variables in corresponding optimalization problems are binary variables. The second algorithm is a lattice-based greedy algorithm with pruning techniques like the alpha-beta pruning which is brand-and-bound algorithm suitable for the lattice structure of the strategy space. In addition, I used hath tables to avoid repeated computations during search processes.

For a multi-stage attack, although an additional stage can substantially increase the accuracy of the re-identification when there is no protection, it makes the attack more vulnerable to our game theoretic protection because the adversary could be tricked into inferring wrong intermediate information and thus mitigates the privacy risk.

Comparing to anonymization models such as K-anonymity and differential privacy, game-based policy-making models are more explainable and transparent. First, for any policy recommended by game models that looks counter-intuitive, its rationale could be clearly explained by examining the risk assessment process. Second, parameters in game models are based on real scenarios with specific meanings instead of a proposed metric that is abstract.

Finally, these conclusions are robust to order-of-magnitude changes in parameters of my game models.

6.3. Limitations and Future Works

There are several limitations or challenges that confronted by all the game models, which provide directions for my future works.

First, all game models assume a single adversary (data recipient). It would be desirable to generalize the model to capture the scenario of multiple data recipients, which would lead to a multi-follower

Stackelberg game. Further considerations of the uncertainties about the payoffs and information of data recipients would lead to a multi-follower Bayesian Stackelberg game. When multiple adversaries are involved, cooperation and competition among them and the public good should also be considered.

Second, the risk assessments are based on the capabilities of an attacker at a specific moment, which might improve in the future. Because the protection model expects a one-time release of data, no matter how accurate the current models on the parameter valuation or adversarial models are, they would likely change over time. Unfortunately, it is nearly impossible to protect the released dataset from future attacks. However, an alternative could be to provide access to the dataset within an accountable environment, such as a registered query-based access system. Since the scenario would no longer be a one-time release of data, a more complex game model would be required, such as the sequential game model.

Third, like most game models, my game models are based on the assumptions that players in the game are rational. In situations where players may make irrational decisions, behavior game theory could be utilized to explain the decision-making processes using real-life experiments.

Fourth, the settings of parameters are subject to change in different scenarios and are difficult to justify. For example, the value of data is likely dependent upon anticipated usages of the data where the recipient could be an insurer or a blackmailer. Sensitivity analyses on these essential parameters are conducted to justify the settings of parameters in the case study. It turns out that the game policy is always the preferred policy, no matter how the values for the parameters vary. Furthermore, the sharer's expected payoff in a game theoretic solution is insensitive to fluctuations of parameters. This suggests that a rough estimate of parameters is sufficient to apply the policy in practice.

Fifth, this game theoretic methodology examined in high-dimensional environments has the potential to be used in other applications with different settings, data types, attacks, or protection methods. First, our game theoretic models could be used in other settings such as public health settings. In public health settings, health and demographic data are incentivized to be shared with third parties or shared to the public to monitor and inform the spread of an epidemic or a pandemic. However, these data are inherently sensitive and need privacy protection. Second, our game theoretic models could be used for privacy-preserving sharing of images, videos, and texts to achieve a better payoff for the sharing by considering fine-grained sharing policies. Third, our game theoretic models could be applied to other multi-stage privacy attacks such as genome reconstruction attacks. Last but not the least, our game theoretic models are compatible with obfuscation (i.e., noise addition) operations and have the potential to supplement or even replace differential privacy models. For example, the parameters epsilon and delta in the differential privacy could be set optimally and transparently using game theoretic approaches with a specific adversarial model in consideration. Local differential privacy models could be used for privacy-preserving sharing of individual-level datasets. Furthermore, by considering a much larger strategy space in terms of the level of noise (discretized) could be added to each attribute in a data record, better tradeoff between data utility and privacy risk could be found comparing to a differential privacy model that is based on fixed parameters.

REFERENCES

[1] All of Us Research Program Investigators. (2019). The “All of Us” research program. *New England Journal of Medicine*, 381(7), 668-676.

[2] Hazel, J. W., & Slobogin, C. (2018). Who Knows What, and When: A Survey of the Privacy Policies Proffered by US Direct-to-Consumer Genetic Testing Companies. *Cornell JL & Pub. Pol'y*, 28, 35.

[3] Eisenstein, M. (2015). Big data: The power of petabytes. *Nature*, 527(7576), S2-S4.

[4] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, 25(1), 37-43.

[5] US Department of Health and Human Services. (2002). Standards for privacy of individually identifiable health information. Final rule. *Federal Register*, 67, 53181-53273.

[6] Wheeland, D. G. (2014). Final NIH genomic data sharing policy. *Federal Register*, 79, 51345-51354.

[7] Erlich, Y., & Narayanan, A. (2014). Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6), 409-421.

[8] Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J. P., ... and Wang, X. (2015). Privacy in the genomic era. *ACM Computing Surveys*, 48, 6.

[9] Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6), 417-428.

[10] Roberts, J. L., Pereira, S., & McGuire, A. L. (2017). Should you profit from your genome?. *Nature Biotechnology*, 35(1), 18-20.

[11] Watson, J. (2013). *Strategy: an introduction to game theory (Third Edition)*. WW Norton.

[12] Shoham, Y. and Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

[13] Von Neumann, J., & Morgenstern, O. (2007). *Theory of Games and Economic Behavior: 60th Anniversary Commemorative Edition*. Princeton University Press.

[14] Shoham, Y. (2008). Computer science and game theory. *Communications of the ACM*, 51, 74-79.

[15] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (Third Edition)*. Prentice Hall.

[16] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.

[17] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676), 354-359.

[18] Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., ... & Bowling, M. (2017). Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337), 508-513.

[19] Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... & Oh, J. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350-354.

[20] Manshaei, M. H., Zhu, Q., Alpcan, T., Bacşar, T. and Hubaux, J. P. (2013). Game theory meets network security and privacy. *ACM Computing Surveys*, 45, 1-39.

[21] Warren, S.D. and Brandeis, L.D. (1890). The right to privacy. *Harvard Law Review*, 4, 193-220.

[22] Acquisti, A., Brandimarte, L. and Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347, 509-514.

[23] US Department of Health and Human Services (2000). Standards for privacy and individually identifiable health information; final rule. *Federal Register*, 65, 82462–82829.

[24] Lin, Z., Owen, A. B., & Altman, R. B. (2004). Genomic Research and Human Subject Privacy. *Science*, 305(5681), 183-183.

[25] Pakstis, A. J., Speed, W. C., Fang, R., Hyland, F. C., Furtado, M. R., Kidd, J. R., & Kidd, K. K. (2010). SNPs for a universal individual identification panel. *Human Genetics*, 127(3), 315-324.

[26] Clayton, E. W., Halverson, C. M., Sathe, N. A., & Malin, B. A. (2018). A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. *PLoS One*, 13(10), e0204417.

[27] Sweeney, L. (1997) Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics*, 25, 98–110.

[28] Golle, P. (2006, October). Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society* (pp. 77-80).

[29] Koot, M., Van't Noordende, G. and De Laat, C. (2010). A study on the re-identifiability of Dutch citizens. In *Proceedings of the 3rd Workshop on Hot Topics in Privacy Enhancing Technologies*. (pp. 35-49)

[30] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., ... & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 4, e1000167.

[31] Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. and Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339, 321-324.

[32] Fung, B. C., Wang, K., Chen, R. and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42, 1-53.

[33] Sweeney L. (2000). Uniqueness of simple demographics in the US population. Technical Report. Carnegie Mellon University.

[34] Harmanci, A., & Gerstein, M. (2016). Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature Methods*, 13(3), 251-256.

[35] Lippert, C., Sabatini, R., Maher, M. C., Kang, E. Y., Lee, S., Arikan, O., ... & Yocum, K. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences*, 114(38), 10166-10171.

[36] Sankararaman, S., Obozinski, G., Jordan, M. I. and Halperin, E. (2009). Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41, 965-967.

[37] Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., ... & Sulem, P. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, 40(9), 1068-1075.

[38] Humbert, M., Ayday, E., Hubaux, J. P. and Telenti, A. (2013, November). Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1141-1152).

[39] Kayser, M., & De Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3), 179-192.

[40] Kantarcioglu, M., Jiang, W., Liu, Y., & Malin, B. (2008). A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on Information Technology in Biomedicine*, 12(5), 606-617.

[41] Xie, W., Kantarcioglu, M., Bush, W. S., Crawford, D., Denny, J. C., Heatherly, R., & Malin, B. A. (2014). SecureMA: protecting participant privacy in genetic association meta-analysis. *Bioinformatics*, 30(23), 3334-3341.

[42] Canim, M., Kantarcioglu, M., & Malin, B. (2011). Secure management of biomedical data with cryptographic hardware. *IEEE Transactions on Information Technology in Biomedicine*, 16(1), 166-175.

[43] Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S. O., ... & Haeussler, M. (2019). Federated discovery and sharing of genomic data using Beacons. *Nature Biotechnology*, 37(3), 220-224.

[44] Gammon, K. (2018). Experimenting with blockchain: Can one technology boost both data integrity and patients' pocketbooks?. *Nature Medicine*, 24(4), 378-381.

[45] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 557-570.

[46] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1, 3-es.

[47] Li, N., Li, T. and Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering* (pp. 106-115).

[48] Dwork C. (2006, July) Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming* (pp. 1–12).

[49] Xia, W., Heatherly, R., Ding, X., Li, J. and Malin, B. (2013, February). Efficient discovery of de-identification policy options through a risk-utility frontier. In *Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy* (pp. 59-70).

[50] Xia, W., Heatherly, R., Ding, X., Li, J., & Malin, B. A. (2015). RU policy frontiers for health data de-identification. *Journal of the American Medical Informatics Association*, 22(5), 1029-1041.

[51] Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... & Robinson, G. E. (2015). Big data: astronomical or genomical?. *PLoS Biology*, 13(7), e1002195.

[52] Phillips, A. M. (2016). Only a click away—DTC genetics for ancestry, health, love… and more: A view of the business and regulatory landscape. *Applied and Translational Genomics*, 8, 16-22.

[53] Taber, K. A. J., Dickinson, B. D., & Wilson, M. (2014). The promise and challenges of next-generation genome sequencing for clinical care. *JAMA Internal Medicine*, 174(2), 275-280.

[54] Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., ... & Brilliant, M. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genetics in Medicine*, 15(10), 761-771.

[55] Aronson, S. J., & Rehm, H. L. (2015). Building the foundation for genomics in precision medicine. *Nature*, 526(7573), 336-342.

[56] Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10), 681-691.

[57] Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1), 27-38.

[58] Hayden, E. C. (2013). Geneticists push for global data-sharing: international organization aims to promote exchange and linking of DNA sequences and clinical information. *Nature*, 498(7452), 16-17.

[59] ACMG Board of Directors. (2017). Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine*, 19(7), 721-722.

[60] Ball, M. P., Bobe, J. R., Chou, M. F., Clegg, T., Estep, P. W., Lunshof, J. E., ... & Church, G. M. (2014). Harvard Personal Genome Project: lessons from participatory public research. *Genome medicine*, 6, 10.

[61] Greshake, B., Bayer, P. E., Rausch, H., & Reda, J. (2014). OpenSNP–a crowdsourced web resource for personal genomics. *PLoS One*, 9(3), e89204.

[62] Ozercan, H. I., Ileri, A. M., Ayday, E., & Alkan, C. (2018). Realizing the potential of blockchain technologies in genomics. *Genome Research*, 28(9), 1255-1263.

[63] McGuire, A. L., Oliver, J. M., Slashinski, M. J., Graves, J. L., Wang, T., Kelly, P. A., ... & Treadwell-Deering, D. (2011). To share or not to share: a randomized trial of consent for data sharing in genome research. *Genetics in Medicine*, 13(11), 948-955.

[64] Oliver, J. M., Slashinski, M. J., Wang, T., Kelly, P. A., Hilsenbeck, S. G., & McGuire, A. L. (2012). Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Public Health Genomics*, 15(2), 106-114.

[65] The European Parliament and the Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union L*, 119, 1-88.

[66] Zerhouni, E. and Nabel, E. G. (2008). Protecting aggregate genomic data. *Science*, 322, 44.

[67] Paltoo, D. N., Rodriguez, L. L., Feolo, M., Gillanders, E., Ramos, E. M., Rutter, J., ... & Caulder, M. (2014). Data use under the NIH GWAS data sharing policy and future directions. *Nature Genetics*, 46(9), 934.

[68] Lunshof, J. E., Chadwick, R., Vorhaus, D. B., & Church, G. M. (2008). From genetic privacy to open consent. *Nature Reviews Genetics*, 9(5), 406-411.

[69] Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., ... & Popova, N. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10), 1181-1186.

[70] Dalenius, T. (1986). Finding a needle in a haystack or identifying anonymous census records. *Journal of Official Statistics*, 2(3), 329.

[71] Sweeney, L. (2000). Simple demographics often identify people uniquely. Technical Report LIDAP-WP3. Available from: https://dataprivacylab.org/projects/identifiability/paper1.pdf [cited 31 October 2020].

[72] Malin, B., & Sweeney, L. (2000, November). Determining the identifiability of DNA database entries. In *Proceedings of the AMIA 2000 Annual Symposium* (pp. 537-541). American Medical Informatics Association.

[73] Malin, B., & Sweeney, L. (2004). How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37(3), 179-192.

[74] Malin, B. (2006, November). Re-identification of familial database records. In *Proceedings of the 2016 AMIA Annual Symposium* (pp. 524-528). American Medical Informatics Association.

[75] Sweeney, L., Abu, A., & Winn, J. (2013). Identifying participants in the personal genome project by name (a re-identification experiment). arXiv preprint arXiv:1304.7605.

[76] Rodriguez, L. L., Brooks, L. D., Greenberg, J. H., & Green, E. D. (2013). The complexities of genomic identifiability. *Science*, 339(6117), 275-276.

[77] Humbert, M., Huguenin, K., Hugonot, J., Ayday, E., & Hubaux, J. P. (2015, June-July). De-anonymizing genomic databases using phenotypic traits. In *Proceedings on Privacy Enhancing Technologies*, 2015(2), 99-114.

[78] Sero, D., Zaidi, A., Li, J., White, J. D., Zarzar, T. B. G., Marazita, M. L., ... & Shriver, M. D. (2019). Facial recognition from DNA using face-to-DNA classifiers. *Nature Communications*, 10(1), 2557.

[79] Wang, R., Li, Y. F., Wang, X., Tang, H. and Zhou, X. (2009, November). Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and Communications Security* (pp. 534-544).

[80] Cai, R., Hao, Z., Winslett, M., Xiao, X., Yang, Y., Zhang, Z. and Zhou, S. (2015). Deterministic identification of specific individuals from GWAS results. *Bioinformatics*, 31, 1701-1707.

[81] Zhang, L., Pan, Q., Wang, Y., Wu, X., & Shi, X. (2017). Bayesian network construction and genotype-phenotype inference using gwas statistics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(2), 475-489.

[82] Visscher, P. M. and Hill, W. G. (2009). The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet*, 5, e1000628.

[83] Jacobs, K. B., Yeager, M., Wacholder, S., Craig, D., Kraft, P., Hunter, D. J., ... and Thomas, G. D. (2009). A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*, 41, 1253-1257.

[84] Braun, R., Rowe, W., Schaefer, C., Zhang, J. and Buetow, K. (2009). Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS Genet*, 5, e1000668.

[85] Im, H. K., Gamazon, E. R., Nicolae, D. L., & Cox, N. J. (2012). On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *The American Journal of Human Genetics*, 90(4), 591-598.

[86] Shringarpure, S. S., & Bustamante, C. D. (2015). Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5), 631-646.

[87] Von Thenen, N., Ayday, E., & Cicek, A. E. (2019). Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics*, 35(3), 365-371.

[88] Dwork, C., Smith, A., Steinke, T., Ullman, J., & Vadhan, S. (2015, October). Robust traceability from trace amounts. *In 2015 IEEE 56th Annual Symposium on Foundations of Computer Science* (pp. 650-669). IEEE.

[89] Wang, Y., Wen, J., Wu, X., & Shi, X. (2016, July). Infringement of individual privacy via mining differentially private GWAS statistics. In *International Conference on Big Data Computing and Communications* (pp. 355-366). Springer, Cham.

[90] Lumley, T., & Rice, K. (2010). Potential for revealing individual-level information in genome-wide association studies. *JAMA*, 303(7), 659-660.

[91] Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., & Ristenpart, T. (2014, August). Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX conference on Security Symposium* (pp. 17-32).

[92] Beaulieu-Jones, B. K., Yuan, W., Finlayson, S. G., & Wu, Z. S. (2018). Privacy-preserving distributed deep learning for clinical data. arXiv preprint arXiv:1812.01484.

[93] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy* (pp. 3-18). IEEE.

[94] Hayes, J., Melis, L., Danezis, G., & De Cristofaro, E. (2017). LOGAN: evaluating privacy leakage of generative models using generative adversarial networks. arXiv preprint arXiv:1705.07663.

[95] Halperin, E., & Stephan, D. A. (2009). SNP imputation in association studies. *Nature Biotechnology*, 27(4), 349-351.

[96] Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7), 499-511.

[97] Nyholt, D. R., Yu, C. E., & Visscher, P. M. (2009). On Jim Watson's APOE status: genetic information is hard to hide. *European Journal of Human Genetics*, 17(2), 147-149.

[98] Edge, M. D., Algee-Hewitt, B. F., Pemberton, T. J., Li, J. Z., & Rosenberg, N. A. (2017). Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proceedings of the National Academy of Sciences*, 114(22), 5671-5676.

[99] Ayday, E., & Humbert, M. (2017). Inference attacks against kin genomic privacy. *IEEE Security and Privacy*, 15(5), 29-37.

[100] Humbert, M., Ayday, E., Hubaux, J. P., & Telenti, A. (2017). Quantifying interdependent risks in genomic privacy. *ACM Transactions on Privacy and Security*, 20(1), 1-31.

[101] Deznabi, I., Mobayen, M., Jafari, N., Tastan, O., & Ayday, E. (2017). An inference attack on genomic data using kinship, complex correlations, and phenotype information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(4), 1333-1343.

[102] Ayoz, K., Ayday, E., & Cicek, A. E. (2020). Genome reconstruction attacks against genomic data-sharing beacons. arXiv preprint arXiv:2001.08852.

[103] Backes, M., Berrang, P., Humbert, M., Shen, X., & Wolf, V. (2018). Simulating the large-scale erosion of genomic privacy over time. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5), 1405-1412.

[104] Ney, P., Ceze, L., & Kohno, T. (2020). Genotype extraction and false relative attacks: security risks to third-party genetic genealogy services beyond identity inference. In *Network and Distributed System Security Symposium*.

[105] Kennett, D. (2019). Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. *Forensic Science International*, 301, 107-117.

[106] Erlich, Y., Shor, T., Pe'er, I., & Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. *Science*, 362(6415), 690-694.

[107] Kim, J., Edge, M. D., Algee-Hewitt, B. F., Li, J. Z., & Rosenberg, N. A. (2018). Statistical detection of relatives typed with disjoint forensic and biomedical loci. *Cell*, 175(3), 848-858.

[108] Ellenbogen, P., & Narayanan, A. (2019). Identification of anonymous DNA using genealogical triangulation. bioRxiv, 531269.

[109] Malin, B. A. (2005). An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *Journal of the American Medical Informatics Association*, 12(1), 28-34.

[110] Tang, H., Jiang, X., Wang, X., Wang, S., Sofia, H., Fox, D., ... & Ohno-Machado, L. (2016). Protecting genomic data analytics in the cloud: state of the art and opportunities. *BMC Medical Genomics*, 9(1), 63.

[111] Wang, S., Jiang, X., Singh, S., Marmor, R., Bonomi, L., Fox, D., ... & Ohno-Machado, L. (2017). Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. *Annals of the New York Academy of Sciences*, 1387(1), 73.

[112] Shi, X., & Wu, X. (2017). An overview of human genetic privacy. *Annals of the New York Academy of Sciences*, 1387(1), 61.

[113] Aziz, M. M. A., Sadat, M. N., Alhadidi, D., Wang, S., Jiang, X., Brown, C. L., & Mohammed, N. (2019). Privacy-preserving techniques of genomic data—a survey. *Briefings in bioinformatics*, 20(3), 887-895.

[114] Mittos, A., Malin, B., & De Cristofaro, E. (2019, July). Systematizing genome privacy research: A privacy-enhancing technologies perspective. In *Proceedings on Privacy Enhancing Technologies*, 2019(1), 87-107.

[115] Berger, B., & Cho, H. (2019). Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biology*, 20(1), 128.

[116] E. Ayday, E. de Cristofaro, J.-P. Hubaux, and G. Tsudik. The Chills and Thrills of Whole Genome Sequencing. *IEEE Computer*, 2013.

[117] Kim, M., & Lauter, K. (2015, December). Private genome analysis through homomorphic encryption. *BMC Medical Informatics and Decision Making*, 15(Suppl 5), S3.

[118] Lauter, K., López-Alt, A., & Naehrig, M. (2014, September). Private computation on encrypted genomic data. In *Proceedings of the 2014 International Conference on Cryptology and Information Security in Latin America* (pp. 3-27). Springer, Cham.

[119] Wang, S., Zhang, Y., Dai, W., Lauter, K., Kim, M., Tang, Y., ... & Jiang, X. (2016). HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS. *Bioinformatics*, 32(2), 211-218.

[120] Hasan, M. Z., Mahdi, M. S. R., Sadat, M. N., & Mohammed, N. (2018). Secure count query on encrypted genomic data. *Journal of Biomedical Informatics*, 81, 41-52.

[121] Cho, H., Wu, D. J., & Berger, B. (2018). Secure genome-wide association analysis using multiparty computation. *Nature Biotechnology*, 36(6), 547-551.

[122] Chen, F., Wang, S., Jiang, X., Ding, S., Lu, Y., Kim, J., ... & Png, E. (2017). Princess: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. *Bioinformatics*, 33(6), 871-878.

[123] Bonte, C., Makri, E., Ardeshirdavani, A., Simm, J., Moreau, Y., & Vercauteren, F. (2018). Towards practical privacy-preserving genome-wide association study. *BMC Bioinformatics*, 19(1), 1-12.

[124] Sadat, M. N., Al Aziz, M. M., Mohammed, N., Chen, F., Jiang, X., & Wang, S. (2018). Safety: secure gwas in federated environment through a hybrid solution. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1), 93-102.

[125] Loukides, G., Gkoulalas-Divanis, A. and Malin, B. (2010). Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107, 7898-7903.

[126] Tramèr, F., Huang, Z., Hubaux, J. P., & Ayday, E. (2015, October). Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp. 1286-1297).

[127] Simmons, S., Sahinalp, C., & Berger, B. (2016). Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Systems*, 3(1), 54-61.

[128] Li, B., Karwa, V., Slavković, A., & Steorts, R. C. (2018). A privacy preserving algorithm to release sparse high-dimensional histograms. *Journal of Privacy and Confidentiality*, 8(1).

[129] Bae, H., Jung, D., Choi, H. S., & Yoon, S. (2020, January). AnomiGAN: Generative Adversarial Networks for Anonymizing Private Medical Data. In *Pacific Symposium on Biocomputing* (Vol. 25, pp. 563-574).

[130] El Emam, K., Jabbouri, S., Sams, S., Drouet, Y. and Power, M. (2006). Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research*, 8, e28.

[131] Domingo-Ferrer, J. and Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11, 195-212.

[132] Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 571-588.

[133] Malin, B. A. (2005). Protecting genomic sequence anonymity with generalization lattices. *Methods of Information in Medicine*, 44(05), 687-692.

[134] Li, G., Wang, Y., & Su, X. (2012). Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. *Computer Methods and Programs in Biomedicine*, 108(1), 1-9.

[135] Heatherly, R. D., Loukides, G., Denny, J. C., Haines, J. L., Roden, D. M., & Malin, B. A. (2013). Enabling genomic-phenomic association discovery without sacrificing anonymity. *PloS One*, 8(2), e53875.

[136] Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report Technical Report SRICSL-98-04, SRI Computer Science Laboratory.

[137] Iyengar, V. S. (2002, July). Transforming data to satisfy privacy constraints. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 279-288).

[138] LeFevre, K., DeWitt, D. J. and Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering* (pp. 25-25).

[139] Nergiz, M. E., Atzori, M. and Clifton, C. (2007, June). Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data* (pp. 665-676).

[140] Dwork, C. (2011, October). The Promise of Differential Privacy A Tutorial on Algorithmic Techniques. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science* (pp. 1-2).

[141] Uhlerop, C., Slavković, A., & Fienberg, S. E. (2013). Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality*, 5(1), 137.

[142] Yu, F., Fienberg, S. E., Slavković, A. B., & Uhler, C. (2014). Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50, 133-141.

[143] Johnson, A., & Shmatikov, V. (2013, August). Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1079-1087).

[144] Simmons, S., & Berger, B. (2016). Realizing privacy preserving genome-wide association studies. *Bioinformatics*, 32(9), 1293-1300.

[145] Raisaro, J. L., Choi, G., Pradervand, S., Colsenet, R., Jacquemont, N., Rosat, N., ... & Hubaux, J. P. (2018). Protecting privacy and security of genomic data in I2B2 with homomorphic encryption and differential privacy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5), 1413-1426.

[146] Raisaro, J. L., Troncoso-Pastoriza, J. R., Misbach, M., Sousa, J. S., Pradervand, S., Missiaglia, E., ... & Hubaux, J. P. (2018). M ed C o: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(4), 1328-1341.

[147] Simmons, S., Berger, B., & Sahinalp, C. (2019, January). Protecting Genomic Data Privacy with Probabilistic Modeling. In Pacific Symposium on Biocomputing. In *Pacific Symposium on Biocomputing* (Vol. 24, p. 403).

[148] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014, December). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672-2680).

[149] Bae, H., Jung, D., & Yoon, S. (2019). AnomiGAN: Generative adversarial networks for anonymizing private medical data. arXiv preprint arXiv:1901.11313.

[150] Dyke, S. O., Linden, M., Lappalainen, I., De Argila, J. R., Carey, K., Lloyd, D., ... & Cutts, T. (2018). Registered access: authorizing data access. *European Journal of Human Genetics*, 26(12), 1721-1731.

[151] Kaye, J., Curren, L., Anderson, N., Edwards, K., Fullerton, S. M., Kanellopoulou, N., ... & Taylor, P. L. (2012). From patients to partners: participant-centric initiatives in biomedical research. *Nature Reviews Genetics*, 13(5), 371-376.

[152] Deuber, D., Egger, C., Fech, K., Malavolta, G., Schröder, D., Thyagarajan, S. A. K., ... & Durand, C. (2019, July). My genome belongs to me: controlling third party computation on genomic data. In *Proceedings on Privacy Enhancing Technologies*, 2019(1), 108-132.

[153] Grishin, D., Obbad, K., Estep, P., Quinn, K., Zaranek, S. W., Zaranek, A. W., ... & Church, G. (2018). Accelerating genomic data generation and facilitating genomic data access using decentralization, privacy-preserving technologies and equitable compensation. *Blockchain Healthc Today*, 1, 1-23.

[154] Shabani, M. (2019). Blockchain-based platforms for genomic data sharing: a de-centralized approach in response to the governance problems?. *Journal of the American Medical Informatics Association*, 26(1), 76-80.

[155] Zhang, Y., Zhao, X., Li, X., Zhong, M., Curtis, C., & Chen, C. (2019, January). Enabling privacy-preserving sharing of genomic data for GWASs in decentralized networks. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining* (pp. 204-212).

[156] Kuo, T. T., Gabriel, R. A., & Ohno-Machado, L. (2019). Fair compute loads enabled by blockchain: sharing models by alternating client and server roles. *Journal of the American Medical Informatics Association*, 26(5), 392-403.

[157] Zhou, X., Peng, B., Li, Y. F., Chen, Y., Tang, H. and Wang, X. (2011, September). To release or not to release: evaluating information leaks in aggregate human-genome data. In *European Symposium on Research in Computer Security* (pp. 607-627). Springer, Berlin, Heidelberg.

[158] Ayday, E., Raisaro, J. L., Hubaux, J. P. and Rougemont, J. (2013, November). Protecting and evaluating genomic privacy in medical tests and personalized medicine. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society* (pp. 95-106).

[159] Raisaro, J. L., Tramer, F., Ji, Z., Bu, D., Zhao, Y., Carey, K., ... & Shringarpure, S. (2017). Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *Journal of the American Medical Informatics Association*, 24(4), 799-805.

[160] Al Aziz, M. M., Ghasemi, R., Waliullah, M., & Mohammed, N. (2017). Aftermath of bustamante attack on genomic beacon service. *BMC Medical Genomics*, 10(2), 43.

[161] El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., ... & Roffey, T. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16, 670-682.

[162] Elliot, M. and Dale, A. (1999). Scenarios of attack: the data intruder’s perspective on statistical disclosure risk. *Netherlands Official Statistics*, 14, 6-10.

[163] Dankar, F. K. and El Emam, K. (2010, March). A method for evaluating marketer re-identification risk. In *Proceedings of the 2010 EDBT/ICDT Workshops* (pp. 1-10).

[164] Banerjee, M., Adl, R. K., Wu, L. and Barker, K. (2011, September). Quantifying privacy violations. In *Workshop on Secure Data Management* (pp. 1-17). Springer, Berlin, Heidelberg.

[165] Lebanon, G., Scannapieco, M., Fouad, M. R. and Bertino, E. (2006, December). Beyond k-anonymity: A decision theoretic framework for assessing privacy risk. In *International Conference on Privacy in Statistical Databases* (pp. 217-232). Springer, Berlin, Heidelberg.

[166] Benitez, K. and Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, 17, 169-177.

[167] Malin, B., Loukides, G., Benitez, K., & Clayton, E. W. (2011). Identifiability in biobanks: models, measures, and mitigation strategies. *Human Genetics*, 130(3), 383.

[168] Humbert, M., Ayday, E., Hubaux, J. P., & Telenti, A. (2014, November). Reconciling utility with privacy in genomics. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society* (pp. 11-20).

[169] Kale, G., Ayday, E., & Tastan, O. (2018). A utility maximizing and privacy preserving approach for protecting kinship in genomic databases. *Bioinformatics*, 34(2), 181-189.

[170] Craig, D. W., Goor, R. M., Wang, Z., Paschall, J., Ostell, J., Feolo, M., ... & Manolio, T. A. (2011). Assessing and managing risk when sharing aggregate genetic variant data. *Nature Reviews Genetics*, 12(10), 730-736.

[171] Wagner, I. (2017). Evaluating the strength of genomic privacy metrics. *ACM Transactions on Privacy and Security*, 20(1), 1-34.

[172] Barth-Jones, D., El Emam, K., Bambauer, J., Cavoukian, A., & Malin, B. (2015). Assessing data intrusion threats. *Science*, 348(6231), 194-195.

[173] El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PloS One*, 6(12), e28071.

[174] El Emam, K., & Arbuckle, L. (2014). *Anonymizing health data: Case studies and methods to get you started. 2nd ed*. O'Reilly Medi, Inc.

[175] Rasmusen, E. and Blackwell, B. (1994). *Games and information, volume 2*. Cambridge.

[176] Luedtke, A., Carone, M., Simon, N., & Sofrygin, O. (2020). Learning to learn from data: Using deep adversarial learning to construct optimal statistical procedures. *Science Advances*, 6(9), eaaw2140.

[177] Başar, T. and Olsder, G. J. (1999). *Dynamic noncooperative game theory, Second Edition*. Society for Industrial and Applied Mathematics.

[178] Tambe, M., Jiang, A. X., An, B., & Jain, M. (2014, March). Computational game theory for security: Progress and challenges. In *AAAI Spring Symposium on Applied Computational Game Theory*.

[179] Tambe, M., Jain, M., Pita, J. A. and Jiang, A. X. (2012, October). Game theory for security: Key algorithmic principles, deployed systems, lessons learned. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing* (pp. 1822-1829). IEEE.

[180] An, B., Brown, M., Vorobeychik, Y. and Tambe, M. (2013, May). Security games with surveillance cost and optimal timing of attack execution. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 223-230).

[181] Pita, J., Jain, M., Marecki, J., Ordóñez, F., Portway, C., Tambe, M., ... and Kraus, S. (2008, May). Deployed ARMOR protection: the application of a game theoretic model for security at the Los Angeles

International Airport. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (Industrial Track)* (pp. 125-132).

[182] Tsai, J., Rathi, S., Kiekintveld, C., Ordonez, F. and Tambe, M. (2009). IRIS-a tool for strategic security allocation in transportation networks. In *Proceedings of the 2009 International Conference on Autonomous Agents and Multi-Agent Systems (Industry Track)* (pp. 37-44).

[183] An, B., Pita, J., Shieh, E., Tambe, M., Kiekintveld, C. and Marecki, J. (2011). Guards and protect: Next generation applications of security games. *ACM SIGecom Exchanges*, 10, 31-34.

[184] Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., DiRenzo, J., ... & Meyer, G. (2012, July). PROTECT: an application of computational game theory for the security of the ports of the United States. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence* (pp. 2173-2179).

[185] Pita, J., Tambe, M., Kiekintveld, C., Cullen, S. and Steigerwald, E. (2011, May). GUARDS: game theoretic security allocation on a national scale. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 37-44).

[186] Yin, Z., Jiang, A. X., Tambe, M., Kiekintveld, C., Leyton-Brown, K., Sandholm, T. and Sullivan, J. P. (2012). TRUSTS: Scheduling randomized patrols for fare inspection in transit systems using game theory. *AI magazine*, 33, 59-59.

[187] Acquisti, A., Dingledine, R. and Syverson, P. (2003, January). On the economics of anonymity. In *International Conference on Financial Cryptography* (pp. 84-102). Springer, Berlin, Heidelberg.

[188] Chen, Y., Sheffet, O., & Vadhan, S. (2014, December). Privacy games. In *International Conference on Web and Internet Economics* (pp. 371-385). Springer, Cham.

[189] Chen, Y., Sheffet, O., & Vadhan, S. (2020). Privacy games. *ACM Transactions on Economics and Computation*, 8, 1-37.

[190] Rajbhandari, L. and Snekkenes, E. A. (2010, August). Using game theory to analyze risk to privacy: An initial insight. In *IFIP PrimeLife International Summer School on Privacy and Identity Management for Life* (pp. 41-51). Springer, Berlin, Heidelberg.

[191] Brückner, M. and Scheffer, T. (2009, December). Nash equilibria of static prediction games. In *Advances in Neural Information Processing Systems* (pp. 171-179).

[192] Brückner, M. and Scheffer, T. (2011, August). Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 547-555).

[193] Loiseau, P. (2016). Combining game theory and statistical learning for security, privacy and network systems (Doctoral dissertation, University of Innsbruck).

[194] Ioannidis, S. and Loiseau, P. (2013, December). Linear regression as a non-cooperative game. In *International Conference on Web and Internet Economics* (pp. 277-290). Springer, Berlin, Heidelberg.

[195] Chessa, M., Grossklags, J. and Loiseau, P. (2015, July). A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. In *2015 IEEE 28th Computer Security Foundations Symposium* (pp. 90-104). IEEE.

[196] Liu, W. and Chawla, S. (2009, December). A game theoretical model for adversarial learning. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 25-30). IEEE.

[197] Oh, S. J., Fritz, M. and Schiele, B. (2017, October). Adversarial image perturbation for privacy protection - A game theory perspective. In *2017 IEEE International Conference on Computer Vision* (pp. 1491-1500). IEEE.

[198] Gianini, G. and Damiani, E. (2008, August). A game-theoretical approach to data-privacy protection from context-based inference attacks: A location-privacy protection case study. In *Workshop on Secure Data Management* (pp. 133-150). Springer, Berlin, Heidelberg.

[199] Freudiger, J., Manshaei, M. H., Hubaux, J. P. and Parkes, D. C. (2009, November). On non-cooperative location privacy: a game-theoretic analysis. In *Proceedings of the 16th ACM conference on Computer and Communications Security* (pp. 324-337).

[200] Liu, X., Liu, K., Guo, L., Li, X. and Fang, Y. (2013, April). A game-theoretic approach for achieving k-anonymity in location based services. In *Proceedings of the 2013 IEEE International Conference on Computer Communications* (pp. 2985-2993). IEEE.

[201] Wang, W. and Zhang, Q. (2014, April). A stochastic game for privacy preserving context sensing on mobile phone. In *Proceedings of the 2014 IEEE Conference on Computer Communications* (pp. 2328-2336). IEEE.

[202] Olteanu, A. M., Huguenin, K., Humbert, M. and Hubaux, J. P. (2016). The sharing game: Benefits and privacy implications of (co)-location sharing with interdependences. EPFL, research report, 2016. [Online]. Available from: https://infoscience. epfl. ch/record/218755 [cited 26 October 2020].

[203] Shokri, R., Theodorakopoulos, G., Troncoso, C., Hubaux, J. P. and Le Boudec, J. Y. (2012, October). Protecting location privacy: optimal strategy against localization attacks. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security* (pp. 617-627).

[204] Shokri, R. (2015, June-July). Privacy games: Optimal user-centric data obfuscation. In *Proceedings on Privacy Enhancing Technologies*, 2015, 299-315.

[205] Shokri, R., Theodorakopoulos, G. and Troncoso, C. (2016). Privacy games along location traces: A game-theoretic framework for optimizing location privacy. *ACM Transactions on Privacy and Security*, 19, 1-31.

[206] Biczók, G. and Chia, P. H. (2013, April). Interdependent privacy: Let me share your data. In *International Conference on Financial Cryptography and Data Security* (pp. 338-353). Springer, Berlin, Heidelberg.

[207] Pu, Y. and Grossklags, J. (2014, November). An economic model and simulation results of app adoption decisions on networks with interdependent privacy consequences. In *International Conference on Decision and Game Theory for Security* (pp. 246-265). Springer, Cham.

[208] Chen, J., Kiremire, A. R., Brust, M. R. and Phoha, V. V. (2014). Modeling online social network users' profile attribute disclosure behavior from a game theoretic perspective. *Computer Communications*, 49, 18-32.

[209] Blocki, J., Christin, N., Datta, A., Procaccia, A. D. and Sinha, A. (2013, August). Audit games. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence* (pp. 41-47).

[210] Blocki, J., Christin, N., Datta, A., Procaccia, A. D. and Sinha, A. (2015, January). Audit games with multiple defender resources. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (pp. 791-797).

[211] Cárdenas, A. A., Amin, S., Schwartz, G., Dong, R. and Sastry, S. (2012, October). A game theory model for electricity theft detection and privacy-aware control in AMI systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing* (pp. 1830-1837). IEEE.

[212] Yan, C., Li, B., Vorobeychik, Y., Laszka, A., Fabbri, D., & Malin, B. (2018, April). Get your workload in order: Game theoretic prioritization of database auditing. In *2018 IEEE 34th International Conference on Data Engineering* (pp. 1304-1307). IEEE.

[213] Yan, C., Li, B., Vorobeychik, Y., Laszka, A., Fabbri, D., & Malin, B. (2019). Database audit workload prioritization via game theory. *ACM Transactions on Privacy and Security*, 22(3), 1-21.

[214] Kargupta, H., Das, K. and Liu, K. (2007, April). A game theoretic approach toward multi-party privacy-preserving distributed data mining. In *Proceedings of the 2007 European Conference on Principles of Data Mining and Knowledge Discovery*.

[215] Kantarcioglu, M., Bensoussan, A. and Hoe, S. C. (2010, November). When do firms invest in privacy-preserving technologies?. In *International Conference on Decision and Game Theory for Security* (pp. 72-86). Springer, Berlin, Heidelberg.

[216] Nix, R. and Kantarciouglu, M. (2011). Incentive compatible privacy-preserving distributed classification. *IEEE Transactions on Dependable and Secure Computing*, 9, 451-462.

[217] Wang, W., Ying, L. and Zhang, J. (2015, September). A game-theoretic approach to quality control for collecting privacy-preserving data. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing* (pp. 474-479). IEEE.

[218] Miyaji, A. and Rahman, M. S. (2011, July). Privacy-preserving data mining: a game-theoretic approach. In *IFIP Annual Conference on Data and Applications Security and Privacy* (pp. 186-200). Springer, Berlin, Heidelberg.

[219] Mohammed, N., Fung, B. C. and Debbabi, M. (2011). Anonymity meets game theory: secure data integration with malicious participants. *The VLDB Journal*, 20, 567-588.

[220] Kumari, V. and Chakravarthy, S. (2016). Cooperative privacy game: a novel strategy for preserving privacy in data publishing. *Human-centric Computing and Information Sciences*, 6, 1-20.

[221] Li, M., Carrell, D., Aberdeen, J., Hirschman, L., Kirby, J., Li, B., ... and Malin, B. A. (2016). Optimizing annotation resources for natural language de-identification via a game theoretic framework. *Journal of Biomedical Informatics*, 61, 97-109.

[222] Wu, X., Wu, T., Khan, M., Ni, Q., & Dou, W. (2017). Game theory based correlated privacy preserving analysis in big data. *IEEE Transactions on Big Data*, 1, 1-1.

[223] Duong, Q., LeFevre, K. and Wellman, M. P. (2010). Strategic modeling of information sharing among data privacy attackers. *Informatica*, 34, 151-158.

[224] Humbert, M., Ayday, E., Hubaux, J. P. and Telenti, A. (2015, January). On non-cooperative genomic privacy. In *International Conference on Financial Cryptography and Data Security* (pp. 407-426). Springer, Berlin, Heidelberg.

[225] Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. V. (2007). *Algorithmic Game Theory*. Cambridge University Press.

[226] Conitzer, V. and Sandholm, T. (2006, June). Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce* (pp. 82-90).

[227] Wan, Z., Vorobeychik, Y., Clayton, E. W., Kantarcioglu, M. and Malin, B. (2020). Game theory for privacy-preserving sharing of genomic data. In *Responsible Genomic Data Sharing* (pp. 135-160). Academic Press.

[228] Prasser, F., Gaupp, J., Wan, Z., Xia, W., Vorobeychik, Y., Kantarcioglu, M., ... & Malin, B. (2017). An open source tool for game theoretic health data de-identification. In *AMIA Annual Symposium Proceedings* (Vol. 2017, p. 1430). American Medical Informatics Association.

[229] Rasmussen-Torvik, L. J., Stallings, S. C., Gordon, A. S., Almoguera, B., Basford, M. A., Bielinski, S. J., ... & Crosslin, D. R. (2014). Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. *Clinical Pharmacology & Therapeutics*, 96(4), 482-489.

[230] Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., ... & Wang, J., & The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68-74.

[231] Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9), 793-795.

[232] Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., ... & Guarino, P. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, 70, 214-223.

[233] Bowton, E., Field, J. R., Wang, S., Schildcrout, J. S., Van Driest, S. L., Delaney, J. T., ... & Karnes, J. H. (2014). Biobanks and electronic medical records: enabling cost-effective research. *Science Translational Medicine*, 6(234), 234cm3-234cm3.

[234] Liu, Y., Wan, Z., Xia, W., Kantarcioglu, M., Vorobeychik, Y., Clayton, E. W., ... & Malin, B. A. (2018). Detecting the Presence of an Individual in Phenotypic Summary Data. In *AMIA Annual Symposium Proceedings* (Vol. 2018, p. 760). American Medical Informatics Association.

[235] Wang, S., Jiang, X., Tang, H., Wang, X., Bu, D., Carey, K., ... & Malin, B. (2017). A community effort to protect genomic data sharing, collaboration and outsourcing. *NPJ Genomic Medicine*, 2(1), 1-6.

[236] Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E. W., Kantarcioglu, M., Ganta, R., Heatherly, R. and Malin, B. A. (2015). A game theoretic framework for analyzing re-identification risk. *PloS One*, 10(3), e0120592.

[237] Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E. W., Kantarcioglu, M. and Malin, B. (2017). Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach. *The American Journal of Human Genetics*, 100(2), 316-322.

[238] Wan, Z., Vorobeychik, Y., Kantarcioglu, M. and Malin, B. (2017). Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services. *BMC Medical Genomics*, 10(2), 87-100.

[239] National Institutes of Health. (2003). Final NIH Statement on Sharing Research Data. Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [cited 26 October 2020].

[240] National Science Foundation. (2018). Frequently Asked Questions (FAQs) for Public Access. Available from: https://www.nsf.gov/pubs/2018/nsf18041/nsf18041.jsp [cited 26 October 2020].

[241] Polonetsky, J. and Tene, O. (2013). Privacy and Big Data: Making Ends Meet. *Stanford Law Review*, 66, 25.

[242] Fiske, S. T. and Hauser, R. M. (2014). Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences*, 111, 13675-13676.

[243] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M. and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 1376.

[244] Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D. and Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107, 22436-22441.

[245] Kosinski, M., Stillwell, D. and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110, 5802-5805.

[246] Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information". *Communications of the ACM*, 53(6), 24-26.

[247] Rothstein, M. A. (2010). Is deidentification sufficient to protect health privacy in research?. *The American Journal of Bioethics*, 10(9), 3-11.

[248] Gellman, R. (2010). The Deidentification Dilemma: A Legislative and Contractual Proposal. *Fordham Intellectual Property, Media and Entertainment Law Journal*, 21(1), 33.

[249] Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57, 1701.

[250] US Department of Health and Human Services. (2011). Human subjects research protections: enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators. *Federal Register*, 76(143), 44512-44531.

[251] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Available from: http://archive.ics.uci.edu/ml [cited 26 October 2020].

[252] US Census Bureau. (2011). 2010 Census Summary File 1 Tennessee[machine-readable data files] / prepared by the US Census Bureau. Available from: http://www.census.gov/prod/cen2010/doc/sf1.pdf [cited 26 October 2020].

[253] Benitez, K., Loukides, G., & Malin, B. (2010, November). Beyond safe harbor: automatic discovery of health information de-identification policy alternatives. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 163-172).

[254] National Institutes of Health. (2007). Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies. NOT-OD-07–088. Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html [cited 26 October 2020].

[255] National Center for Biotechnology Information. (2014). The database of Genotypes and Phenotypes (dbGaP). Available from: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000360.v1.p1 [cited 26 October 2020].

[256] National Institutes of Health. (2014). NIH Research Portfolio Online Reporting Tools (RePORT). Available from: http://projectreporter.nih.gov/reporter.cfm [cited 20 October 2014].

[257] US Department of Health and Human Services. (2013). HIPAA Breach Notification Rule. Available from: http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule [cited 26 October 2020].

[258] Wang, J., Xia, C., Wang, Y., Ding, S., & Sun, J. (2012). Spatial prisoner's dilemma games with increasing size of the interaction neighborhood on regular lattices. *Chinese Science Bulletin*, 57(7), 724-728.

[259] Zhu, C. J., Sun, S. W., Wang, L., Ding, S., Wang, J., & Xia, C. Y. (2014). Promotion of cooperation due to diversity of players in the spatial public goods game with increasing neighborhood size. *Physica A: Statistical Mechanics and its Applications*, 406, 145-154.

[260] Schatz, M. C. (2015). Biological data sciences in genome research. *Genome Research*, 25(10), 1417-1422.

[261] Kidd, J. M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., ... & McLaughlin, S. F. (2012). Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *The American Journal of Human Genetics*, 91(4), 660-671.

[262] Shringarpure, S. S., Bustamante, C. D., Lange, K., & Alexander, D. H. (2016). Efficient analysis of large datasets and sex bias with ADMIXTURE. *BMC Bioinformatics*, 17(1), 1-6.

[263] Trinidad, S. B., Fullerton, S. M., Bares, J. M., Jarvik, G. P., Larson, E. B., & Burke, W. (2010). Genomic research and wide data sharing: views of prospective participants. *Genetics in Medicine*, 12(8), 486-495.

[264] Paltoo, D. N., Rodriguez, L. L., Feolo, M., Gillanders, E., Ramos, E. M., Rutter, J., ... & Caulder, M. (2014). Data use under the NIH GWAS data sharing policy and future directions. *Nature Genetics*, 46(9), 934.

[265] Jiang, X., Zhao, Y., Wang, X., Malin, B., Wang, S., Ohno-Machado, L., & Tang, H. (2014). A community assessment of privacy preserving techniques for human genomes. *BMC Medical Informatics and Decision Making*, 14(S1), S1.

[266] Weil, C. J., Mechanic, L. E., Green, T., Kinsinger, C., Lockhart, N. C., Nelson, S. A., ... & Buccini, L. D. (2013). NCI think tank concerning the identifiability of biospecimens and "omic" data. *Genetics in Medicine*, 15(12), 997-1003.

[267] Kaye, J., & Hawkins, N. (2014). Data sharing policy design for consortia: challenges for sustainability. *Genome Medicine*, 6(1), 1-8.

[268] National Institutes of Health. (2014). NIH genomic data sharing policy. Notice Number NOT-OD-14-124. Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html [cited 26 October 2020].

[269] Pita, J., Jain, M., Ordónez, F., Portway, C., Tambe, M., Western, C., ... & Kraus, S. (2009). Using game theory for Los Angeles airport security. *AI Magazine*, 30(1), 43-43.

[270] Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., DiRenzo, J., ... & Meyer, G. (2012, June). Protect: A deployed game theoretic system to protect the ports of the united states. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 13-20).

[271] IBM Security. (2016). 2016 Cost of Data Breach Study: Global Analysis. Available from: http://www.ibm.com/security/data-breach/ [cited 18 November 2016].

[272] IBM Security. (2020). Cost of a Data Breach Report 2020. Available from: http://www.ibm.com/security/data-breach/ [cited 26 October 2020].

[273] Smit, P. R., Meijer, R. F., & Groen, P. P. J. (2004). Detection rates, an international comparison. *European Journal on Criminal Policy and Research*, 10(2-3), 225-253.

[274] US Equal Employment Opportunity Commission. (2010). Genetic Information Nondiscrimination Act of 2008, Final Rule. *Fed Regist*, 75(216), 68912-68939.

[275] US Equal Employment Opportunity Commission. (2011). Regulations To Implement the Equal Employment Provisions of the Americans With Disabilities Act, as Amended; Final Rule. *Fed Regist*, 76(58), 16978-17017.

[276] Rehm, H. L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews Genetics*, 14(4), 295-300.

[277] Green, E. D., Guyer, M. S., and National Human Genome Research Institute. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333), 204-213.

[278] Sanderson, S. C., Linderman, M. D., Suckiel, S. A., Diaz, G. A., Zinberg, R. E., Ferryman, K., ... & Schadt, E. E. (2016). Motivations, concerns and preferences of personal genome sequencing research participants: baseline findings from the HealthSeq project. *European Journal of Human Genetics*, 24(1), 14-20.

[279] Hull, S. C., Sharp, R. R., Botkin, J. R., Brown, M., Hughes, M., Sugarman, J., ... & Wilfond, B. S. (2008). Patients' views on identifiability of samples and informed consent for genetic research. *The American Journal of Bioethics*, 8(10), 62-70.

[280] Kaufman, D. J., Murphy-Bollinger, J., Scott, J., & Hudson, K. L. (2009). Public opinion about the importance of privacy in biobank research. *The American Journal of Human Genetics*, 85(5), 643-654.

[281] Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., ... & Pasternak, S., & International HalMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851-861.

[282] Felch, J. (2008). DNA databases blocked from the public. *Los Angeles Times*, 29, A31.

[283] Knoppers, B. M. (2014). International ethics harmonization and the global alliance for genomics and health. *Genome Medicine*, 6(2), 1-3.

[284] Torres-Español, M., Anvar, S. Y., & Sobrido, M. J. (2016). Variations in the genome: the mutation detection 2015 meeting on detection, genome sequencing, and interpretation. *Human Mutation*, 37(10), 1106-1109.

[285] Bellman, R. (1957). A Markov decision process. Journal of Mathematical Mechanics, 6, 679-684.

[286] Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. Technology Press of Massachusetts Institute of Technology, Cambridge, MA, USA.

[287] Letchford, J., & Vorobeychik, Y. (2013, May). Optimal interdiction of attack plans. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 199-206).

[288] Xia, W., Kantarcioglu, M., Wan, Z., Heatherly, R., Vorobeychik, Y., & Malin, B. (2015, October). Process-driven data privacy. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1021-1030).

[289] Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10), 1095-1100.

[290] Fink, A. M. (1964). Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I (Mathematics)*, 28(1), 89-93.

[291] Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994* (pp. 157-163). Morgan Kaufmann.

[292] Shabani, M., & Borry, P. (2018). Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *European Journal of Human Genetics*, 26(2), 149-156.

[293] Shabani, M., & Marelli, L. (2019). Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO reports*, 20(6), e48316.

[294] Anindya, I. C., Roy, H., Kantarcioglu, M., & Malin, B. (2017, November). Building a Dossier on the Cheap: Integrating Distributed Personal Data Resources Under Cost Constraints. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1549-1558).

[295] Threlkeld. Free databases Ysearch and Mitosearch closing May 24[Blog]. (2018). Available from: https://casestone.com/threlkeld/home/latest-news/94-free-databases-ysearch-and-mitosearch-closing-may-24 [cited 31 October 2020].

[296] Malin, B., Karp, D., & Scheuermann, R. H. (2010). Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative Medicine*, 58(1), 11-18.

[297] US Equal Employment Opportunity Commission. (2016). Genetic Information Nondiscrimination Act of 2008, Final Rule. *Fed Regist*, 81, 31143-31159.

[298] Clayton, E. W., Evans, B. J., Hazel, J. W., & Rothstein, M. A. (2019). The law of genetic privacy: applications, implications, and limitations. *Journal of Law and the Biosciences*, 6(1), 1-36.

[299] Branson, J., Good, N., Chen, J. W., Monge, W., Probst, C., & El Emam, K. (2020). Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations. *Trials*, 21(1), 1-9.

[300] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., ... & Liu, B. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos Med*, 12(3), e1001779.

[301] Artyushina, A. (2020, August). The EU is launching a market for personal data. Here's what that means for privacy. In MIT Technology Review. Available from: https://www.technologyreview.com/2020/08/11/1006555/eu-data-trust-trusts-project-privacy-policy-opinion [cited 31 October 2020].

[302] Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M., ... & Tam, P. K. H. (2003). The International HapMap Project. *Nature*, 426(6968), 789-796.

[303] Peng, B., & Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18), 3686-3687.

[304] Peng, B., & Amos, C. I. (2008). Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics*, 24(11), 1408-1409.

[305] Comenetz, J. (2016). Frequently occurring surnames in the 2010 census. US Census Bureau. Available from: https://www.census.gov/topics/population/genealogy/data/2010_surnames.html [cited 31 October 2020].

[306] US Census Bureau (2019). Population, population change, and estimated components of population change: April 1, 2010 to July 1, 2019 (NST-EST2019-alldata). US Census Bureau. Available from: https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html [cited 31 October 2020].

[307] Sánchez-Diz, P., Alves, C., Carvalho, E., Carvalho, M., Espinheira, R., García, O., ... & Silva, C. (2008). Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study. *International Journal of Legal Medicine*, 122(6), 529-533.

[308] US Census Bureau (2016). Source of income in 2015- people 15 years old and over, by income of specified type in 2015, age, race, Hispanic origin, and sex, in Current Population Survey, 2016 Annual Social and Economic Supplement, US Census Bureau. Available from: https://www2.census.gov/programs-surveys/cps/tables/pinc-08/2016/pinc08_1_1_1.xls [cited 31 October 2020].

[309] Sraders, A. (2019) What is the middle class? Income and range. TheStreet. Available from: https://www.thestreet.com/personal-finance/what-is-middle-class-14833259 [cited 31 October 2020].

[310] Kimura, M., & Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences*, 75(6), 2868-2872.

[311] Martin, J. A., Hamilton, B. E., Osterman, M. J. K., Driscoll, A. K. (2019). Births: Final data for 2018 [Table 3]. *National Vital Statistics Reports*, 68, 13. https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_13-508.pdf [cited 31 October 2020].

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Derivation of Game Theoretic Methods

Table A.1 Notation used throughout this section, ordered by their chronological introduction.

Notation	Description
n	The number of individuals in the study, or the size of the set I
m	The number of SNPs available for sharing, or the size of the set J
R	The reference dataset sampled from the underlying population containing the study dataset
$r[i,j]$	The number of minor alleles in SNP j of individual i in the reference dataset
n_r	The number of individuals in the reference dataset
$\tilde{f}[j]$	The manipulated minor allele frequency of SNP j in the study dataset
L_U	The utility loss of shared data, equals 1 minus the utility score U
v	A vector in the L_p space
$u[j]$	The utility associated with SNP j
k	The number of shared SNPs
T	The set of targeted individuals
X	The target dataset
$x[i,j]$	The number of minor alleles in SNP j of target i
n_x	The number of targets available to the recipient
b	A combination of targets as an attacking strategy, represented by a binary vector of size n_x
$s(i)$	The ground truth whether target i is in the study, or the membership indicator
$t(i)$	The ground truth whether individual i in the study dataset is targeted, or the targeting indicator
N	The ground truth of the number of successfully attacked targets
R_P	The privacy risk associated with the shared data
P	The privacy protection score of the shared data (equal to 1 minus the privacy risk R_P)
Y_R	The ground truth of the recipient's payoff
B_R	The ground truth of the recipient's benefit
C_R	The recipient's cost
$b^*(g)$	The recipient's optimal attacking strategy, given the sharer's strategy g, for all targets
$\bar{Y}_R$	The recipient's payoff, as estimated by the recipient
$\bar{B}_R$	The recipient's benefit, as estimated by the recipient
$\bar{N}_R$	The number of successfully attacked targets, as estimated by the recipient
n_p	The number of individuals in the underlying population containing this study
θ_l	The threshold of the likelihood ratio test
Y_S	The ground truth of the sharer's payoff
C_S	The ground truth of the sharer's cost
ω	The privacy-utility ratio, $\omega = L_S n_x p / H$
$\hat{P}$	The estimated privacy protection score of the shared data
Y_R^D	The ground truth of the recipient's payoff (attacking only targets in the study dataset)
C_R^D	The recipient's cost (attacking only targets in the study dataset)
$\bar{Y}_R^D$	The recipient's payoff, as estimated by the recipient (attacking only targets in the study dataset)
$\bar{B}_R^D$	The recipient's benefit, as estimated by the recipient (attacking only targets in the study dataset)
$\bar{N}_R^D$	The number of successfully attacked targets as estimated by the recipient (regarding the study dataset)
$\bar{f}[j]$	The minor allele frequency of SNP j in the underlying population containing this study

In this section, we derive the payoff functions for the genomic data sharing Stackelberg game. In the process, we provide justification for the representation of the functions. Table A.1 provides a summary of the notation commonly used throughout this section (in addition to those shown in Table 3.1 in Chapter 3), ordered by their chronological introduction below.

A.1.1 Oracle

For computational purposes, we begin by laying out several assumptions. First, we assume that the allele frequencies of all single nucleotide polymorphisms (SNPs), in the entire population as well as several subpopulations, can be queried easily. In this setting, we define a subpopulation as a group of people who share a certain set of attributes that are deemed not to be sensitive. Second, we assume that individual-level genotype information is accessible at a certain cost.

Real-world representations of such an oracle include the 1000 Genomes Project [230] and the International HapMap Project[28] [302].

A.1.2 Shared Data

Let us assume that the sharer has collected a pool of genotypes D (i.e., the study dataset) with the same sensitive phenotype (e.g., a positive diagnosis of a sexually transmitted disease, such as syphilis or acquired immune deficiency syndrome (AIDS)) containing n individuals and m SNPs. The sharer is incentivized to share the summary statistics along with the sensitive phenotype associated with this pool for research publications and grant applications, as has become expected for genome-wide association studies (GWAS), and more recently whole genome and exome studies. As has been shown in various studies, attacks can leverage such summary statistics to infer if an individual is a participant of a study to breach their privacy.

It should be recognized that there are mainly two kinds of summary statistics published in GWAS: a) measurements of linkage disequilibrium (LD) (i.e., in the form of a correlation matrix) and b) minor allele

[28] On June 16, 2016, the legacy HapMap site was taken offline by the National Center for Biotechnology Information (NCBI), due to security vulnerabilities that were exposed. However, the archived HapMap data will continue to be available via an NIH-managed FTP site, hosted at ftp://ftp.ncbi.nlm.nih.gov/hapmap/.

frequencies (MAFs). A correlation matrix can be exploited to reconstruct the entire dataset as shown in attack of Wang and colleagues.[3] As a result, the sharer only releases the MAFs for m SNPs in the pool, $f = [f[1], \cdots, f[j], \cdots, f[m]]$ and the size of the pool n. As noted above, it is assumed that there is a reference dataset R sampled from the underlying population containing this pool, in which the MAFs of these SNPs are $\hat{f} = [\hat{f}[1], \cdots, \hat{f}[j], \cdots, \hat{f}[m]]$. Note that $\forall j, \hat{f}[j] \leq 0.5$.

Now, let us say that the study dataset D containing n individuals and m SNPs is represented as a matrix:

$$D = \begin{bmatrix} d[1,1] & \cdots & d[1,m] \\ \vdots & \ddots & \vdots \\ d[n,1] & \cdots & d[n,m] \end{bmatrix} \tag{A.1}$$

in which $d[i,j] \in \{0,1,2\}$ is a ternary variable representing the number of minor alleles in the jth SNP of the ith individual in the study dataset. The summary statistics f, based on the dataset D, are thus calculated as:

$$f[j] = \frac{1}{2n} \sum_{i=1}^{n} d[i,j], \quad \forall j \in [1, \cdots, m]. \tag{A.2}$$

We represent the reference dataset R, which contains n_r targets across m SNPs, as a matrix as well:

$$R = \begin{bmatrix} r[1,1] & \cdots & r[1,m] \\ \vdots & \ddots & \vdots \\ r[n_r,1] & \cdots & r[n_r,m] \end{bmatrix} \tag{A.3}$$

in which $r[i,j] \in \{0,1,2\}$ is a ternary variable representing the number of minor alleles in the jth SNP of the ith individual in the reference dataset. The summary statistics $\hat{f}$, based on the dataset R, are thus calculated as:

$$\hat{f}[j] = \frac{1}{2n_r} \sum_{i=1}^{n_r} r[i,j], \quad \forall j \in [1, \cdots, m]. \tag{A.4}$$

A.1.3 Sharer's Strategy Space and Data Utility Score

The sharer has the freedom to apply any data manipulation strategies to the allele frequencies before releasing the data. These strategies are varied and can include (but are not limited to) noise addition, reducing the precision of allele rates, complete suppression (also referred to as redaction). All of these strategies enhance the resilience of summary data against known attacks (e.g., Sankararaman *et al.*'s attack[4]) at the cost of lowering the quality of the data. This loss of data utility can be defined in a L_p

space as the distance between the shared manipulated MAFs $\tilde{f} = [\tilde{f}[1], \cdots, \tilde{f}[j], \cdots, \tilde{f}[m]]$ and the original MAFs f:

$$L_U(\tilde{f}) = \|\tilde{f} - f\|_p \tag{A.5}$$

where the L_p norm is defined as:

$$\|v\|_p = \left(\sum_{j=1}^{m} |v[j]|^p\right)^{1/p} \tag{A.6}$$

in which $v = [v[1], \cdots, v[j], \cdots, v[m]]$ represents any vector in the L_p space.

We define the utility score of the shared data $U \in [0,1]$ as a function of shared frequencies $\tilde{f}$, given the original frequencies f:

$$U(\tilde{f}) = 1 - \frac{L_U(\tilde{f})}{\max_{\tilde{f}} L_U(\tilde{f})}. \tag{A.7}$$

For simplicity, let us set $p = 1$ (i.e., we use L_1-norm). Then the utility score is additive across all frequencies. Furthermore, we assume that the sharer only applies a data suppression (i.e., we assume that only truthful MAF frequencies are disclosed). Now, let us say that if the sharer suppresses (*i.e.*, does not share) the j^{th} frequency, then the recipient will query it from the oracle for the corresponding frequency in the reference population, $\hat{f}[j]$. A data suppression sharing strategy can be represented as $g = [g[1], \cdots, g[j], \cdots, g[m]]$, when $g[j]$ is a binary variable. Note that $g[j] = 0$ if the sharer suppresses the j^{th} frequency and $g[j] = 1$ if the sharer releases it. Recognize that the size of the strategy space is thus 2^m. As a result, if only data suppression strategies are considered, then the shared frequencies $\tilde{f}$ can be represented as a function of the strategy g:

$$\tilde{f}[j](g) = f[j]g[j] + \hat{f}[j](1 - g[j]). \tag{A.8}$$

By substituting Equation A.6 and Equation A.8 into Equation A.5, the data utility loss L_U can be represented as a function of the suppression strategy g:

$$L_U(g) = \sum_j \left|f[j]g[j] + \hat{f}[j](1 - g[j]) - f[j]\right| = \sum_j (1 - g[j])\left|\hat{f}[j] - f[j]\right|. \tag{A.9}$$

We define the utility associated with j^{th} allele frequency as $u[j] = \left|\hat{f}[j] - f[j]\right|$. The maximal data utility loss is maximized when $L_U = \sum_j u[j]$, which occurs when the sharer does not release anything (*i.e.*, $g = 0$ or $\tilde{f} = \hat{f}$). By substituting Equation A.9 into Equation A.7, the utility score of the shared dataset U can be represented as a function of the sharer's strategy g:

$$U(g) = 1 - \frac{\sum_j (1 - g[j])u[j]}{\sum_j u[j]} = \frac{\sum_j g[j]u[j]}{\sum_j u[j]} = \frac{\sum_j g[j]\left|\hat{f}[j] - f[j]\right|}{\sum_j\left|\hat{f}[j] - f[j]\right|} \tag{A.10}$$

which is precisely Equation 3.6 combined with Equation 3.8 in Chapter 3.

For comparison with Sankararaman *et al.*'s SNP suppression policy,[4] we need to redefine the utility score as a function of the number of shared SNPs k (assuming that the sharer only releases the top-scoring k SNPs). After reordering the SNPs according to their associated utility u, the sharer's strategy can be represented as a step function of $k \in [1, \cdots, m]$:

$$g[j](k) = \begin{cases} 1, & j \leq k. \\ 0, & j > k. \end{cases} \tag{A.11}$$

Notice that the size of the strategy space will reduce to m. Thus, the utility of the shared data as a function of k is:

$$U(k) = \frac{\sum_{j=1}^{k} u[j]}{\sum_{j=1}^{m} u[j]} = \frac{\sum_{j=1}^{k} \left|\hat{f}[j] - f[j]\right|}{\sum_{j=1}^{m} \left|\hat{f}[j] - f[j]\right|}. \tag{A.12}$$

A.1.4 Recipient's Target Set

We define the target set T as the set of identified genomes the recipient has obtained and are available for the membership inference attack. Let us say that the target dataset X, which contains n_x targets and m SNPs, is represented as a matrix:

$$X = \begin{bmatrix} x[1,1] & \cdots & x[1,m] \\ \vdots & \ddots & \vdots \\ x[n_x,1] & \cdots & x[n_x,m] \end{bmatrix} \tag{A.13}$$

in which $x[i,j] \in \{0,1,2\}$ is a ternary variable that represents the number of minor alleles in the j^{th} SNP site of i^{th} target.

A.1.5 Recipient's Strategy Space

We define a membership inference attack as the attempt of a recipient of the data to infer whether a targeted (identified) individual is in the study. The attack is successful if, and only if, the target is actually in the study. The recipient’s strategy can be represented as a combination of targets, or a vector $b = [b[1], \cdots, b[i], \cdots, b[n_x]]$ in which each element is a binary variable, where $b[i]$ is 1 if target i is attacked and 0 otherwise. Since the sharer does not know the target set, they are concerned with the decision whether each individual in the study is attacked conditional on it being a target, which is represented by a vector $a = [a[1], \cdots, a[i], \cdots, a[n]]$ in which each element is a binary variable, with $a[i] = 1$ if individual i in the study is attacked and 0 otherwise.

A.1.6 Privacy Protection Score

The ground truth about which targets are in the study is represented as $s = [s[1], \cdots, s[i], \cdots, s[n_x]]$, where $s[i]$ is 1 if target i is actually in the study and 0 otherwise. The ground truth about which individual in the study is targeted is represented as $t = [t[1], \cdots, t[i], \cdots, t[n]]$, where $t[i]$ is 1 if individual i in the study is targeted and 0 otherwise. Thus, the number of successfully attacked targets N can be defined as:

$$N(a) = \sum_{i=1}^{n} a[i]t[i] = \sum_{i=1}^{n_x} b[i]s[i] = N(b). \tag{A.14}$$

The privacy risk R_P is simply defined as the number of successfully attacked targets N divided by its maximal value $\max_a N(a)$ (i.e., the number of individuals in the study that are targeted). Accordingly, the privacy protection score P is defined as:

$$P(a) = 1 - R_P(a) = 1 - \frac{N(a)}{\max\limits_a N(a)} = 1 - \frac{\sum_{i=1}^{n} a[i]t[i]}{\sum_{i=1}^{n} t[i]}. \tag{A.15}$$

A.1.7 Recipient's Payoff and the Optimal Strategy

In our model, the recipient is economically motivated by his or her monetary payoff Y_R. This is defined as the difference between his or her benefit B_R and his or her cost C_R. Both of these are represented as functions of his or her strategy b:

$$Y_R(b) = B_R(b) - C_R(b). \tag{A.16}$$

Let us assume that the recipient gains a certain amount of money $G_R[i]$ if attacked target i is actually in the study. Then the recipient’s benefit B_R can be represented as a function of his or her strategy b:

$$B_R(b) = \sum_{i=1}^{n_x} G_R[i]b[i]s[i]. \tag{A.17}$$

For simplicity, let us further assume that the recipient gains the same amount of money G_R for each successfully attacked target. Then, the recipient’s benefit $B_R(b)$ is proportional to the number of successfully attacked targets N:

$$B_R(b) = G_R \sum_{i=1}^{n_x} b[i]s[i] = G_R \cdot N(b). \tag{A.18}$$

In our model, we assume (without loss of generality) that the recipient mainly has two types of cost: 1) a cost for accessing a target c_a and 2) an expected cost due to a penalty c_p. The latter is calculated as the product of the crime detection rate (usually about twenty percent[5]) and the penalty which is enforced by regulation, laws and/or a data use agreement (DUA) signed by the recipient. Let us assume that the crime detection rate is 0.2. It is notable that, once an attack is detected (regardless of its success status), it is regarded as a violation of the DUA and would be penalized. Thus, the recipient's cost $C_R(b)$ is represented as a function of his or her strategy b:

$$C_R(b) = \left(c_a + c_p\right)\sum_{i=1}^{n_x} b[i]. \tag{A.19}$$

As a result, the recipient's payoff Y_R is:

$$Y_R(b) = G_R\sum_{i=1}^{n_x} b[i]s[i] - \left(c_a + c_p\right)\sum_{i=1}^{n_x} b[i] = \sum_{i=1}^{n_x} b[i]\left(G_R \cdot s[i] - c_a - c_p\right). \tag{A.20}$$

If the recipient is aware of the membership (*i.e.*, indicator $s[i]$ for each target i), then his or her optimal strategy would be the one that maximizes the payoff:

$$b^* = \underset{b}{\operatorname{argmax}}\, Y_R(b). \tag{A.21}$$

This would imply:

$$b^*[i] = \begin{cases} 1, & G_R \cdot s[i] - c_a - c_p > 0, \\ 0, & G_R \cdot s[i] - c_a - c_p \leq 0, \end{cases} \quad \forall i \in T. \tag{A.22}$$

However, in the scenario we posit in the main manuscript, the recipient is unaware of the membership indicator s. Instead, he or she can only approximate the payoff Y_R by estimating the number of successfully attacked targets:

$$\bar{Y}_R(b) = \bar{B}_R(b) - C_R(b) = G_R \cdot \bar{N}_R(b) - C_R(b) \tag{A.23}$$

in which the number of successfully attacked targets N is estimated using its expected value:

$$\bar{N}_R(b) = E\big(N(b)\big) = \sum_{i=1}^{n_x} b[i]E(s[i]). \tag{A.24}$$

In this scenario, the expected value of the membership indicator is the probability that target i is in the study. A Bayesian approach is used to estimate the posterior probability of membership given the target's genotype information, which is based on a likelihood ratio test introduced by Sankararaman *et al.* [36].

$$E(s[i]) = Pr\{i \in D|X[i]\}, \quad \forall i \in T. \tag{A.25}$$

A prior probability of membership $p[i] = Pr\{i \in D\}$ is needed for the Bayesian approach, which could be estimated using the sampling rate n/n_p, if the target set is randomly sampled from the underlying population containing this pool whose size is n_p:

$$p[i] \cong p = \frac{1}{n_x}\sum_{i=1}^{n_x} s[i] \cong \frac{n}{n_p}, \quad \forall i \in T. \tag{A.26}$$

The likelihood ratio for target i with alleles $x[i] = [x[i,1], \cdots, x[i,j], \cdots, \ x[i,m]]$ is a function of the sharing strategy g. This is the reason why the recipient's strategy is influenced by the sharer's strategy in the Stackelberg game:

$$l(i,g) = \frac{\prod_{j\in J} f[j]^{x[i,j]g[j]}(1-f[j])^{(2-x[i,j])g[j]}}{\prod_{j\in J} \hat{f}[j]^{x[i,j]g[j]}\left(1-\hat{f}[j]\right)^{(2-x[i,j])g[j]}}, \quad \forall i \in T. \tag{A.27}$$

With some rearrangement, it can be seen that this is equivalent to:

$$l(i,g) = \prod_{j\in J}\left(\frac{f[j]}{\hat{f}[j]}\right)^{x[i,j]g[j]}\left(\frac{1-f[j]}{1-\hat{f}[j]}\right)^{(2-x[i,j])g[j]}, \quad \forall i \in T. \tag{A.28}$$

It is often easier (and more numerically stable) to deal with the log-likelihoods, which yields the log-likelihood difference:

$$\begin{aligned}\log\big(l(i,g)\big) = \sum_{j\in J} & x[i,j]g[j]\big(\log f[j] - \log \hat{f}[j]\big) \\ & + (2-x[i,j])g[j]\big(\log(1-f[j]) - \log\big(1-\hat{f}[j]\big)\big).\end{aligned} \tag{A.29}$$

As such, the posterior probability that target $i \in D$ based on genotype information $X[i]$ is derived according to the Bayes' Rule:

$$Pr\{i \in D|X[i]\} = \frac{Pr\{X[i]|i \in D\}Pr\{i \in D\}}{Pr\{X[i]\}}. \tag{A.30}$$

Note that

$$Pr\{X[i]\} = \prod\nolimits_{j\in J} \hat{f}[j]^{x[i,j]g[j]}\left(1-\hat{f}[j]\right)^{(2-x[i,j])g[j]} \tag{A.31}$$

and

$$Pr\{X[i]|i \in D\} = \prod\nolimits_{j\in J} f[j]^{x[i,j]g[j]}(1-f[j])^{(2-x[i,j])g[j]} \tag{A.32}$$

if we make the approximation that this probability, $Pr\{X[i]|i \in D\}$, is equivalent to the probability of drawing an individual randomly from D, and all SNPs are independent. We highlight that this is the same approximation made by Sankararaman *et al.* [36].

Thus, the posterior probability is a function of the sharing strategy g, which is denoted and represented as:

$$\pi(i,g) = Pr\{i \in D|X[i]\} = p[i]l(i,g). \tag{A.33}$$

The recipient's optimal strategy would be a function of the sharing strategy g, that maximizes his or her estimated payoff:

$$b^*(g) = \underset{b}{\operatorname{argmax}}\, \hat{Y}_R(b). \tag{A.34}$$

The recipient's optimal strategy on any target is independent of the strategy selected for other targets:

$$b^*[i](g) = \begin{cases} 1, & G_R \cdot p[i]l(i,g) - c_a - c_p > 0 \\ 0, & G_R \cdot p[i]l(i,g) - c_a - c_p \leq 0 \end{cases} \quad \forall i \in T \tag{A.35}$$

which is equivalent to:

$$b^*[i](g) = \begin{cases} 1, & l(i,g) > \dfrac{(c_a + c_p)}{G_R \cdot p} \\ 0, & l(i,g) \leq \dfrac{(c_a + c_p)}{G_R \cdot p} \end{cases} \quad \forall i \in T \tag{A.36}$$

if $\forall i \in T, p = p[i]$, which implicitly determines the best threshold of the likelihood ratio test for the recipient:

$$\theta_l = \frac{(c_a + c_p)}{G_R \cdot p}. \tag{A.37}$$

A.1.8 Sharer's Payoff and the Optimal Strategy

The objective of a sharer is to simultaneously optimize both privacy protection and data utility. There are two classes of approaches that can be invoked to solve such a multi-objective optimization problem. The first is to optimize one metric while constraining the others, as is done by Sankararaman *et al.*[4] Specifically, they fix the maximal privacy risk and maximize the data utility (in terms of the number of released SNP allele frequencies).

An alternative is to combine multiple objectives together as a single objective. One way to combine utility and privacy is to consider their monetary impact, in the form of the sharer's payoff Y_S. This is the difference between the sharer's benefit B_S and his or her cost C_S:

$$Y_S(g,a) = B_S(g) - C_S(a). \tag{A.38}$$

Let us assume the sharer's benefit B_S is proportional to the utility of the shared data.

$$B_S(g) = H \cdot U(g) \tag{A.39}$$

where H is the maximal benefit the sharer can obtain, when all of the MAFs are shared as their original values. In this case, the sharer gains 0 when no genomic summary data are shared.

On the other side, the sharer ensures all participants' privacy and safety through a contract that enforces that the sharer pays any loss caused by the data sharing. Let us assume the loss L_S for each individual in the study is the same and constant. Then the sharer's cost C_S is:

$$C_S(a) = L_S \cdot N(a) = L_S \sum_{i=1}^{n} a[i]t[i]. \tag{A.40}$$

Notice that the sharer's cost is proportional to the privacy risk score. This can be simplified based on Equation A.15 and Equation A.25:

$$C_S(a) = L_S \cdot N(a) = L_S\, R_P(a) \max_a N(a) = L_S\big(1 - P(a)\big) n_x p. \tag{A.41}$$

Notice further that H, L_S, n_x, and p are all fixed parameters instead of functions of the sharer's or the recipient's strategy. If we define a privacy-utility ratio $\omega = L_S n_x p / H$, the sharer's payoff could be represented as a function of the utility score U and the privacy score P:

$$Y_S(g, a) = H \cdot U(g) - \omega \mathrm{H}\big(1 - P(a)\big) = H(U(g) + \omega P(a) - \omega). \tag{A.42}$$

Since the sharer does not know the targeting indicator t, he or she can only estimate the payoff by estimating the number of successfully attacked targets:

$$\hat{Y}_S(g, a) = B_S(g) - \hat{C}_S(a) = B_S(g) - L_S \cdot \hat{N}_S(a). \tag{A.43}$$

The number of successfully attacked targets is estimated using an expected value:

$$\hat{N}_S(a) = E\big(N(a)\big) = \sum_{i=1}^{n} a[i] E\big[t[i]\big]. \tag{A.44}$$

The expected value of the targeting indicator is the probability that individual i in the study dataset is targeted:

$$E(t[i]) = \tau[i] \cong \tau = \frac{1}{n} \sum_{t=1}^{n} t[i] = \frac{1}{n} \sum_{i=1}^{n_x} s[i] = \frac{n_x p}{n} \cong \frac{n_x}{n_p}. \tag{A.45}$$

Thus, the sharer's payoff is represented as Equation 3.16 in Chapter 3.

The sharer's optimal strategy is the strategy that maximizes his or her estimated payoff:

$$g^* = \operatorname*{argmax}_g \hat{Y}_S\big(g, a^*(g)\big) = \operatorname*{argmax}_g \left(H \cdot U(g) - L_S \sum_{i=1}^{n} a^*[i](g)\tau[i] \right). \tag{A.46}$$

According to Equation A.45 and Equation A.46, we have:

$$g^* = \operatorname*{argmax}_g \left(H \cdot U(g) - L_S n_x p \cdot \frac{1}{n} \sum_{i=1}^{n} a^*[i](g) \right) = \operatorname*{argmax}_g \Big(U(g) + \omega \hat{P}\big(g, a^*(g)\big) \Big) \tag{A.47}$$

where the privacy-utility ratio is $\omega = L_S n_x p / H$, and the estimated privacy protection score is defined as:

$$\hat{P}(a) = 1 - \frac{1}{n} \sum_{i=1}^{n} a[i]. \tag{A.48}$$

A.1.9 Recipient's Payoff and the Optimal Strategy from the Sharer's Perspective

The recipient's optimal attacking strategy $a^*(g)$ can be simulated by the sharer, assuming all individuals in the study are targeted (*i.e.*, the set of targets include all individuals in the study):

The recipient's optimal attacking strategy is simply:

$$a^*[i](g) = b^*[i](g) \quad \forall i \in I \tag{A.49}$$

According to Equation A.35 and Equation A.36, we have:

$$a^*[i](g) = \begin{cases} 1, & G_R \cdot p[i]l(i,g) - c_a - c_p > 0, \\ 0, & G_R \cdot p[i]l(i,g) - c_a - c_p \leq 0, \end{cases} \quad \forall i \in I \tag{A.50}$$

where the likelihood ratio $l(i,g)$ is redefined as:

$$l(i,g) = \frac{\prod_{j \in J} f[j]^{d[i,j]g[j]}(1 - f[j])^{(2-d[i,j])g[j]}}{\prod_{j \in J} \hat{f}[j]^{d[i,j]g[j]}\left(1 - \hat{f}[j]\right)^{(2-d[i,j])g[j]}}, \quad \forall i \in I. \tag{A.51}$$

According to Equation A.36, Equation A.37, and Equation A.51, if $\forall i \in T, p = p[i]$, we have:

$$a^*[i](g) = \begin{cases} 1, & l(i,g) > \dfrac{\left(c_a + c_p\right)}{G_R \cdot p} = \theta_l, \\ 0, & l(i,g) \leq \dfrac{\left(c_a + c_p\right)}{G_R \cdot p} = \theta_l, \end{cases} \quad \forall i \in I \tag{A.52}$$

Since the sharer does not know the target set, the recipient's payoff resulting from attacking only targets in the study, as estimated by the sharer, can be defined as a function of a, over the set of individuals in the study I.

In doing so, the ground truth of the recipient's payoff (regarding only the study dataset) is defined as:

$$Y_R^D(a) = B_R(a) - C_R^D(a) \tag{A.53}$$

in which the ground truth of the recipient's benefit (regarding only the study dataset) is:

$$B_R(a) = G_R \sum_{i=1}^{n} a[i]t[i] = G_R \cdot N(a) \tag{A.54}$$

and the ground truth of the recipient's cost (from attacking individuals in the study) is:

$$C_R^D(a) = \left(c_a + c_p\right) \sum_{i=1}^{n} a[i]t[i]. \tag{A.55}$$

The recipient's payoff, as estimated by the recipient (regarding only the study dataset), is defined as:

$$\bar{Y}_R^D(a) = \bar{B}_R^D(a) - C_R^D(a) \tag{A.56}$$

The recipient's estimated payoff estimated by the sharer (regarding only the study dataset) is defined as:

$$\hat{Y}_R(g,a) = \hat{B}_R(g,a) - \hat{C}_R(a) \tag{A.57}$$

in which the recipient's cost estimated by the sharer (regarding only the study dataset) is:

$$\hat{C}_R(a) = E\left(C_R^D(a)\right) = \left(c_a + c_p\right)\sum_{i=1}^{n} a[i]\tau[i]. \tag{A.58}$$

and the recipient's estimated benefit estimated by the sharer (regarding only the study dataset) is:

$$\hat{B}_R(g, a) = E\left(\bar{B}_R^D(a)\right) = G_R \cdot \hat{N}_R(g, a) \tag{A.59}$$

in which the number of successfully attacked targets estimated by the recipient, estimated by the sharer (regarding only the study dataset), is:

$$\hat{N}_R(g, a) = E\left(\bar{N}_R^D(b)\right) = \sum\nolimits_{i\in I} p[i]l(i, g)a[i]\tau[i]. \tag{A.60}$$

in which the number of successfully attacked targets, as estimated by the recipient (regarding only the study dataset) is:

$$\bar{N}_R^D(b) = \sum\nolimits_{i\in T} p[i]l(i, g)b[i]s[i] = \sum\nolimits_{i\in I} p[i]l(i, g)a[i]t[i] = \bar{N}_R^D(a). \tag{A.61}$$

The recipient's estimated payoff can also be represented as:

$$\hat{Y}_R(g, a) = \sum_{i\in I} \hat{Y}_R[i](g, a[i]) \tag{A.62}$$

in which the recipient's estimated quantile payoff is represented as:

$$\hat{Y}_R[i](g, a[i]) = \left(G_R \cdot p[i]l(i, g) - c_a - c_p\right)a[i]\tau[i], \quad \forall i \in I. \tag{A.63}$$

Based on this equation, the recipient's optimal strategy can be represented as:

$$a^*[i](g) = \operatorname*{argmax}_{a[i]} \hat{Y}_R[i](g, a[i]) \quad \forall i \in I. \tag{A.64}$$

$$a^*(g) = \operatorname*{argmax}_{a} \hat{Y}_R(g, a). \tag{A.65}$$

A.2 SNP Filtering Pipeline

We created a pipeline to filter out SNPs that lacked sufficient representation in the population and thus were not reliable for privacy-related inferences. The pipeline is shown in Figure A.1 and consists of the following steps:

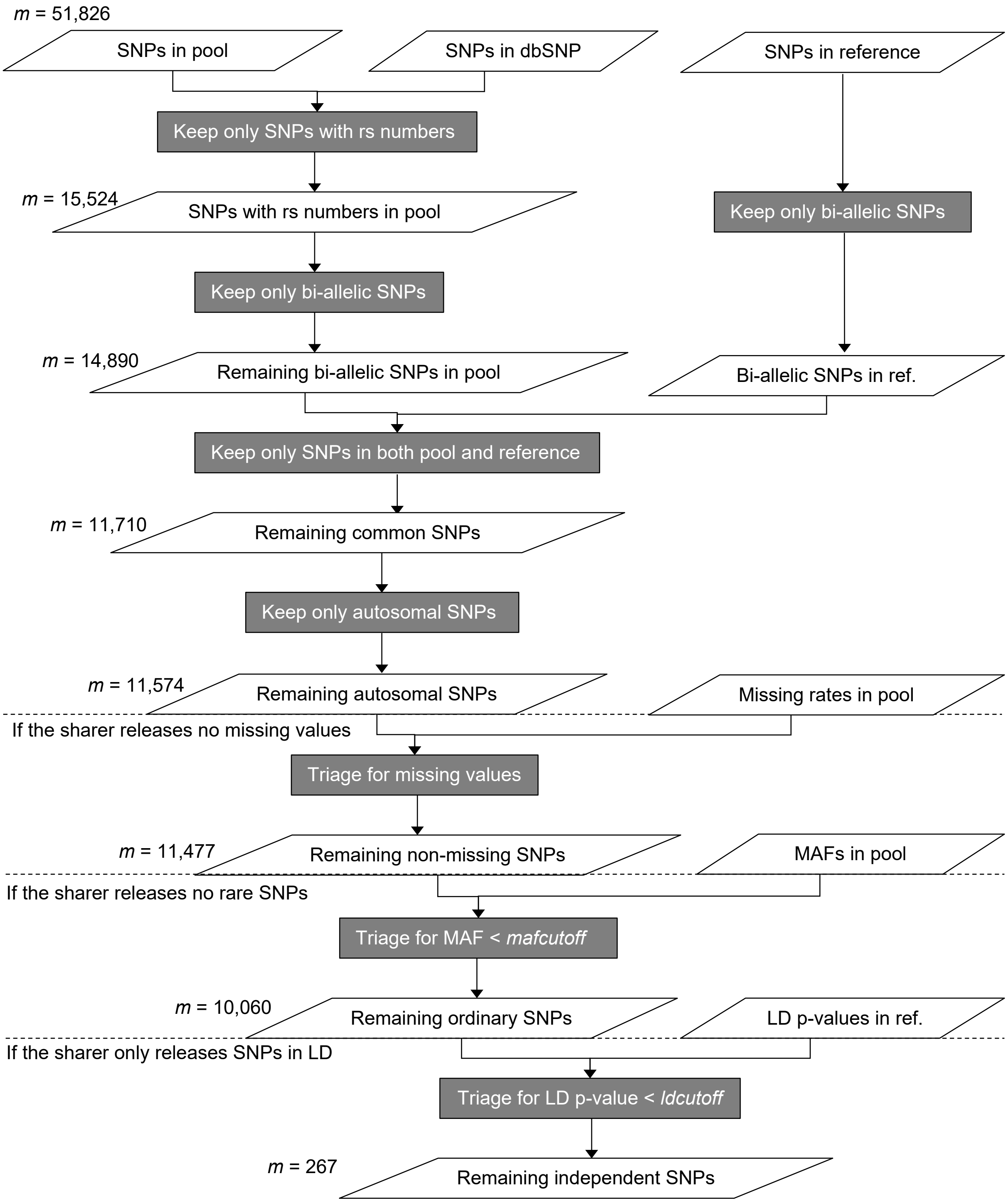

m = 51,826
SNPs in pool
SNPs in dbSNP
SNPs in reference
Keep only SNPs with rs numbers
m = 15,524
SNPs with rs numbers in pool
Keep only bi-allelic SNPs
Keep only bi-allelic SNPs
m = 14,890
Remaining bi-allelic SNPs in pool
Bi-allelic SNPs in ref.
Keep only SNPs in both pool and reference
m = 11,710
Remaining common SNPs
Keep only autosomal SNPs
m = 11,574
Remaining autosomal SNPs
Missing rates in pool
If the sharer releases no missing values
Triage for missing values
m = 11,477
Remaining non-missing SNPs
MAFs in pool
If the sharer releases no rare SNPs
Triage for MAF < *mafcutoff*
m = 10,060
Remaining ordinary SNPs
LD p-values in ref.
If the sharer only releases SNPs in LD
Triage for LD p-value < *ldcutoff*
m = 267
Remaining independent SNPs

Figure A.1 Workflow for the SNP Filtering process in the SPHINX case study.

m is the number of SNPs. SNP, single nucleotide polymorphism. MAF, minor allele frequency. LD, linkage disequilibrium. SPHINX, Sequence and Phenotype Integration Exchange.

- Step 1: Retain SNPs which
    1) have an *rs* number,
    2) are bi-allelic,
    3) are autosomal, and
    4) exist in both the SPHINX and 1000 Genomes datasets;
- Step 2: Retain SNPs with a missing rate smaller than $\bar{\theta} = 0.1$;
- Step 3: Retain SNPs, in the pool, that have a minor allele frequency (MAF) at least as large as $mafcutoff = 0.0001$;
- Step 4: Retain only statistically independent SNPs according to Sankararaman *et al*. [36]. This accomplished through an iterative application of the following subroutines until no SNPs remain:
    - Step 4.1: Compute correlations and corresponding p-values for each SNP pair;
    - Step 4.2: Compute the utility score for each SNP in the form of the absolute difference between MAF in pool and MAF in reference;
    - Step 4.3: Select the SNP with the highest score and filter out all other SNPs that are correlated with it (with a p-value smaller than $ldcutoff = 0.05$) and select the next SNP with the highest score.

Initially, SPHINX contained 51,826 SNPS. After running the data through the filtering pipeline, 267 independent SNPs were retained. Without step 4, if the recipient makes decisions based on all available SNPs, he or she would tend to always attack all individuals in the target set. Figure A.1 provides details on how many SNPs were filtered at each step.

A.3 SPHINX Case Study Experiments

In this section, we provide details of the experimental findings from the SPHINX case study in support of claims made in the main manuscript. We begin with a description of the utility weights applied to the SNPs that survived the filtering pipeline. Next, we report on the likelihood ratios for the most discriminatory SNPs for the SPHINX participants. Finally, we show how the sharer's payoffs are influenced as a function of the number of SNPs released. All game theoretic solutions were discovered by the genetic algorithm presented in Chapter 3.

A.3.1 SNP Utility Weights

To execute the game for the SPHINX resource, we must compute the utility weights w for the 267 remaining SNPs. These weights are shown in Figure A.2. It can be seen that the top 42 (or 15.73%) SNPs cover 91.53% of the overall utility. The remaining majority (84.2%) of the SNPs have very small utility (lower than 0.25%).

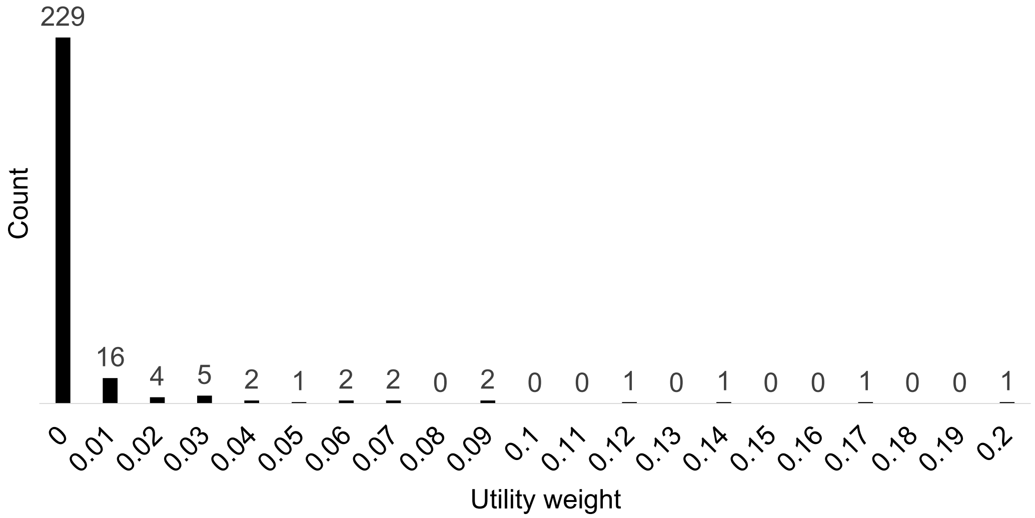


Figure A.2 Histogram of the utility weights of remaining SNPs after filtering in the SPHINX case study. SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

A.3.3 Likelihood Ratios

Next, we compute the log-likelihood ratio for each SNP for each individual in the target set. Since the sharer does not know the target set, he or she optimizes his or her expected payoff based on the estimated target set (which is a combination of the study and reference datasets). Figure A.3 represents the log-likelihood ratios for the top 42 SNPs for all individuals.

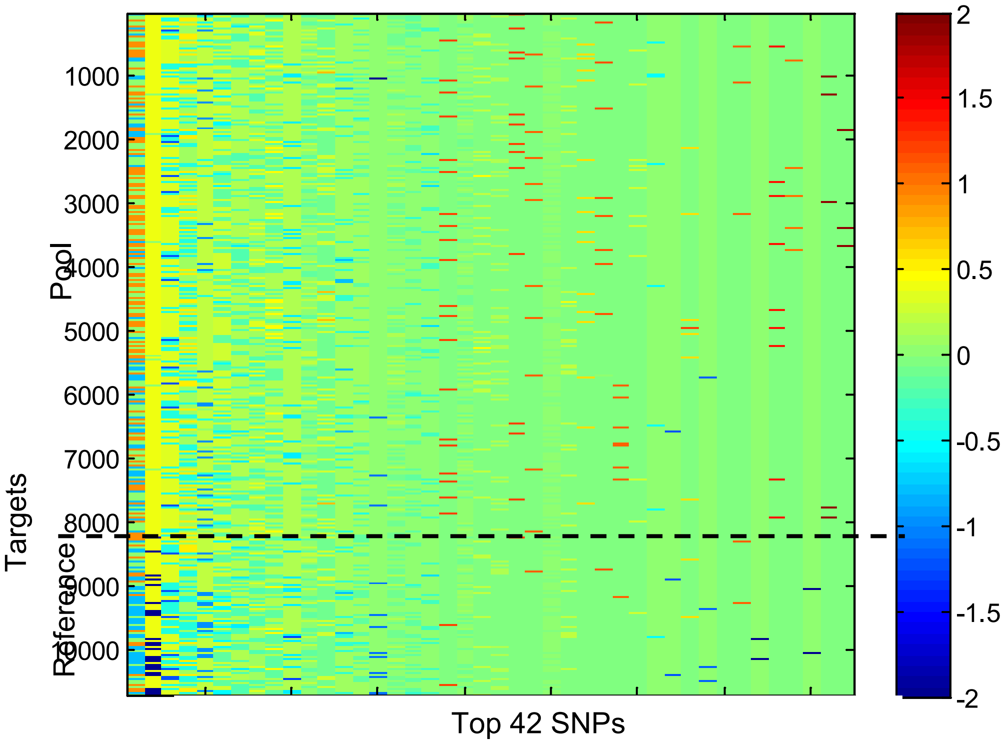


Figure A.3 Log-likelihood ratios for the top 42 SNPs in the estimated target dataset of the SPHINX case study.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.3, the warm colors (*e.g.*, red, orange, yellow) imply that the corresponding target is more likely to be in the pool dataset, while the cool colors (*e.g.*, blue, cyan) imply that the corresponding target is more likely to be in the underlying population. It can be seen that the log-likelihood ratios based on the first several SNPs have more distinguishability than the other SNPs.

A.3.3 Sharer's Payoff in Case Study

The sharer's payoff is computed according to Equation 3.16 in Chapter 3. **Error! Reference source not found.** shows the sharer's estimated payoffs as the sharer releases a growing set of the top SNPs in the SPHINX case study. It can be seen that the sharer's best strategy in the restricted game (where only the top, ordered SNPs can be released) is to release all SNPs for a maximal payoff of $\hat{Y}_S\left(k^*, a^*(k^*)\right) = \hat{Y}_S\left(267, a^*(267)\right) = \$33{,}731$.

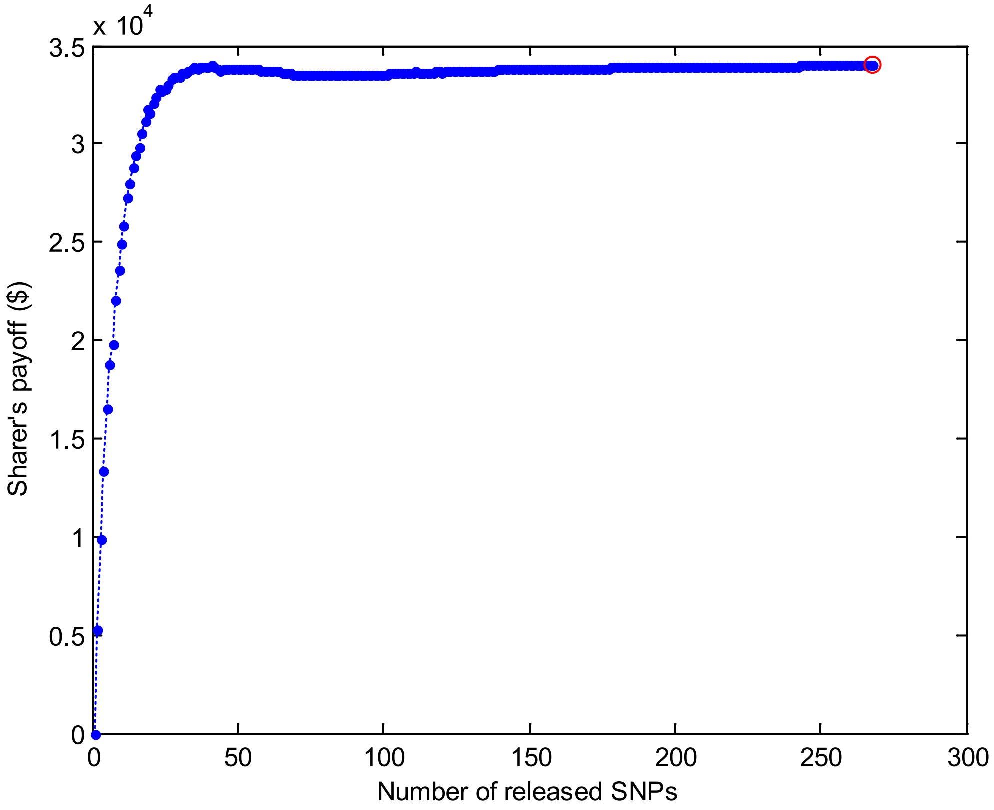


Figure A.4 The sharer's payoffs with a varying number of released SNPs in the SPHINX case study.

The red circle indicates the maximal payoff. SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

By contrast, Sankararaman *et al.*'s approach [36], with $\alpha = 0.1$ and $\beta = 0.5$ leads to the release of 69 and a smaller payoff for the sharer of $\hat{Y}_S\big(69, a^*(69)\big) = \$18{,}798$.

A.4 SPHINX Sensitivity Analysis

In this section, we report on a series of sensitivity analyses on various parameters of the game. These include: 1) expected cost of penalty to the recipient, 2) the maximal benefit for the sharer H only, 3) the gain to the recipient per successful attack G_R only, 4) the loss to the sharer per successful attack L_S only, 5) the gain to the recipient and the loss to the sharer per successful attack assuming they are equal ($G_R = L_S$), 6) the number of targets n_x only, and 7) the prior probability that each target is from the pool dataset

p only. There is no sensitivity analysis on the recipient's cost of access c_a, because it has the same property with the recipient's expected cost of penalty c_p.

A.4.1 Cost of Penalty to the Recipient

Releasing all SNPs is not always the best strategy if the expected cost of penalty changes, especially in the relaxed game. First, we show how various policies induce payoffs under varying penalties. Figure A.5 shows the optimal proportion of released SNPs with the expected cost of penalty varying from \$0 to \$300 in the relaxed game of the SPHINX case study. From Figure A.5 we can see that the optimal proportion of released SNPs changes along with the expected cost of penalty. The optimal proportions of released SNPs in most cases are smaller than 0.9. In the case where the expected cost of penalty to the recipient is \$180, the optimal number of released SNPs is about 0.7.

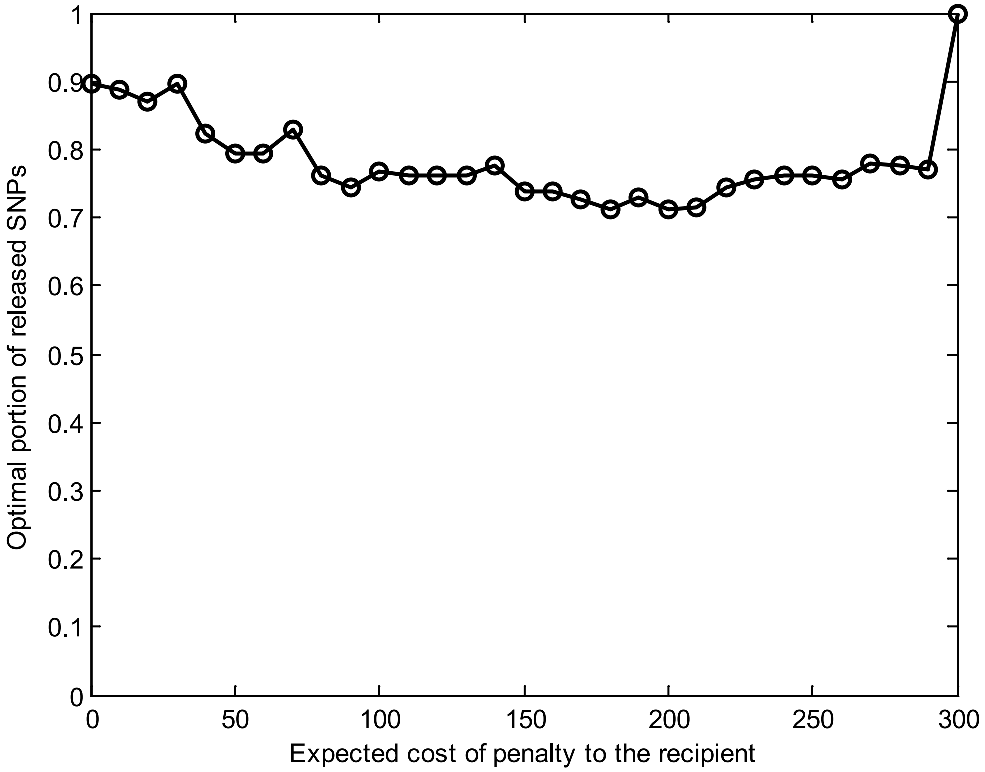


Figure A.5 Optimal portion of released SNPs with a varying penalty in the relaxed game protecting the SPHINX dataset.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

Figure A.6 communicates the same result as Figure 3.10 in Chapter. We present this result for completeness and consistency in this section. Figure A.6 shows the sharer's expected payoff as the expected cost of penalty varies from \$0 to \$300, under four different policies: 1) the game theoretic policy, 2) the game theoretic solution that ensures no attack is successful, 3) the data use agreement (DUA) with a varying penalty (where all SNPs are released), and 4) a SNP suppression policy (based on Sankararaman *et al.*'s approach[4] with no penalty). It can be seen that the game theoretic policy outperforms other policies. The no attack variation of the game theoretic solution yields a relatively high payoff for the sharer in comparison to the SNP suppression strategy when the penalty is greater than \$25.

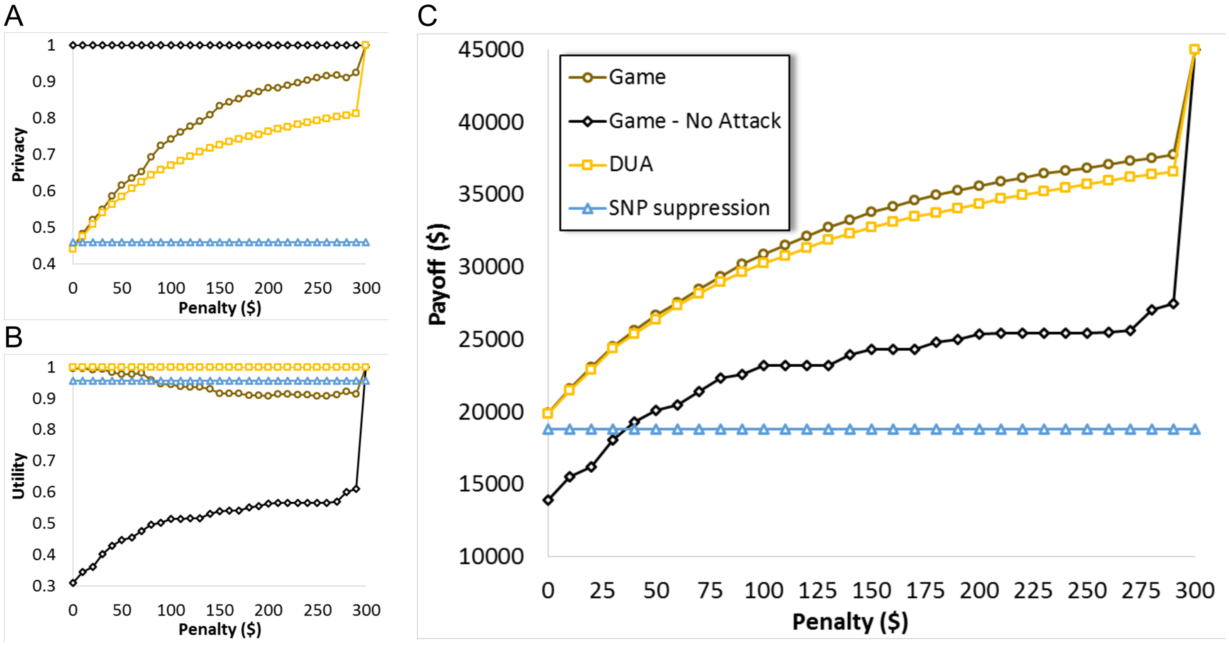


Figure A.6 Comparisons of four protection policies with a varying penalty against the genomic inference attack targeting the SPHINX dataset.

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (yellow lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines) with no penalty, with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

A.4.2 Maximal Benefit to the Sharer

Next, we consider how the maximal benefit to the sharer affects the sharer's optimal strategy and expected payoff. The default value of the maximal benefit to the sharer H is \$45,000. Figure A.7 and Figure A.8 depict the proportion of SNPs released and the payoffs, respectively, as we vary the maximal benefit from \$0 to \$90,000 (holding all other parameters to their default values).

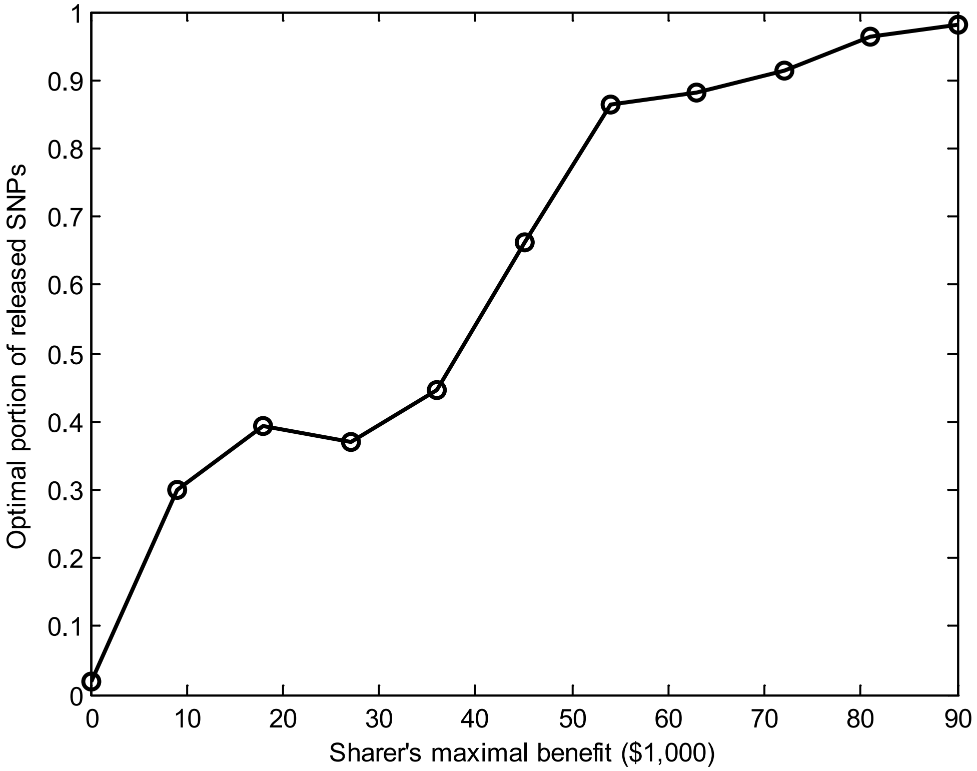


Figure A.7 Optimal portion of released SNPs as a function of the sharer's maximal benefit in the relaxed game protecting the SPHINX dataset.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.7, it can be seen that the sharer's optimal strategy is sensitive to the optimal portion of released SNPs. Specifically, the optimal portion of released SNPs tends to increase along with the sharer's loss.

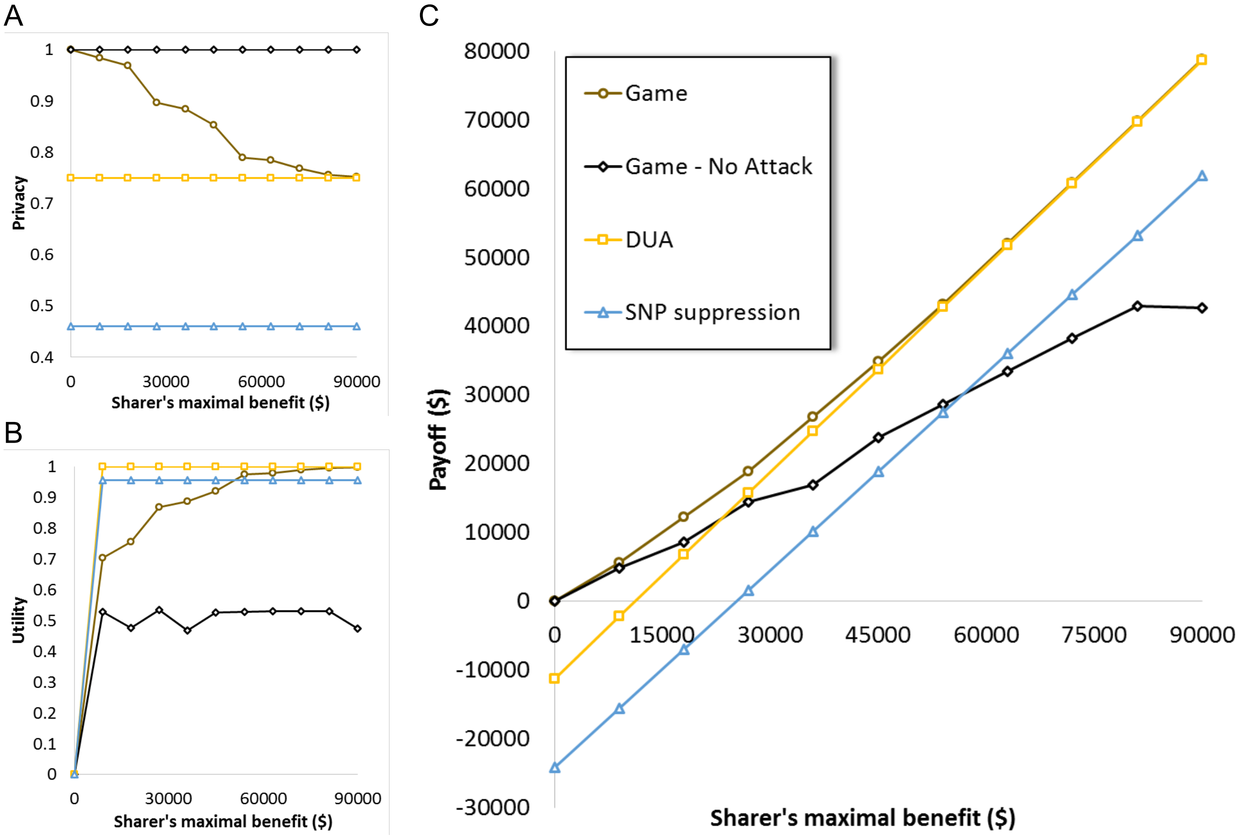


Figure A.8 Comparisons of four protection policies against the genomic inference attack targeting the SPHINX study pool, while varying the sharer's maximal benefit.

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (gold lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.8, it can be seen that the sharer's payoff increases with his or her maximal benefit. Still, as expected, the optimal game theoretic policy outperforms the other policies in terms of the sharer's expected payoff. Furthermore, the advantage of the optimal game theoretic policy over the DUA policy (or the SNP suppression policy) tends to increase as the sharer's maximal benefit decreases.

A.4.3 Gain to the Recipient per Successful Attack

Next, we investigated how the gain to the recipient per successful attack affects the sharer's optimal strategy and expected payoff. The default value of the gain to the recipient per successful attack G_R was

\$360. Figure A.9 and Figure A.10 depict the proportion of SNPs released and the payoffs, respectively, as the gain to the recipient varies from \$0 to \$4,860 (holding all other parameters to their default values).

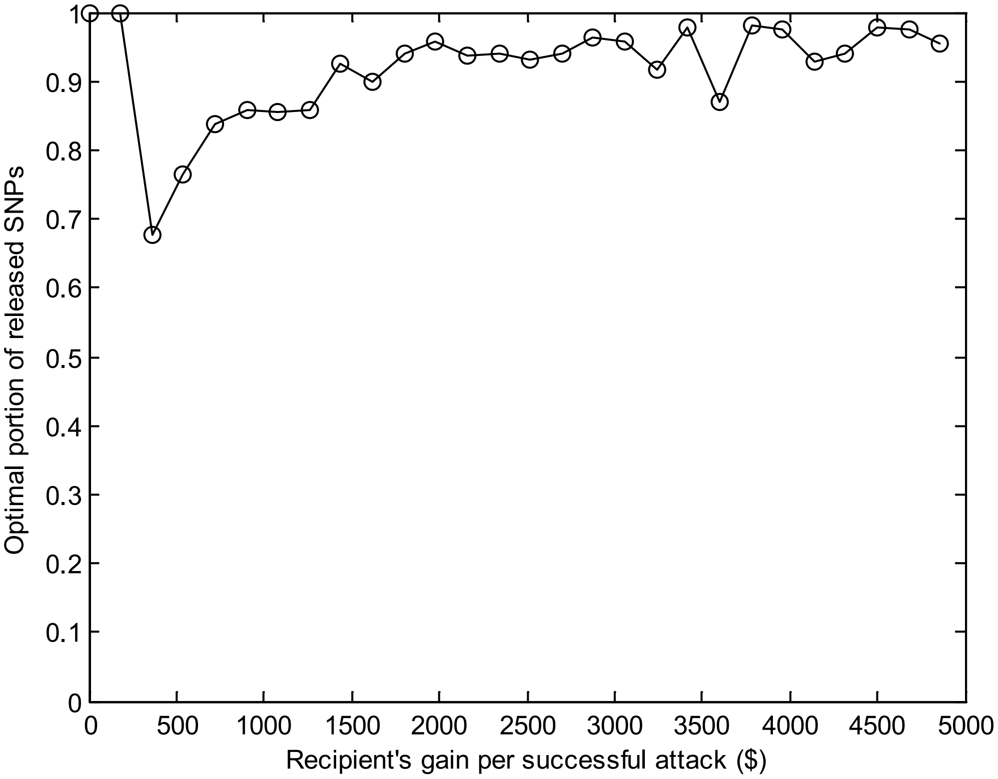


Figure A.9 Optimal portion of released SNPs as a function of the recipient's gain per successful attack in the relaxed game protecting the SPHINX dataset.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.9, it can be seen that the sharer's optimal strategy is sensitive to changes in the gain to the recipient per successful attack G_R. However, the sharer's strategy almost always releases more than 70% SNPs and releases all SNPs if the gain to the recipient per successful attack G_R is smaller than \$180.

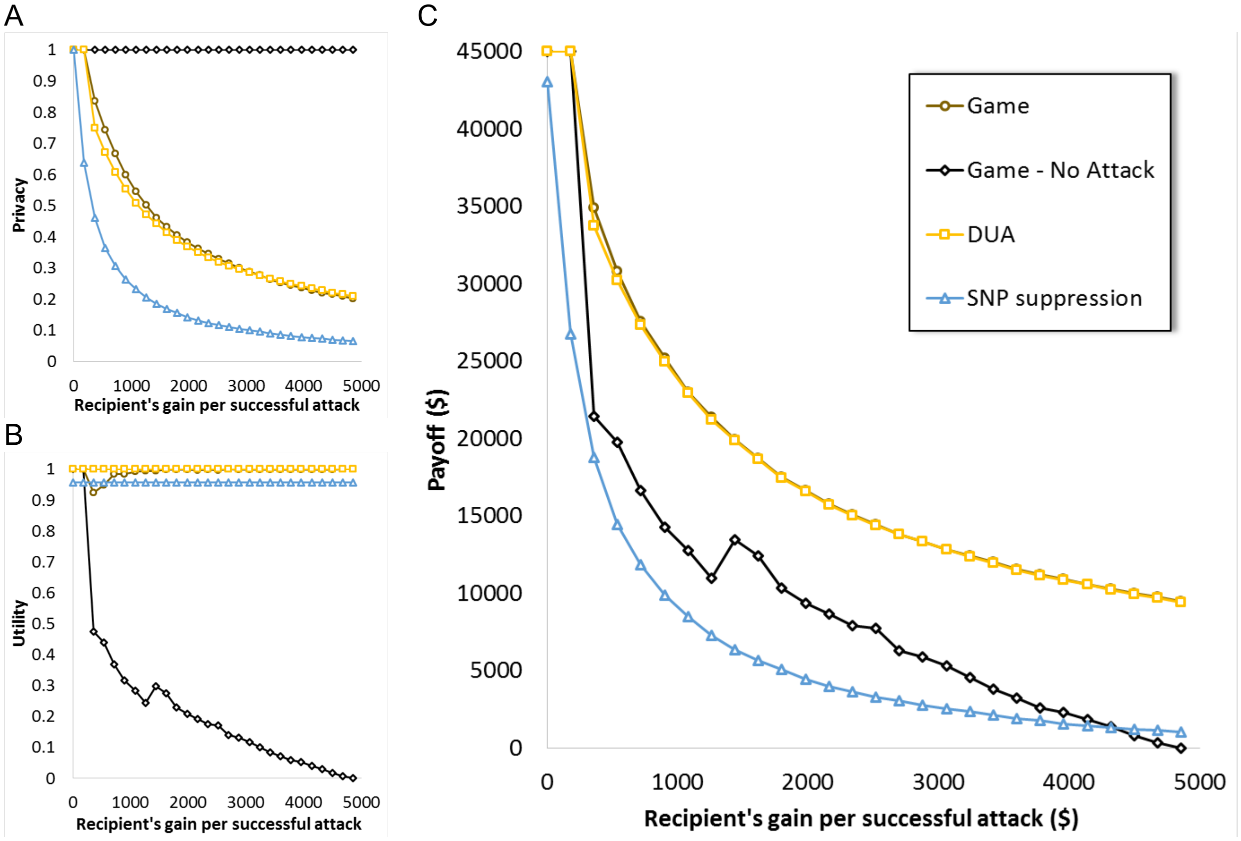


Figure A.10 Comparisons of four protection policies against the genomic inference attack targeting the SPHINX study pool, while varying the recipient's gain per successful attack.

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (gold lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.10, it can be seen that the sharer's payoff decreases as the gain to the recipient increases. As expected, the optimal game theoretic policy outperforms the other policies in terms of both the sharer's expected payoff and the privacy score. Furthermore, the optimal game theoretic policy retains 99% utility when the gain to the recipient per successful attack is smaller than $180 or larger than $1,080. In addition, the sharer's payoff under the optimal game theoretic policy can hardly be eliminated even if the gain to the recipient per successful attack keeps increasing, while the sharer's payoff under the no-risk game theoretic policy, that guarantees no attack will be successful, starts to be eliminated when the gain to the recipient reaches $4,860 per successful attack, which is unlikely true in our framework (without the consideration of economic factors like potential inflation).

A.4.4 Loss to the Sharer per Successful Attack

Next, we investigated how the loss to the sharer per successful attack affects the sharer's optimal strategy and expected payoff. The default value of the loss to the sharer per successful attack L_S is \$360. Figure A.11 and Figure A.12 depict the proportion of SNPs released and the payoffs, respectively, as the loss to sharer varies from \$0 to \$4,860 (holding all other parameters to their default values).

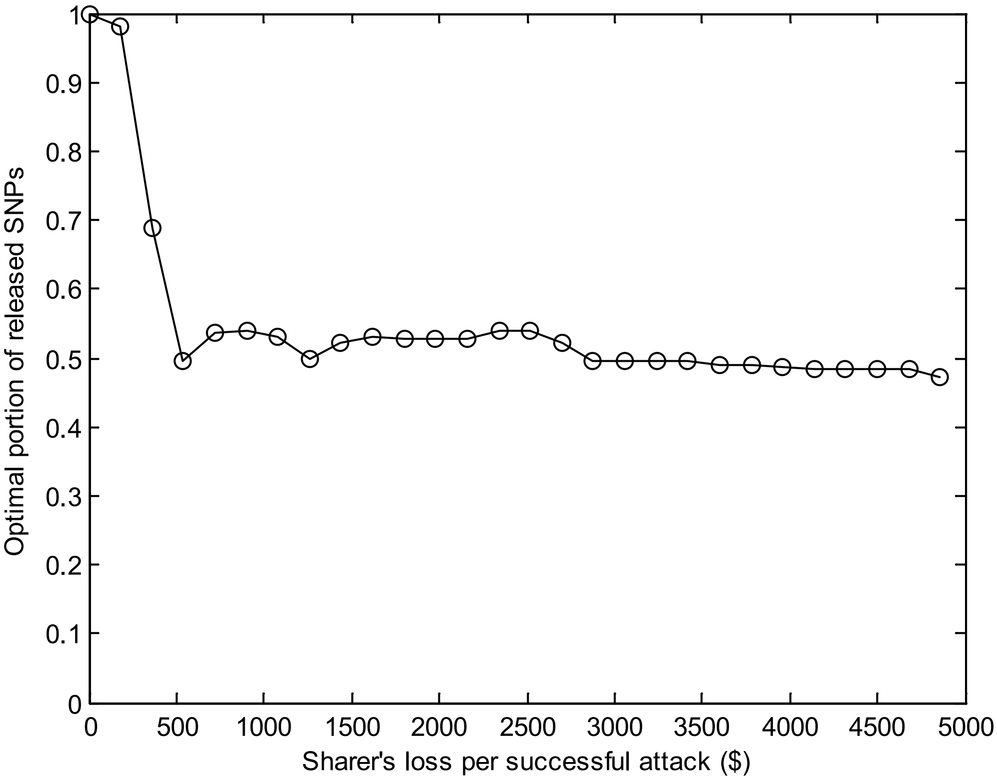


Figure A.11 Optimal portion of released SNPs as a function of the sharer's loss per successful attack in the relaxed game protecting the SPHINX dataset.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.11, it can be seen that the sharer's optimal strategy is sensitive to the change of his or her loss per successful attack L_S. The optimal portion of released SNPs tends to decrease as the sharer's loss increases, and converges to 22.5% after the loss to the sharer per successful attack reaches \$4,140.

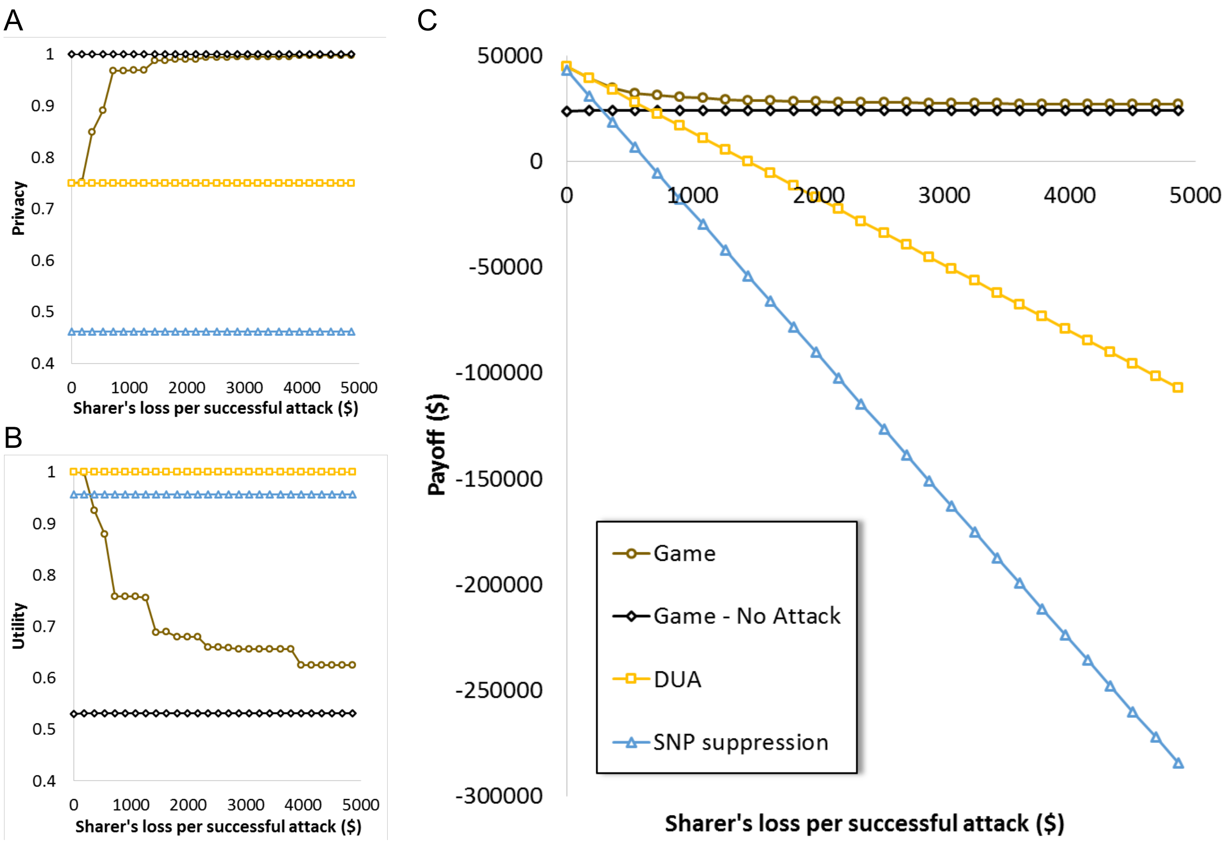


Figure A.12 Comparisons of four protection policies against the genomic inference attack targeting the SPHINX study pool, while varying the sharer's loss per successful attack.

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (gold lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.12, it can be seen that the sharer's payoff decreases as his or her loss increases. Still, as expected, the optimal game theoretic policy outperforms the other policies in terms of the sharer's expected payoff and the privacy score. Furthermore, the optimal game theoretic policy retains 99% privacy when the loss to the sharer is larger than $1,800 per successful attack. In addition, the sharer's payoffs under both game theoretic policies converge to $23,935, while the sharer's payoffs under the other two policies decrease linearly, as the loss to the sharer increases. Thus, the advantage of the optimal game theoretic policy over the DUA policy (or the SNP suppression policy) tends to increase as the loss to the sharer increases.

A.4.5 Gain to the Recipient (equals Loss to the Sharer) per Successful Attack

Next, we investigated how the gain to the recipient per successful attack affects the sharer's optimal strategy and expected payoff if the loss to the sharer is always equal to the gain to the recipient. The default value of the gain to the recipient (the loss to the sharer) per successful attack is $360. Figure A.13 and Figure A.14 depict the proportion of SNPs released and the payoffs, respectively, as the gain to the recipient (the loss to the sharer) varies from $0 to $4,860 (holding all other parameters to their default values).

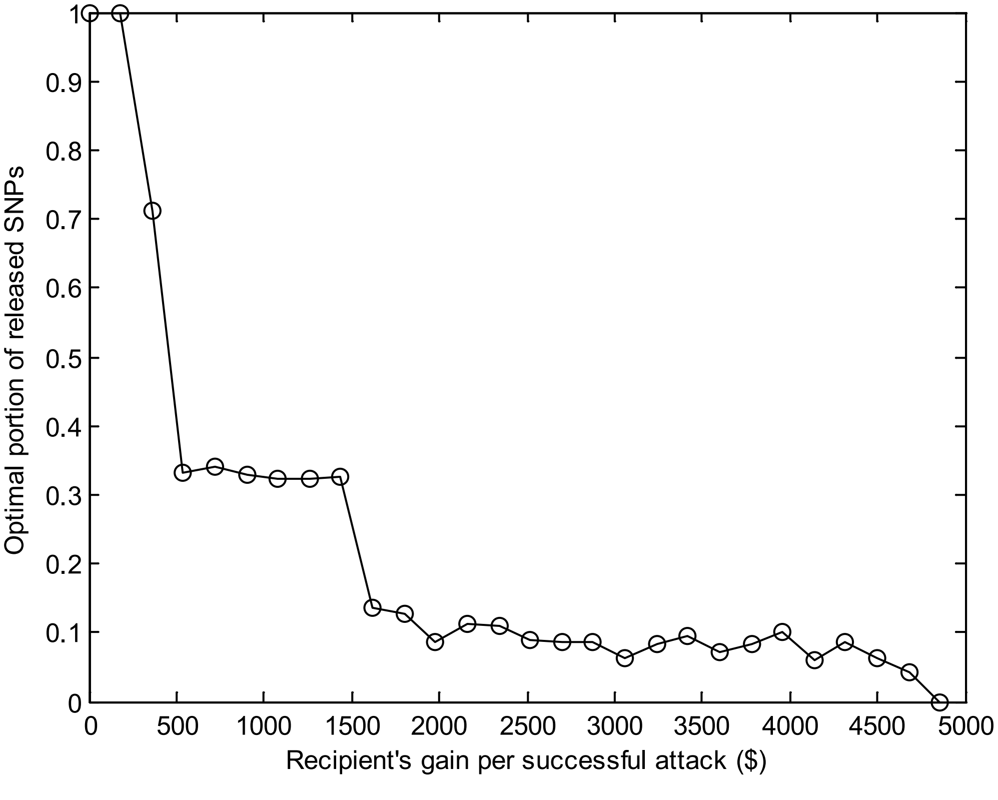


Figure A.13 Optimal portion of released SNPs as a function of the recipient's gain (equals the sharer's loss) per successful attack in the relaxed game protecting the SPHINX dataset.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.13, it can be seen that the sharer's optimal strategy is sensitive to the change of the gain to the recipient per successful attack. The optimal number of released SNPs tends to decrease as the sharer's loss increases, and converges to 0 after the loss to the sharer per successful attack reaches $4,860.

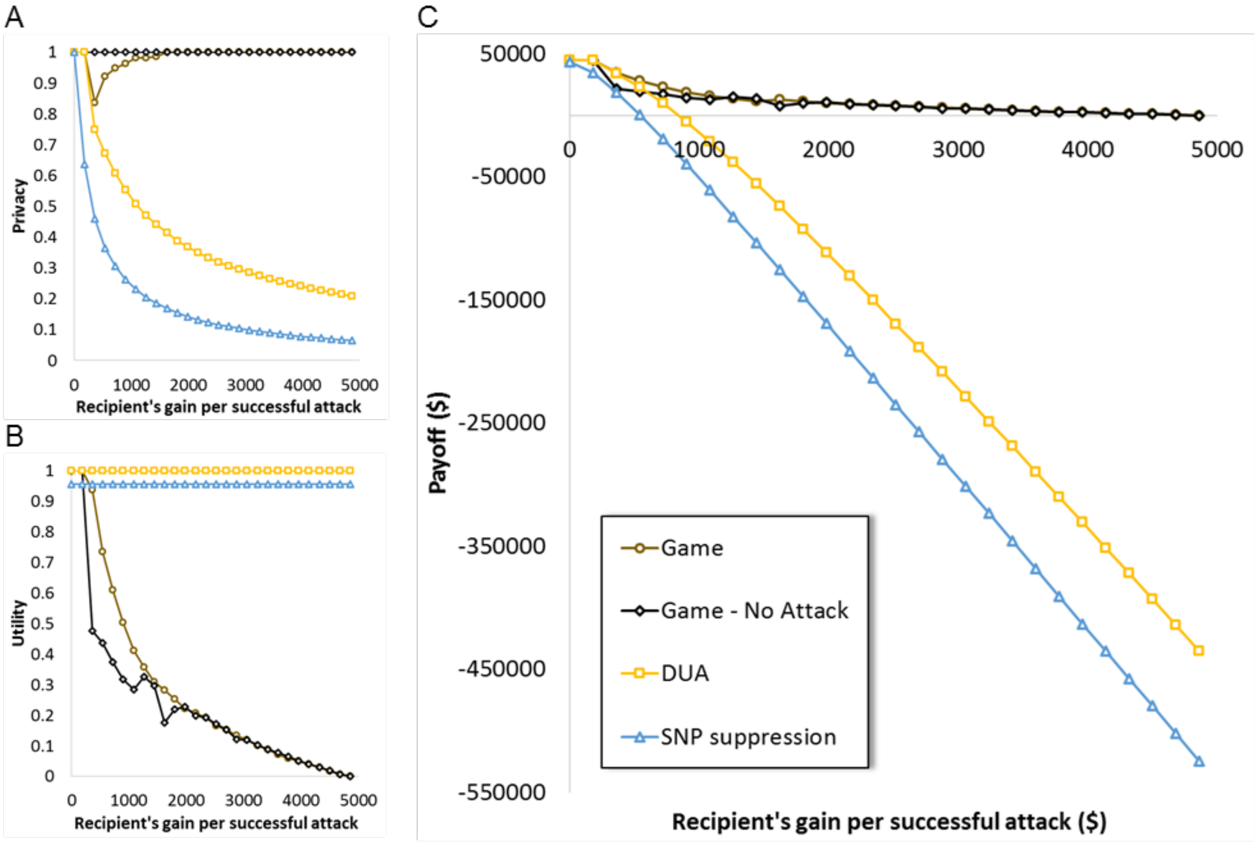


Figure A.14 Comparisons of four protection policies against the genomic inference attack targeting the SPHINX study pool, while varying the recipient's gain and the sharer's loss per successful attack (assuming the sharer's loss is always equal to the recipient's gain).

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (gold lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.14, it can be seen that the sharer's payoff decreases as his or her loss increases. Still, as expected, the optimal game theoretic policy outperforms the other policies in terms of the sharer's expected payoff and the privacy score. Furthermore, the optimal game theoretic policy retains all privacy when the loss to the sharer is smaller than \$180 or larger than \$1,620 per successful attack. In addition, the data utility and the sharer's payoffs under both game theoretic policies converge towards 0 (becomes 0 when the loss to the sharer per successful attack reaches \$4,860), while the sharer's payoffs decrease linearly under the other two policies, as the loss to the sharer increases. Thus, the advantage of the optimal game theoretic policy over the DUA policy (or the SNP suppression policy) tends to increase as the loss to the sharer increases.

A.4.6 Number of Targets

Next, we investigated how the number of targets affects the sharer's optimal strategy and expected payoff. The default value of the number of the targets n_x is 2,500. Figure A.15 and Figure A.16 depict the proportion of SNPs released and the payoffs, respectively, as the number of targets varies from 0 to 5,000.

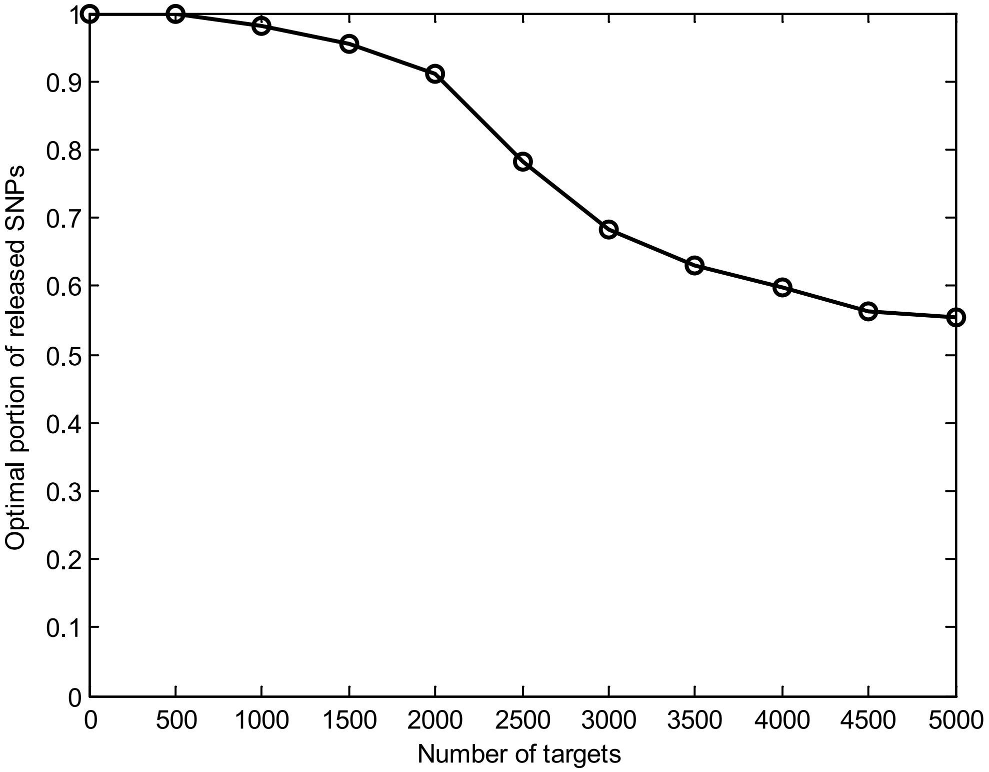


Figure A.15 Optimal number of released SNPs as a function of the number of targets in the relaxed game protecting the SPHINX dataset.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.15, it can be seen that the sharer's optimal strategy is not that sensitive to the number of targets n_x. The optimal number of released SNPs decreases as the number of targets increases, but at a slower pace than what was observed in the sensitivity analyses above.

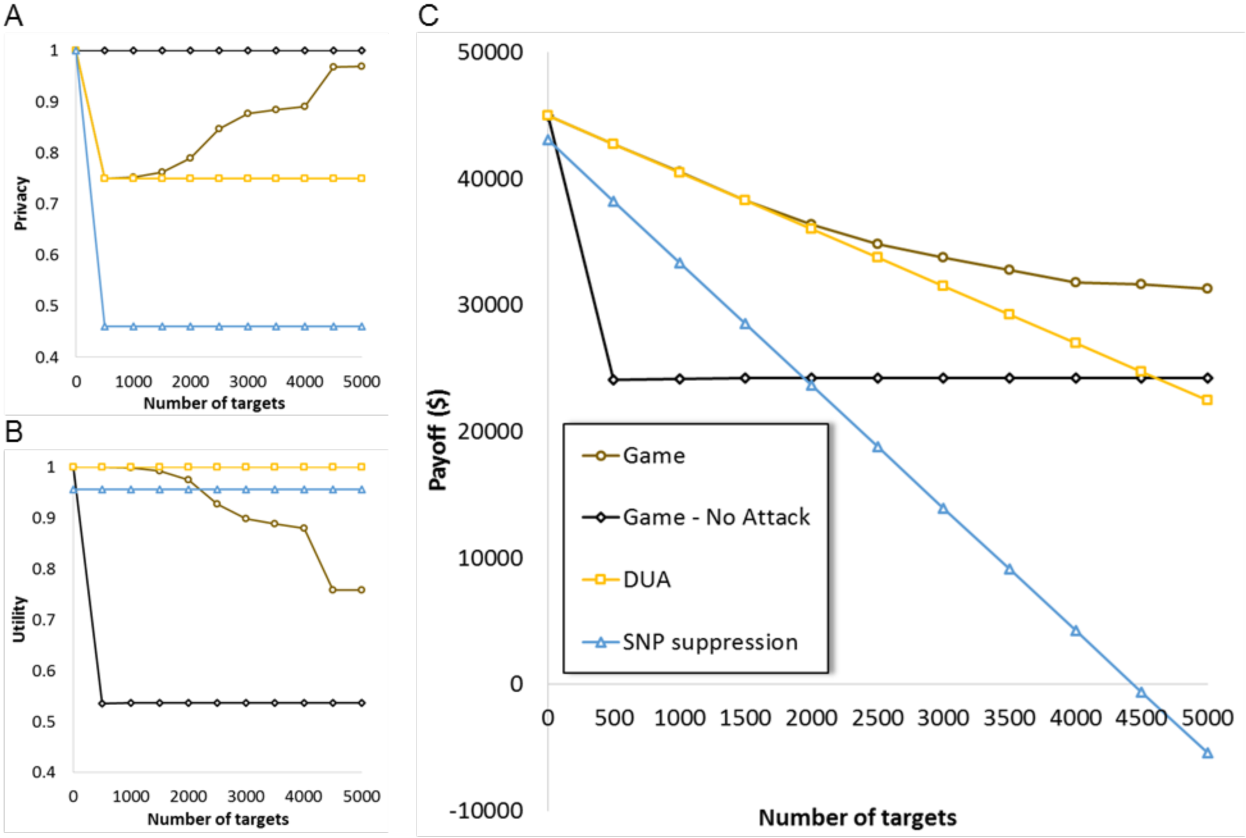


Figure A.16 Comparisons of four protection policies against the genomic inference attack targeting the SPHINX study pool, while varying the number of targets.

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (gold lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.16, it can be seen that the sharer's payoff decreases as the number of targets increases. As expected, the optimal game theoretic policy outperforms the other policies in terms of the sharer's expected payoff. Furthermore, the advantage of the optimal game theoretic policy over the DUA policy (or the SNP suppression policy) tends to increase along with the number of targets.

A.4.7 Prior Probability that Each Target is From the Pool

Next, we investigated how the prior probability, that each target is from the pool, affects the sharer's optimal strategy and expected payoff. The default value of the prior probability p is 0.05. Specifically, as described in Chapter 3, we set the priors to simulate four genome sequencing programs: (i) the Precision Medicine Initiative (PMI) [231], which will have a prior of 0.003 for 1 million participants out of 318 million US citizens, (ii) the Million Veteran Program (MVP) [232], which has a prior of 0.02 for 400,000 participants out of 21 million US military veterans, (iii) the BioVU de-identified DNA repository program of the Vanderbilt University Medical Center [233], which has a prior of 0.1 for 250,000 participants out of two million individuals with electronic medical records at the institution, and (iv) the Rare Diseases Clinical Research Network or RDCRN, which we assume has a prior of 0.5 for coverage of half of the possible population. For context, we compare these with the 0.05 prior probability relied upon in the experiments for SPHINX described above.

Figure A.17 and Figure A.18 depict the proportion of SNPs released and the payoffs, respectively, as the prior probability varies over the set {0.003, 0.002, 0.05, 0.1, 0.5} (holding all other parameters to their default values). Figure A.18 communicates the same result as Figure 3.11 in Chapter 3. We present this result for completeness and consistency in this section.

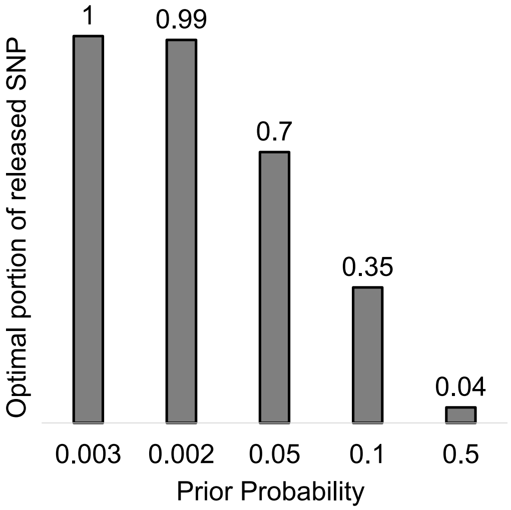


Figure A.17 Optimal portion of released SNPs as a function of the prior probability in the relaxed game protecting the SPHINX dataset.

SNP, single nucleotide polymorphism. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.17, it can be seen that the sharer's optimal strategy is sensitive to the prior probability p. The optimal number of released SNPs decreases as the prior probability increases. However, the sharer's strategy is releasing all SNPs if the prior probability is smaller than 0.002.

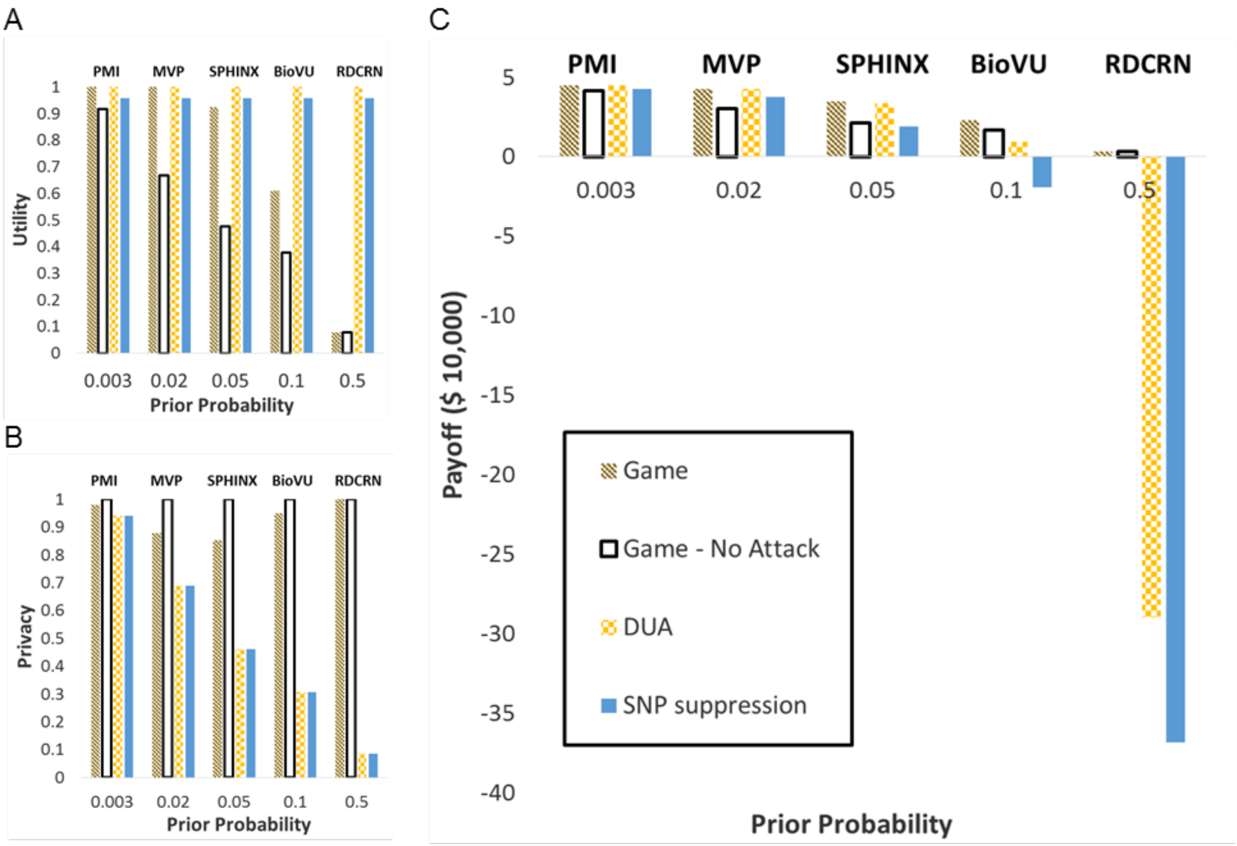


Figure A.18 Comparisons of four protection policies for a range of genomic data sharing programs with varying prior probabilities against the genomic inference attack targeting the SPHINX study pool.

The compared policies include: 1) the optimal game theoretic solution (brown bars filled with downward diagonal pattern), 2) the game theoretic solution that ensures no attack is successful (black bars with no fill), 3) the data use agreement (DUA) (gold bars filled with checkerboard pattern), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue bars with a solid fill), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. (PMI, Precision Medicine Initiative. MVP, Million Veteran Program. SPHINX, Sequence and Phenotype Integration Exchange. BioVU, de-identified biorepository of Vanderbilt University Medical Center. RDCRN, Rare Diseases Clinical Research Network.)

In Figure A.18, it can be seen that the sharer's payoff decreases as the prior probability increases. As expected, the optimal game theoretic policy outperforms the other policies in terms of the sharer's expected payoff. Furthermore, the advantage of the optimal game theoretic policy over the DUA policy (or the SNP suppression policy) tends to increase along with the prior probability.

A.4.8 Summary

In summary, each of the parameters affects the sharer's expected payoff in a monotonic manner. Empirically, the sensitivity analyses indicate that the sharer's expected payoff increases as c_a, c_p, or H increases, as G_R, L_S, n_x, or p decreases. In other words, the sharer's expected payoff increases as θ_l increases or ω decreases. Analytically, we can come to the same conclusion according to Equation A.47 and Equation A.52 in **Appendix A.1**. If θ_l increases, Equation A.52 implies that the recipient would tend not to attack, while Equation A.47 implies that the sharer's payoff would increase. If ω decreases, Equation A.47 implies that the sharer's payoff would increase. Thus, the sharer's payoff is positively correlated to the likelihood ratio threshold θ_l, and negatively correlated to the privacy-utility ratio ω.

A.5 SPHINX Robustness Analysis

In the case study and sensitivity analyses, the sharer knows the recipient's gain per successful attack (i.e., the worth of the data to the recipient) and the number of targets. However, in practice, the sharer is likely to be uncertain about these two parameters. Thus, in this section, we conduct robustness analyses on: 1) the recipient's gain per successful attack G_R, and 2) the number of targets n_x, to see how uncertainties in the sharer's knowledge of the recipient affect the sharer's expected payoff and the game policy.

A.5.1 Recipient's Gain per Successful Attack

First, we investigated how the bias in the sharer's knowledge on the recipient's gain per successful attack affects the sharer's optimal strategy and expected payoff if the loss to the sharer is always equal to the gain to the recipient. The default value of the gain to the recipient (the loss to the sharer) per successful attack G_R is \$360. Figure A.19 depicts the payoff as the actual gain to the recipient (the loss to the sharer) per successful attack varies from \$0 to \$2,520, while the sharer believes this parameter is set to \$360 (holding all other parameters to their default values).

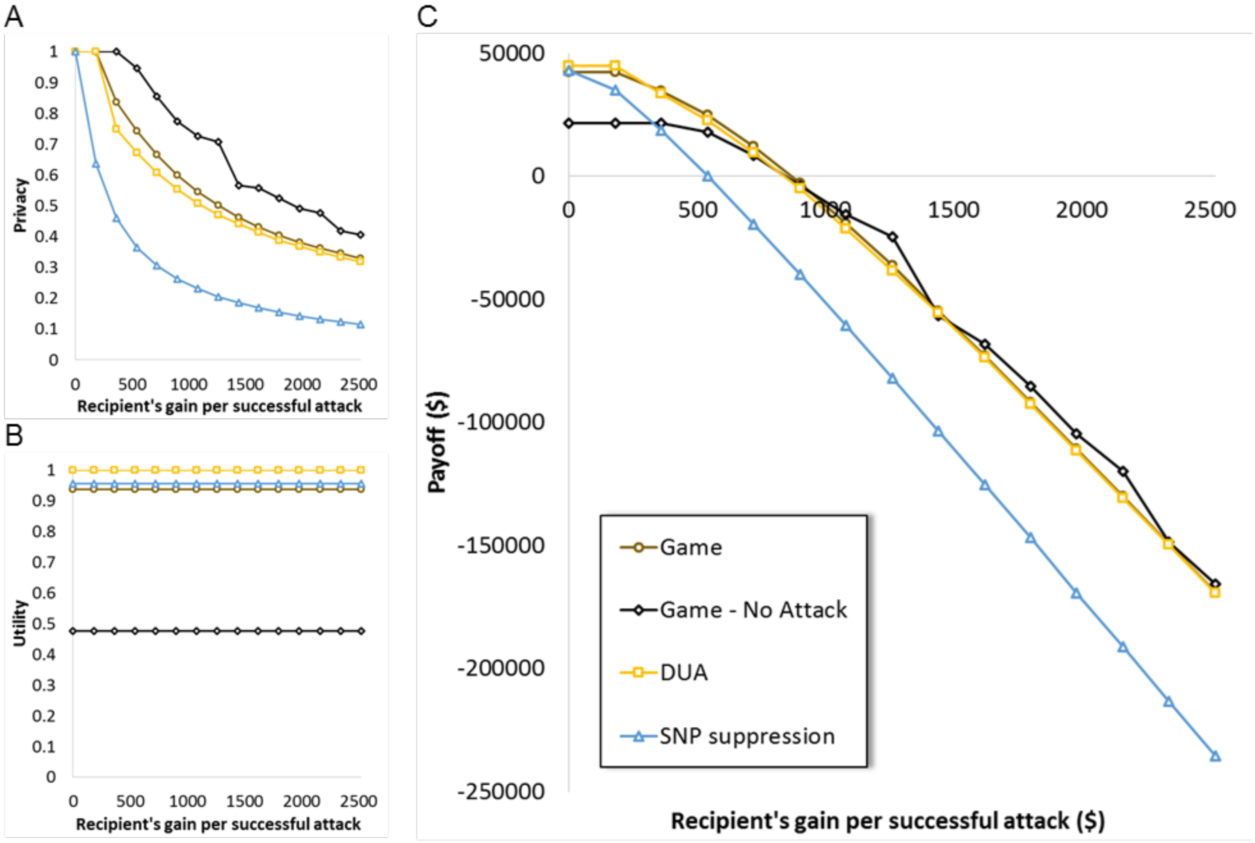


Figure A.19 Comparisons of four protection policies against the genomic inference attack targeting the SPHINX study pool in the robust analysis on the recipient's gain per successful attack (assuming the sharer's loss is always equal to the recipient's gain).

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (gold lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.19, we can see that the sharer's payoff decreases as the actual gain to the recipient increases. Although the optimal game theoretic policy is not always the best, it is better than the others in most cases in terms of the sharer's payoff, especially when the gain to the recipient per successful attack is larger than $180 and smaller than $900. Furthermore, the game theoretic policies always outperform the DUA policy and the SNP suppression policy in terms of the privacy score.

A.5.2 Number of Targets

Next, we investigated how the bias in the sharer's knowledge about the number of targets affects the sharer's optimal strategy and expected payoff. The default value of the number of the targets n_x is 2,500. Figure A.20 depicts the payoff as the actual number of targets varies from 0 to 5,000, but the sharer believes this parameter is set to 2,500 (holding all other parameters to their default values).

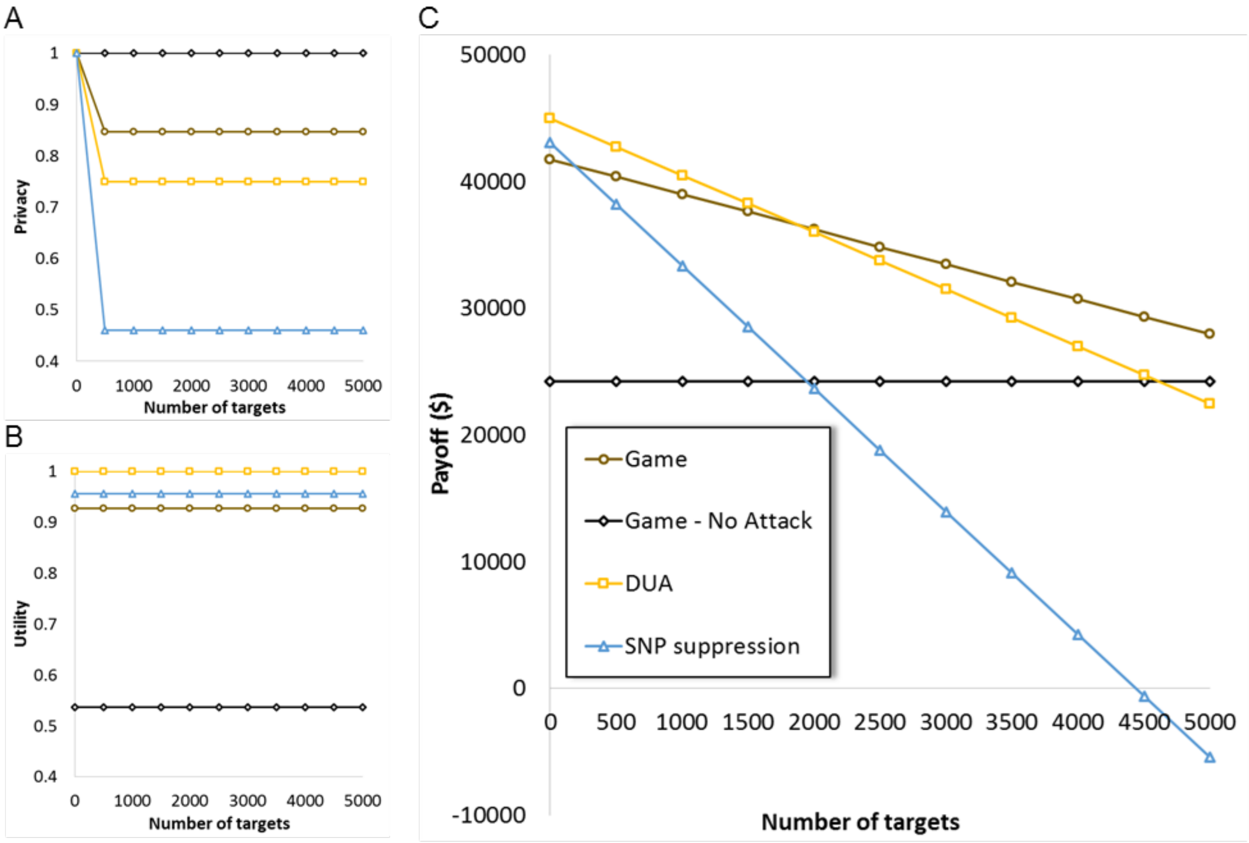


Figure A.20 Comparisons of four protection policies against the genomic inference attack targeting the SPHINX study pool in the robust analysis on the number of targets.

The compared policies include: 1) the optimal game theoretic solution (brown lines), 2) the game theoretic solution that ensures no attack is successful (black lines), 3) the data use agreement (DUA) (gold lines), and 4) the single nucleotide polymorphism (SNP) suppression solution (blue lines), with respect to (A) the expected privacy score of the shared data, (B) the utility score of the shared data, and (C) the sharer's expected overall payoff,. The overall payoff results from combining (i) the privacy protection afforded to the targeted individuals and (ii) the utility in the set of SNPs that are shared. SPHINX, Sequence and Phenotype Integration Exchange.

In Figure A.20, it can be seen that the sharer’s payoff decreases as the actual number of targets increases. Although the game theoretic policy is not always the best, it is better than others in most cases, especially when the actual number of targets are larger than 2,000. The payoffs for the game theoretic policy assures no attack is minimally influenced. Furthermore, the optimal game theoretic policy and the no attack variant are much better than the DUA policy and the SNP suppression policy in terms of the privacy score.

A.5.3 Summary

These results indicate that the sharer’s expected payoff is quite robust with respect to both the gain to the recipient and the number of targets. The payoffs are not completely invariant, but uncertainty in such knowledge often leads to the same overall results of the game.

Appendix B

APPENDIX FOR CHAPTER 5

B.1 Data Sanitization Process for Craig Venter's Data and the Ysearch Dataset

To protect the privacy of the corresponding subjects and enable replications of our investigation, we sanitized the original Venter and Ysearch datasets (i.e., modified for privacy protection) before conducting the experiment, but without affecting the data utility for the demonstration purpose. Specifically, each non-missing value was substituted by a unique random number (within an attribute-dependent range) for all records regarding each genomic attribute. In addition, we masked each genomic attribute's name and replaced each surname in the dataset (except the surname Venter) with a uniquely random artificial surname (again, except the surname Venter).

Now we explain how the value substitution worked in more detail. For example, let us say that the value of Y-STR α is within the set of {2, 4, 5, 6} for all records in the experiments, then a random substitution mapping for Y-STR α could be {2->3, 4->2, 5->5, 6->4} for all records. In addition, let us say that the value of Y-STR β is within the set of {7, 8, 9, 10} for all records, then a random substitution mapping for Y-STR β could be {7->3, 8->6, 9->4, 10->5} for all records. Note that, we did not allow two original values to be substituted by the same value.

The name replacements worked in a similar way. For instance, let us say that the set of surnames in the dataset is {Smith, Johnson, Williams, Venter, Leonard}, then a random replacement mapping could be { Smith -> Fisher, Johnson -> Barnes, Williams -> Butler, Venter -> Venter, Leonard -> Swanson}. Still, we did not allow two original surnames to be substituted by the same artificial surname.

In these ways, in addition to not releasing all attributes' names, we made it difficult for a data recipient to recover the original data and kept all the data utility to demonstrate the attack. All the effectiveness measures of the experiment conducted on the original dataset are the same as those of the experiment conducted on the modified dataset.

B.2 Data Preparation Process in the Experiments based on a Large-scale Simulated Population

We simulated a genetic genealogical population using an individual-based forward-time population genetics simulation tool, simuPOP 1.1.8.3 (v2017) [303, 304] (simupop.sourceforge.net), from which we derived three datasets. The data preparation pipeline for datasets used in the experiments is shown in Figure B.1. The inputs of the population simulation engine include probability distributions and statistics of corresponding attributes. The distributions of surnames [305] and demographics such as states [306] are based upon the 2010 US census data. From the census data, the top 1,000 surnames were selected to generate the population. 16 Y-STR markers were selected according to a study on Y-STR mutations [307]. The Y-STR statistics, including the distribution of tandem repeats and mutation rates, were queried in 2018 from the Y-Chromosome Haplotype Reference Database (YHRD)[29] and summarized in Table B.1.

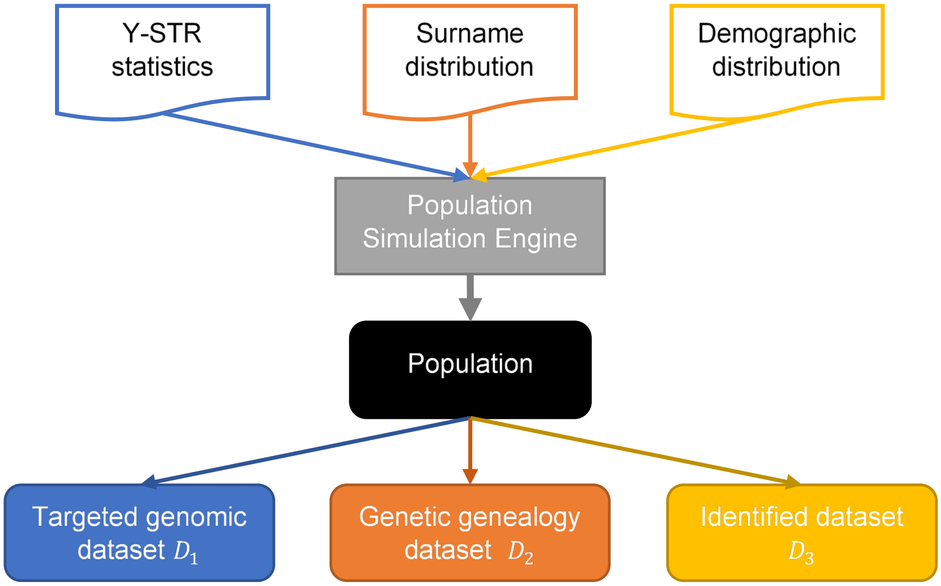


Figure B.1 Preparation pipeline for datasets used in the simulation-based experiments.

With inputs of probability distributions and statistics of corresponding attributes, the population simulation engine generates simulated population, from which three datasets are derived. Y-STR stands for Y-chromosome short tandem repeat.

[29] Y-Chromosome Haplotype Reference Database (YHRD). https://yhrd.org.

Table B.1 Statistics for 16 Y-STR genomic attributes used in the population simulation.

Y-STR	Range of repeats	Mutation rate (*.001) (2018)	Mutation No. (2018)	Mutation rate (*.001) (2008)	Mutation No. (2008)
DYS19	[6, 20]	2.202	37 (in 16,801)	2.381	23 (in 9,658)
DYS385	[6, 28]	2.508	70 (in 27,911)	2.081	31 (in 14,896)
DYS389I	[9, 17]	2.724	41 (in 15050)	1.781	14 (in 7,862)
DYS389II	[23, 36]	4.327	65 (in 15,021)	2.803	22 (in 7,849)
DYS390	[16, 30]	2.083	34 (in 16323)	2.298	21 (in 9,140)
DYS391	[5, 16]	2.531	41 (in 16197)	3.081	28 (in 9,089)
DYS392	[5, 20]	0.496	8 (in 16129)	0.552	5 (in 9,053)
DYS393	[7, 18]	1.068	16 (in 14975)	0.893	7 (in 7,842)
DYS437	[9, 22]	1.320	15 (in 11363)	1.498	7 (in 4,672)
DYS438	[5, 19]	0.351	4 (in 11384)	0.425	2 (in 4,709)
DYS439	[5, 19]	5.459	62 (in 11358)	5.762	27 (in 4,686)
DYS448	[13, 25]	1.385	11 (in 7940)	1.590	2 (in 1,258)
DYS456	[9, 24]	4.408	35 (in 7940)	4.769	6 (in 1,258)
DYS458	[10, 24]	6.172	49 (in 7939)	6.359	8 (in 1,258)
DYS635	[12, 30]	4.211	37 (in 8787)	3.754	8 (in 2,131)
GATA H4.1	[6, 20]	3.010	27 (in 8971)	2.180	5 (in 2,294)

Y-STR stands for Y-chromosome short tandem repeat. DYS stands for DNA Y-chromosome segment.

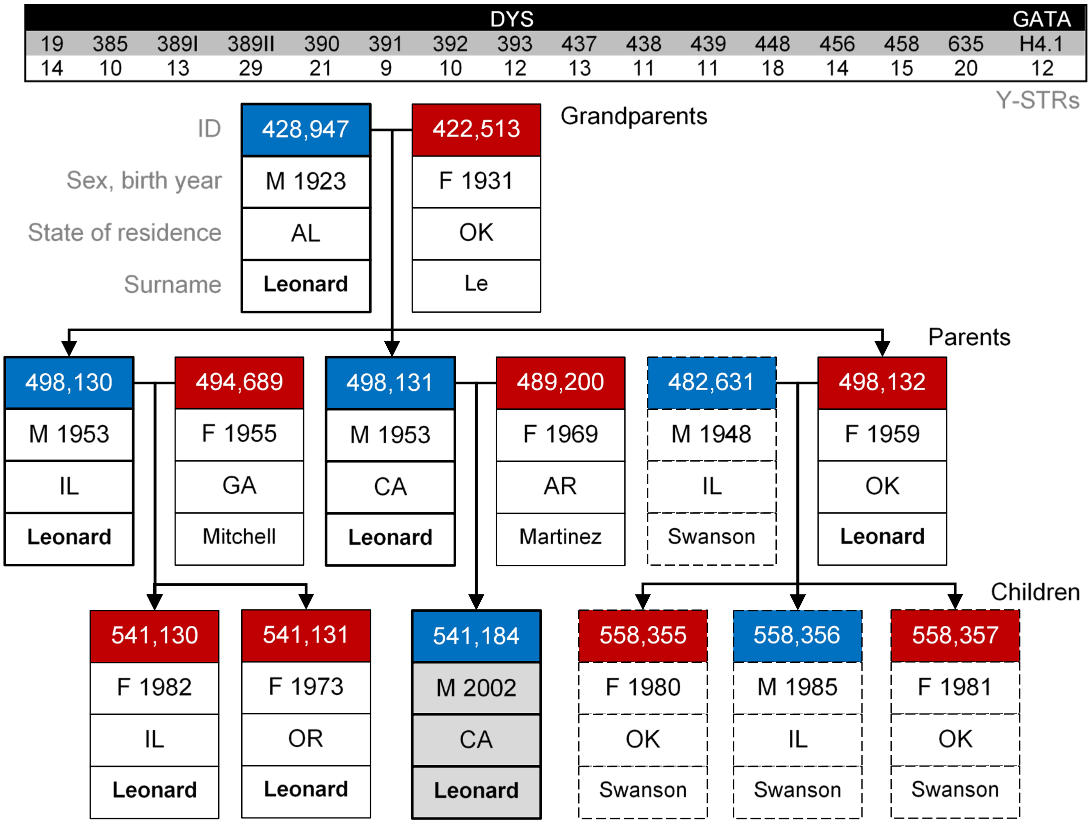
The process of population generation using simuPOP was as follows. The population data were initialized with the following attributes: ID, father's ID, mother's ID, sex, subpopulation, birth year, state of residence, surname, income level, socioeconomic level, and 16 Y-STRs. IDs are integers starting from 0. We set the Y-STRs, state of residence, surname, and income level for everyone in the first generation according to corresponding distributions, respectively. The distribution of income level (every $2500 per level) is based upon a survey [308] conducted in 2015 by the US Census Bureau. The level of socioeconomic status (i.e., upper-class, middle-class, or lower-class) was set according to an individual's income level. The middle-class range was set to be [$40,500, $99,999] according to numbers [309] from the Pew Research Center and the US Census Bureau. Everyone's birth year was set uniformly at random from the 30-year range of [1620, 1649]. Each sex group has the same size, and each subpopulation has the same size as well.

Afterward, we set the mating scheme in the simulation engine that chooses parents from a prenatal generation and generates offspring from chosen parents. Parents were chosen from their respective sex groups. Instead of using a random chooser or any pre-defined chooser, we customized the chooser according to the following procedure. First, we select an individual from the pool of candidates and record the socioeconomic level, and then we select a pair of parents from this socioeconomic level uniformly at random. For those parents who do not meet a particular set of conditions, a new pair of parents will be re-selected if a random number generated from the range of (0,1) is larger than 0.2. The selection loop ends if those conditions are met, or if the random number generated in the current loop is smaller than 0.2. The set of conditions include that the father is not two years younger than the mother, and that the father is not 12 years older than the mother, and that the father's income level is not lower than the mother's. A selected individual has a probability of 80% to be moved out of the candidate pool unless everyone else in the candidate pool has either different sex or different socioeconomic level from the selected individual.

For each pair of parents, the number of children was determined by a zero-truncated Poisson distribution with the parameter $\lambda = 1.6$, which has an expected value of $\lambda/\left(1 - e^{-\lambda}\right) = 2.0$ and a standard deviation of 1.4, according to the average number of own children in families in the US in 1977. Children's attributes were set as follows: The sex of each child was assigned randomly with equal probabilities of male and female. The genotypes of each child were transmitted from parents following Mendel's laws. And the surname of each child was inherited from the father. To simulate each mutation event, we used a stepwise mutation model [310] to increase or decrease the number of repeats for each locus of a new genomic sequence.

To simulate each birth event, we set the child's birth year as the sum of the mother's birth year and the mother's age during the birth event, the latter of which was generated based on a distribution of ages that women become mothers. According to a report from the US National Center for Health Statistics [311], the number of births for women ages 15-19, 20-24, 25-29, 30-34 are 0.18 (5%), 0.73 (19%), 1.10 (29%), 1.09 (29%) millions in 2018, respectively. However, if the father's resulting age is smaller than 16, the child's birth year would be set as the birth year of the father plus 16. To simulate the change of socioeconomic level across generations, we assumed that the expected income level of an individual (in middle age) is equal to the rounded average income level of his or her parents plus a number selected from the set of [-1, 1] uniformly at random unless the resulting income level of the individual is out of bounds. To simulate the change of state of residence, we assumed a child's state of residence has a probability of 33.33% to be different from their parents' state of residence. The new state of residence was selected according to the corresponding distribution of state. Otherwise, the child would have a probability of 50% sharing the same state of residence with their father and the same probability sharing the same state of residence with their mother.

The simulation for the mating process for a prenatal generation in a subpopulation stopped when enough people are generated for the next generation. We assumed that every subpopulation has the same number of people in each generation before migration events. To simulate migration events among subpopulations, we used a migration-by-proportion model to migrate a fixed 10% proportion of each subpopulation to each other populations every five generations. The entire population generation process stopped when enough people are generated for the last generation. The family tree of a randomly selected family (the Leonard family) across the last three generations, with the corresponding primary attributes, is shown in Figure B.2. Note that the surname before marriage (i.e., maiden name) is shown for each female member in the family.



DYS															GATA
19	385	389I	389II	390	391	392	393	437	438	439	448	456	458	635	H4.1
14	10	13	29	21	9	10	12	13	11	11	18	14	15	20	12

Figure B.2 The family tree of a simulated family (the Leonard family) across the last three generations with 20 primary attributes in the display.

The 20 primary attributes include ID, sex, birth year, state of residence, surname, and 16 Y-STRs. The Y-STRs of the data subject with ID 541,184 are shown at the top. Maiden names instead of surnames are shown for female members. For example, the maiden name of the person with ID 422,513 (i.e., grandmother of the data subject with ID 541,184) is Le instead of Leonard (which is not a typo). Y-STR stands for Y-chromosome short tandem repeat. DYS stands for DNA Y-chromosome segment.

The generated population has a total of 600,000 records with 26 attributes and the family structures (i.e., pedigrees), from which we outputted a male population of 90,064 records (last three generations) with 20 primary attributes, including ID, surname, two demographic attributes (i.e., birth year and state of residence) and 16 genomic attributes (i.e., 16 Y-STRs).

In each experiment, we randomly selected three datasets from the population. We first selected 5,000 records uniformly at random and kept a set of attributes (namely, ID, first name (randomly generated), surname, birth year, and state of residence) as the identified dataset D_3. Then we selected 1,000 records uniformly at random from these 5,000 records in D_3 and kept another set of attributes (namely, ID, surname, birth year, state of residence, and Y-STRs) as the targeted genomic dataset D_1. Note that the surname attribute in this simulated dataset will not be released because it rarely exists in a real-world public dataset. Finally, we selected 10,000 records uniformly at random from the population, excluding these 5,000 records in D_3, and kept a set of attributes (namely, ID, surname, Y-STRs) as the genetic genealogy dataset D_2. We noticed that a real world genetic genealogy dataset D_2 (e.g., Ysearch) contains many missing values, so we let a specific portion (e.g., 30%) of genomic values in this dataset be missing values.