When Others Impinge upon Your Privacy: Interdependent Risks and Protection in a Connected World

THÈSE N^{O} 6515 (2015)

PRÉSENTÉE LE 13 MARS 2015

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE POUR LES COMMUNICATIONS INFORMATIQUES ET LEURS APPLICATIONS 1

PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Mathias HUMBERT

acceptée sur proposition du jury:

Prof. E. Telatar, président du jury
Prof. J.-P. Hubaux, directeur de thèse
Prof. J. Fellay, rapporteur
Prof. A. Juels, rapporteur
Prof. R. Molva, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

We recognize the integral and interdependent nature of the Earth, our home.
Preamble to the Rio Declaration on Environment and Development, 1992

To Suzette and Aude

Abstract

Privacy is defined as the right to control, edit, manage, and delete information about oneself and decide when, how, and to what extent this information is communicated to others. Therefore, every person should ideally be empowered to manage and protect his own data, individually and independently of others. This assumption, however, barely holds in practice, because people are by nature biologically and socially interconnected. An individual's identity is essentially determined at the biological and social levels. First, a person is biologically determined by his DNA, his genes, that fully encode his physical characteristics. Second, human beings are social animals, with a strong need to create ties and interact with their peers. Interdependence is present at both levels. At the biological level, interdependence stems from genetic inheritance. At the social level, interdependence emerges from social ties. In this thesis, we investigate whether, in today's highly connected world, individual privacy is in fact achievable, or if it is almost impossible due to the inherent interdependence between people.

First, we study interdependent privacy risks at the social level, focusing on online social networks (OSNs), the digital counterpart of our social lives. We show that, even if an OSN user carefully tunes his privacy settings in order to not be present in any search directory, it is possible for an adversary to find him by using publicly visible attributes of other OSN users. We demonstrate that, in OSNs where privacy settings are not aligned between users and where some users reveal a (even limited) set of attributes, it is almost impossible for a specific user to hide in the crowd. Our navigation attack complements existing work on inference attacks in OSNs by showing how we can efficiently find targeted profiles in OSNs, which is a necessary precondition for any targeted attack. Our attack also demonstrates the threat on OSN-membership privacy.

Second, we investigate upcoming interdependent privacy risks at the biological level. More precisely, due to the recent drop in costs of genome sequencing, an increasing number of people are having their genomes sequenced and share them online and/or with third parties for various purposes. However, familial genetic dependencies induce indirect genomic privacy risks for the relatives of the individuals who share their genomes. We propose a probabilistic framework that relies upon graphical models and Bayesian inference in order to formally quantify genomic privacy risks. Then, we study the interplay between rational family members with potentially conflicting interests regarding the storage security and disclosure of their genomic data. We consider both purely selfish and altruistic behaviors, and we make use of multi-agent influence diagrams to efficiently derive equilibria in the general case where more than two relatives interact with each other. We also propose an obfuscation mechanism in order to reconcile utility with privacy in

genomics, in the context where all family members are cooperative and care about each other's privacy.

Third, we study privacy-enhancing systems, such as anonymity networks, where users do not damage other users' privacy but are actually needed in order to protect privacy. In this context, we show how incentives based on virtual currency can be used and their amount optimized in order to foster cooperation between users and eventually improve everyone's privacy. We derive our analytical findings by relying upon Markov chains, game theory, and Markov decision processes. This last part demonstrates that other people can also play a beneficial role in privacy.

We conclude that the quest for online privacy is chimerical because of the lack of individual control over data. As a consequence, unless cooperation between people quickly expands, we should consider that online privacy is steadily vanishing, and start designing novel mechanisms for the upcoming post-privacy era. We should finally redefine privacy, which is, beyond an individual right, now part of the commons.

Keywords : interdependent privacy, genomic privacy, online social networks (OSNs), incentives, cooperation, Bayesian inference, graphical models, obfuscation mechanism, game theory, Markov chains, Markov decision processes, multi-agent influence diagrams

Résumé

La protection de la vie privée est définie comme le droit de contrôler, d'éditer, de gérer, et d'effacer l'information nous concernant, ainsi que de décider quand, comment, et dans quelle mesure cette information peut être communiquée à des tiers. Par conséquent, chaque individu devrait idéalement avoir les moyens de gérer et de protéger ses propres données, individuellement et indépendamment des autres. Cependant, cette hypothèse n'est que peu valable en pratique, car nous sommes par nature interconnectés biologiquement et socialement. Or, notre identité est essentiellement déterminée par nos sphères biologique et social. Premièrement, un individu est déterminé par son ADN, ses gènes, qui codent entièrement ses caractéristiques physiques. Deuxièmement, l'homme est un animal social, avec un besoin profond de créer des liens et d'interagir avec ses semblables. Nous sommes interdépendants à ces deux niveaux de notre identité. Au niveau biologique, l'interdépendance est le résultat de notre héritage génétique. Au niveau social, l'interdépendance est liée à nos liens sociaux. Dans cette thèse, nous étudions si, dans notre monde hyperconnecté, la protection de la vie privée est possible individuellement, ou si ceci est rendu quasi impossible par l'interdépendance inhérente à notre humanité.

Tout d'abord, nous étudions les risques de confidentialité liés à notre interdépendance au niveau social, en se focalisant sur les réseaux sociaux en ligne (comme Facebook), qui sont la projection numérique de notre vie sociale. Nous montrons que, même si un utilisateur définit avec soin ses paramètres de confidentialité afin de ne pas être présent dans l'annuaire de recherche, il est possible pour un attaquant de le retrouver en utilisant les attributs des autres utilisateurs accessibles publiquement. Nous démontrons que, dans les réseaux sociaux où les réglages de confidentialités ne sont pas similaires entre les utilisateurs et où certains utilisateurs révèlent un ensemble d'attributs (même restreint), il est pratiquement impossible pour un utilisateur spécifique de se cacher dans la masse des utilisateurs. Notre attaque de navigation complémente les travaux existants sur les attaques d'inférence dans les réseaux sociaux, en montrant comment l'on peut trouver efficacement des profils cibles dans les réseaux sociaux, ce qui est une condition nécessaire à n'importe quelle attaque ciblée. Notre attaque démontre également la menace qui pèse sur la confidentialité de l'adhésion à un réseau social.

Deuxièmement, nous examinons les risques de confidentialité liés à notre interdépendance au niveau biologique. En particulier, grâce à la baisse rapide des coûts de séquençage du génome, un nombre croissant d'individus font séquencer leur génome et le partagent en ligne et/ou avec des tiers. Cependant, les dépendances génétiques familiales entraînent des risques indirects pour la confidentialité des données génomiques des

membres d'une famille dont certains individus partagent leurs propres génomes. Nous proposons un modèle probabiliste qui s'appuie sur les modèles graphiques et l'inférence bayésienne pour quantifier formellement les risques liés aux données génomiques. Ensuite, nous étudions l'interaction entre des agents rationnels appartenant à une même famille, avec des intérêts potentiellement contradictoires concernant la sécurité et le partage des données génomiques. Nous considérons à la fois des comportements égoïste et altruiste, et utilisons des diagrammes d'influence multi-agents afin de calculer efficacement des équilibres dans le cas général où plus de deux membres d'une même famille intéragissent entre eux. Nous proposons également un mécanisme de brouillage afin de réconcilier l'utilité avec la protection des données génomiques, dans un contexte où tous les membres de la famille sont coopératifs et se soucient de la confidentialité des données génomiques des autres membres.

Troisièmement, nous étudions des systèmes de protection de la vie privée, comme les réseaux anonymes, où les autres utilisateurs ne nuisent pas notre vie privée mais sont au contraire nécessaires à la protection de cette vie privée. Dans ce contexte, nous montrons comment des incitations basées sur une monnaie virtuelle peuvent être utilisées et leur quantité optimisée afin d'encourager la coopération entre les utilisateurs et en fin de compte améliorer l'anonymat de tous les utilisateurs. Nous obtenons nos résultats analytiques en nous appuyant sur des chaînes de Markov, de la théorie des jeux, ainsi que sur des processus de décision markoviens. Cette dernière partie démontre que les autres individus peuvent aussi jouer un rôle positif dans la protection de la vie privée.

En conclusion, nous estimons que la quête de la confidentialité en ligne est une chimère à cause du manque de contrôle individuel sur les données personnelles. Par conséquent, à moins que la coopération entre les individus ne se développe rapidement, nous devrions considérer que la confidentialité en ligne est en train de disparaître, et devrions commencer à concevoir de nouveaux mécanismes pour l'ère post-confidentialité à venir. Nous devrions finalement redéfinir (le droit à) la vie privée, qui est désormais, au-delà d'un droit individuel, partie intégrante de nos biens communs.

Mots-Clés : interdépendance dans la protection des données, protection des données génomiques, réseaux sociaux en ligne, incitations, coopération, inférence bayésienne, modèles graphiques, mécanisme de brouillage, théorie des jeux, chaînes de Markov, processus de décision markoviens, diagrammes d'influence multi-agents

Acknowledgements

First and foremost, I am extremely grateful to my advisor, Prof. Jean-Pierre Hubaux, for giving me the opportunity to work on fascinating research topics, and for providing me with the freedom that I needed to satisfy my thirst for knowledge. Thanks also for always encouraging me to work on my weaknesses, and for providing me with precious advice in the journey towards this Ph.D. thesis.

I would like to thank my thesis committee members Prof. Jacques Fellay, Prof. Ari Juels, Prof. Refik Molva, and Prof. Emre Telatar for their time and effort spent reviewing this dissertation. It was an immense honor to have these brilliant minds sit on my committee.

I am very thankful to my co-authors for our fruitful collaborations and their contributions to this thesis: Prof. Erman Ayday, Prof. Hossein Manshaei, Prof. Matthias Grossglauser, Dr Amalio Telenti, Dr Julien Freudiger, Dr Reza Shokri, Théophile Studer and Francisco Santos. I am also grateful to Dr Olivier Levêque, Prof. Boi Faltings, Prof. Jean-Yves Le Boudec, Prof. Jens Grossklags, Dr Abdelberi Chaabane, Dr Mohamed Ali Kaafar, Dr Paul J. McLaren, Dr Jacques Rougemont, Prof. Vincent Mooser, Dr Cihangir Tezcan, Kévin Fahy, Roger Michoud for their helpful input and ideas. Thanks also to our secretaries and system administrators for all the great support they provided me during my thesis. I am especially grateful to Holly for her unceasing help in the improvement of my English.

I am thankful to all my colleagues and friends at LCA for sharing with me the joys and challenges of the Ph.D. process: Igor, Kévin, Mohamed, Sébastien, Julien H., Vincent, Nicolas, Dan, Christina, Lucas, Nevena, Julien F., Berker, Reza, Marcin, Hossein, Murtuza, Maxim, Erman, Jean Louis, Zhicong, Alexandra, Anh, Italo, and Huang. Special thanks to my great officemates, Igor and Kévin, for their friendship, moral and technical support, and the discussions and fun shared in BC210. My Ph.D. years would not have been as pleasant without the support of my friends outside LCA: Yann, Jean-Baptiste, Sergio, Arman, Caterina, Nathalie, Mirjam, Célia, Célien, and Aditya. This list is certainly not exhaustive. I also would like to mention here my gratitude to all the wonderful people I have met at Amnesty International, at U.C. Berkeley, as well as all the people who have contributed to give meaning to my life.

I am indebted to my (step)family for their unconditional support and love. I dedicate this thesis to my mother for her support during all my studies, and for all the invaluable things she has taught me. Last but not least, my heartfelt thanks go to Aude, for being on my side during the toughest periods of my doctoral experience, for her love, and the happiness and meaning she has brought into my life.

Contents

Chapter 1

Introduction

Since its popularization in the 1990s, the Internet has dramatically changed the world we live in. Among various benefits, the Web has enabled decentralized information and communication on a large scale, thus diminishing censorship and control over public opinion by political or economic powers. The Arab Spring of 2011 is certainly the best example of the positive impact of the Internet on free speech, civic rights, political freedom and democracy. Social media, such as Facebook or Twitter, were instrumental in the organization of the protests, and in the dissemination of information. The other side of this digital revolution is in the new forms of surveillance and control it enables. Billions of dollars are invested by surveillance agencies in both democratic and authoritarian regimes to intercept and analyze communication data, and in some countries, to censor political or religious content. Increasing computing and storage capabilities enable global data mining and lead to perpetual electronic surveillance.

Surveillance, however, is at odds with privacy, a fundamental human right recognized by the Swiss Federal Constitution (article 13), by the European Convention on Human Rights (article 8), and by the Universal Declaration of Human Rights (article 12). In many aspects, privacy (including anonymity) is a condition for democracy, as it is essential to the preservation of freedom of speech. The protection of sources in journalism also ultimately relies upon privacy and anonymity. Recent revelations about mass surveillance by Western government agencies have shed light on the right to privacy, generating intense debate about the limits of this right and the balance between privacy and (national) security. These revelations also highlight the privacy risks caused by trading our data off in return for free services, such as Web search, e-mails, or social media. We should keep in mind that by doing so, we, as Internet users, are also, and perhaps primarily, fuelling the current immense privacy erosion.

By the number of their users and the scale of deliberate data disclosure, online social networks (OSNs) are probably the most relevant example illustrating this phenomenon. By providing their billion users with platforms for sharing their pictures, videos, interests, political views, emotions, and other intimate data, online social networks fulfill the human need for social recognition, gratification and publicity. Social media especially encourage data over-sharing, as their business model essentially relies on targeted advertising, thus users' data. But this trend seriously threatens users' privacy. First, according to revelations about the PRISM program, companies such as Facebook and Google per-

mit U.S. and British intelligence to directly tap into their central servers to track foreign targets [1]. Second, government agencies of authoritarian regimes, even though they do not have direct access to the OSNs' servers, also exploit OSNs to infiltrate protesters' social networks. Indeed, several Syrian activists reported having been arrested and forced to reveal their Facebook passwords [141]. Third, recruiters are known to look up OSN profiles of job applicants, potentially leading to hiring discriminations. Some employers and colleges even request the Facebook passwords of job applicants and students in order to get full access to their profiles [156].

Following the over-sharing behavior driven by OSNs, some people have started publicly disclosing their most intimate biological data, i.e. their genetic data. With the help of rapidly developing technology, the cost of DNA sequencing has dramatically decreased. This has allowed the availability and use of genomic data in research, healthcare, law enforcement, and recreational applications. Moreover, individuals can now obtain the sequencing of a significant part of their DNA (genotype) for less than $100 via direct-to-consumer genetic testing. These individuals can then use this genomic data to learn about their ancestries, their predispositions to diseases, and even their (genetic) compatibilities with potential partners. Following the trend exemplified by online social networks, some individuals bring the disclosure of personal data to new heights, by revealing their genomic data on genome-sharing websites (such as OpenSNP [2] or Personal Genome Project [3]). Today, there are already thousands of genotypes available online, and this number continues to increase. The sharing of genomic data might be seen as more benevolent than the egocentric storytelling of OSNs,[1] but this does not at all diminish the huge privacy risks inherent to this very sensitive information.

First of all, as genomic data carries information about our predisposition to diseases and physical traits, it can be used to infer future physical conditions. As a consequence, access to this data can potentially lead to serious discriminations in health insurance, life insurance, or mortgages. Furthermore, genomic data could also be used to discriminate people in their work, sports, and eventually in their whole life ambitions, as very well portrayed by the 1997 movie *Gattaca*. Moreover, as it also carries information about kinship, genomic data can lead to familial nightmares, for instance, divorce caused by the discovery of illegitimate offspring [5]. Last but not least, as DNA bears detailed information about our ethnicity and ancestries, it could be leveraged by racist movements to discriminate people based on their genetic origins. We should definitely not underestimate this risk, which is of low likelihood, but whose effect would be of extreme magnitude. The European tragedy of the Holocaust should remind us that ethnic differences can be exploited for evil ends. We can also imagine how the systematic racial segregation could have been even worsened if detailed genetic profiles had fallen into the hands of Nazi authorities [6]. Such tragedies could occur again, against any ethnic minority, especially with the rise of far-right parties and hate ideas in today's Europe. Our duty is to limit the risk of such systematic segregation by preventing the leakage and dissemination of genomic data.

Following Alan Westin's definition [171], privacy is the right to control, edit, manage, and delete information about oneself and decide when, how, and to what extent information is communicated to others. Therefore, each person should be empowered to manage and preserve his own privacy, individually and independently of others. How-

[1] A recent survey showed that the first motivation of individuals who publicly share their genotypes was to help research [4].

ever, there are many online activities where the privacy attitudes of others matter as much as our own. The best example is probably e-mail service providers that have most of our e-mails, even if we do not use these e-mail providers ourselves [7]. Unfortunately, at both social and biological levels of our lives, which together determine our identity, there exist mutual privacy risks. At the social level, interdependence emerges from our social ties, our friends, acquaintances or colleagues. Consequently, by their behavior in OSNs, they can reveal information about us that we cannot control. At the biological level, interdependence is inherent to genetic inheritance that relates our genome to those of our family members. Interdependence is an integral part of mankind, and privacy is no exception.

In this thesis, we investigate how these interdependent risks affect our privacy at both of the aforementioned levels of lives; we also propose countermeasures and solutions to mitigate indirect threats caused by others. We conclude this thesis with the study of privacy-enhancing systems where others, instead of damaging our own privacy, are actually needed to preserve it, thus demonstrating that interdependence can also be beneficial to privacy.

Contributions

In this thesis, we explore the problem of **interdependent privacy risks and protection in today's highly connected world**. We primarily study mutual privacy risks incurred by individuals who are, by nature, biologically and socially interconnected. We also propose solutions that eventually all require some degree of cooperation among individuals. Finally, we demonstrate how others can play a positive role for individuals' privacy when appropriate incentives are put in place.

Our contributions are as follows:

1. We study interdependent privacy risks in online social networks. We propose a navigation privacy attack, where an external adversary exploits the public social links and public attributes of users to find a target user. We describe a search algorithm that, in order to efficiently navigate towards target users, relies upon geographical attributes as well as occupation attributes. Our results show that the majority of users of two prominent OSNs can be found with our algorithm. This study complements the existing work on inference attacks in OSNs by showing that an OSN user cannot hide simply by excluding himself from a central directory or search function, even in a network with more than one billion users. Our findings also demonstrate that with the current privacy policies of most OSNs there is no OSN-membership privacy; for instance, the Syrian activists or job applicants mentioned above would have no chance of denying the existence of their OSN accounts. We suggest countermeasures that could easily be implemented by OSN operators in order to prevent this uncontrollable loss of privacy.

2. Considering the emergence of direct-to-consumer genetic testing and the resulting increasing use of genomic data for various purposes, we tackle the novel problem of genomic privacy. The first step in this endeavor is to formally measure the threats. To this end, we propose a probabilistic framework that relies upon Bayesian inference to quantify the genomic privacy and health privacy risks, including those induced by a person's relatives. We show that interdependence within a family

can have a serious impact on the family members' privacy levels. We illustrate the significance of the threat by carrying out a cross-website attack, using OSNs as a side channel to gather kinship information. Moreover, in the context of personal genomics, we study the interactions between relatives with different interests and behaviors regarding the storage security and sharing of their genomic data. We consider only purely selfish relatives who are willing to maximize only their own utility. We extend the game-theoretic framework to also take into account the altruistic behavior that can stem from familial ties. We also propose to rely upon multi-agent influence diagrams in order to efficiently predict equilibrium behaviors in the general scenario where more than two family members interact with each other. Finally, we develop a genomic-privacy preserving mechanism based on obfuscation that allows individuals to share (part of) their genomic data, while preserving some of the genomic privacy of their relatives. We emphasize that this is possible only if the family members care about each other, thus would cooperate with each other.

3. To end on a positive note, we study cooperative privacy-enhancing technologies where others do not compromise our privacy but actually help improve it. There are plenty of such systems where we need to rely upon our peers to protect our privacy. Indeed, one of the current most popular privacy-enhancing tools (Tor [56]) is based on the cooperation of others. The cost of cooperation causes a lack of such benevolent agents, which remains one of the main issues in Tor. We suggest that monetary incentives could be put in place in order to foster cooperation of more users in anonymity networks and in other privacy-enhancing systems. Under this assumption, we study the optimal amount of virtual money needed to maximize the efficiency of the system. To achieve this goal, we propose a scrip system model, which notably enables us to demonstrate that threshold strategies lead to a stable equilibrium. We evaluate the effect of various parameters on the optimal amount of money. Finally, we apply our analytical findings to real-world applications, such as anonymity networks. This part of the thesis demonstrates that our novel scrip system can help improve fairness and efficiency of cooperative privacy-enhancing systems via well-designed monetary incentives.

Thesis Outline

Following the three main areas of contributions described above, this thesis contains three parts. Part I discusses the interdependent privacy risks in online social networks. In particular, we show in Chapter 2 how any external adversary can find target users by exploiting publicly revealed information and misaligned privacy settings of OSN users. In Part II, we study interdependent privacy in the genomic context. In Chapter 3, we focus on the quantification of genomic privacy risks inherent to kinship. In Chapter 4, we analyze the interplay between family members with non-cooperative behaviors in the genomic-privacy context. In Chapter 5, we describe a genomic-privacy preserving mechanism that relies upon some degree of cooperation between relatives. In Part III, we study systems where interdependence can be beneficial for privacy. In Chapter 6, in particular, we investigate how incentives can be optimized in order to encourage cooperative

behavior in mutual privacy protection, thus eventually maximizing fairness and efficiency of the privacy-enhancing systems.

Publications

Chapter 2 is an extended version of [94]. Chapter 3 is an extension of [89], whereas Chapter 4 is based on the results of [91]. Chapter 5 contains the findings of [90]. Finally, Chapter 6 rests on the results of [93].

Part I

Interdependent Privacy in Online Social Networks

Chapter 2

Navigating around Privacy in Online Social Networks

2.1 Introduction

Over the last few years, online social networks (OSNs) have revolutionized the way people behave and interact with each other over the Internet. OSNs enable the majority of users to not just be passive consumers of the Web, but to become active producers of content, and to be storytellers of their own lives for the first time online. The other side of the coin is that privacy breaches are intrinsically bound to OSNs, and new forms of surveillance and control have emerged with OSNs. Recruiters are now known to look up Facebook profiles of job applicants, and hiring discrimination based on OSNs has become a serious threat [19, 66]. Some employers and colleges even request the Facebook passwords of job applicants and student athletes in order to get full access to their profiles [156]. OSNs have also been exploited by government agencies of authoritarian regimes to infiltrate protesters' social networks. Several Syrian activists have notably reported having been arrested and forced to reveal their Facebook passwords [141]. These practices are only the tip of the iceberg of privacy erosion caused by OSNs.

The first, straightforward method for finding an individual in an online social network is to rely on a central directory, if available. Obviously, a user u trying to keep his profile private would opt not to be listed in such a directory or, if this privacy option is not available,[1] make use of a pseudonym. The second method to reach u is to rely on the social links between users and to navigate via these links towards u. This approach works if some of u's friends show their friend lists publicly (thereby exposing u), which is the default setting in most OSNs.

In order to find a hidden user, an attacker could search the whole public social graph. However, such an exhaustive search, despite guaranteeing to find any user in the giant component,[2] would certainly be too expensive for OSNs that contain hundreds of millions users, notably because of the anti-crawling features deployed by virtually all OSNs. To

[1] Since the end of 2012, Facebook does not allow its members to remove themselves from the search directory, even though this is considered to be an important privacy setting [75].

[2] This holds if the search starts from the giant component and the target is in this component too. This is a fair assumption for current OSNs; for example, in 2011, researchers found that 99.91% of users belonged to the giant component in Facebook [164].

reduce the search cost, the attacker can decide to crawl only a targeted subset of OSN users. In this chapter, we evaluate the feasibility of such an attack for *large* networks and ultimately answer the following question: Is it possible to find a target profile by navigating a small fraction of the whole network, by relying on public attributes of queried profiles? Answering this question is crucial for privacy, because reaching the target profile or its neighborhood is *the necessary precondition* for any targeted attack such as the inference of hidden attributes (e.g., political or religious views) through other personal attributes [45, 130], or through friends' public attributes [115, 135, 54].

To the best of our knowledge, this is the first work proposing to make use of social links between users to find a target profile in an OSN. Our *navigation attack* is generic in order to apply to any attribute-enhanced OSN (such as Facebook, Google+, or Twitter). We propose a search algorithm that relies on a space of attributes and distance heuristics based on A^* search [83]. The categories of attributes and their priorities can be adapted to any kind of OSN. We show how the attack can be efficiently carried out, given the OSN visibility and privacy policies, and the users' privacy choices, by implementing it in the two largest OSNs, Facebook and Google+. For this OSN, building upon results on navigation and routing in social networks, the attack first relies on geographical attributes only, making use of additional types of attributes (such as education or work) as soon as it reaches the target's city. Our results demonstrate that 66.5% of Facebook users are findable by crawling a median number of users smaller than 400, and 59% of Google+ users are findable by crawling a median number of users small than 300. This shows that it is very difficult to hide in an OSN, however large it is, and thus to prevent targeted attacks and/or to deny the existence of a profile. Moreover, targets' cities are reached in 92% and 93.5% of the cases by crawling a median number of 13 and 8 users, in Facebook and Google+, respectively. This shows the efficiency of geographic navigation in Facebook and Google+. We propose two main explanations for the failed cases. First, the targets least likely to be discovered are those who have a small number of friends, or have privacy-cautious friends (not revealing too much information), or friends whose revealed information is not similar to their own information. Second, users living in larger cities tend to be harder to discover than others, especially in Facebook. Whereas the latter reason is inherent to the structure of the OSN and to the limit we impose on the number of crawled users, the former is essentially due to the privacy settings of the targets' friends and the OSN dynamics. Our results demonstrate that homophily in social networks [131, 25] does not only allow to infer hidden attributes of OSN users locally, but also allows to efficiently navigate toward the target. Note that we do not assume any prior knowledge on the network structure and the users' distribution in the network. Moreover, by starting the navigation from a random user in the network, we consider the worst-case scenario for the attacker, and provide a lower-bound on the attack efficiency. It is clear that the use of advanced search filters or source users closer to the target can only further benefit the attacker. We briefly show how this can dramatically reduce the search cost. Finally, we show that simple countermeasures exist and could be implemented directly by the OSN operators.

2.2 Model

OSN Model Online social networks can be described as social links between online users who own a personal profile. Formally, an OSN can be defined as a graph $G = (V, E)$,

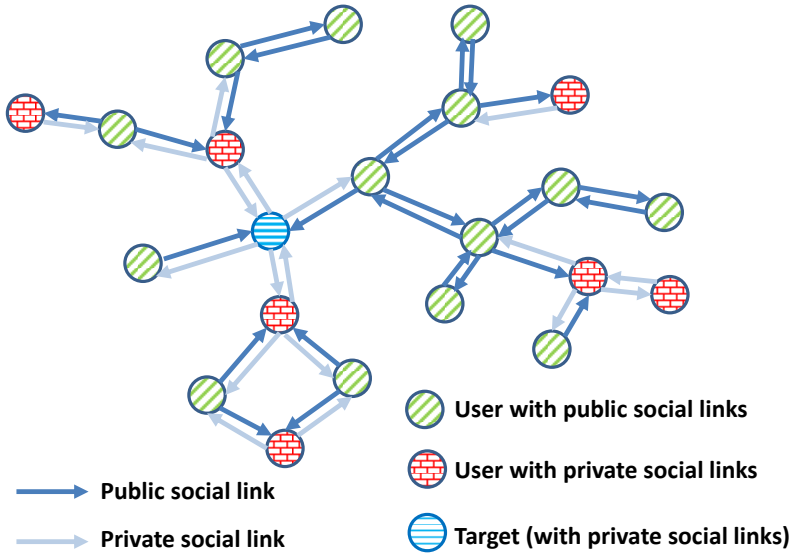


Figure 2.1: OSN model. The target keeps his social links private, but two of his neighbors make these links public.

where the vertex set, V, represents the set of users[3] and E, the edge set, their social links. Each user $u \in V$ is endowed with a set of attributes $\mathcal{A}_u$ that is a subset of the set $\mathcal{A}$ of the available attributes (gender, birthdate, education, city, ...). OSNs with symmetric social links requiring mutual consent, such as Facebook or LinkedIn, can be modeled as undirected graphs, whereas OSNs with asymmetric social links, such as Twitter or Google+, can be represented as directed graphs.[4]

In most OSNs, users can decide to what extent and with whom they share information by appropriately tuning their privacy settings. For instance, in Facebook users can reveal personal attributes to *friends* only, to *friends of friends*, or to *everyone* in the OSN. The same settings are generally available for their list of social links. Embedding users' privacy settings on their social links into the original social graph G induces a directed public subgraph D, where directed edges are those whose tail vertices have publicly available social links. Formally, $D = (V, E_d)$, with $E_d = \{(u, v) | (u, v) \in E, \Gamma(u) \neq \varnothing\}$, where $\Gamma(u)$ represents the out-neighbors of $u \in D$. Note that we make the conservative assumption that all privacy settings except the public one (e.g., *everyone* in Facebook) are private (e.g., *friends*, *friends of friends*), as we cannot access the information if we are not part of a user's close neighborhood. Figure 2.1 shows a simple example of an OSN with 22 users, among which 7 have private social links.

Adversary Model The attacker can be any external curious entity that wants to collect data or infer information about a target t. We assume that the attacker controls at least one node and can thus have access to information publicly visible in the OSN. In order to reach his target, the attacker will search the public subgraph D, relying on all public

[3] In the rest of the chapter, we will alternatively write *user*, *node* or *vertex* to refer to a member of the OSN.

[4] Note that Facebook now also allows asymmetric social links, by enabling users to become subscribers of other users.

social links and other public personal attributes (such as place of residence and work, educational affiliations, hobbies, etc.). We assume this attacker to have prior knowledge on the values of a subset $\mathcal{A}'_t$ of t's personal attributes, that he will use to navigate towards the target. As the attacker will reach the target through the target's social links (friends, friends of friends, ...), he will also discover at least one friend of the target, which can be useful for friend-based inference attacks [135, 54, 173]. Finally, note that the attacker we consider in this work is passive, in that he does not subvert any user account or interact with other OSN users, e.g., to create social ties with them.

2.3 Approach

We present here our navigation attack and algorithm. This attack is generic in order to apply to any attribute-enhanced OSN. We suppose that the attacker cannot rely on any search directory to find the target or to jump towards any user close to the target and that the navigation's starting point is randomly selected. This helps us evaluate the feasibility of a navigation attack in the worst-case scenario, and provide an upper-bound on the number of nodes that need to be crawled before reaching a target in general. In Subsec. 2.5.2, we nevertheless show how the attacker can take advantage of search filters to quicken the navigation.

In the generic scenario, the attacker navigates from user to user through public social links, until he reaches the target. He makes an informed decision about the next user to visit by relying on information publicly revealed by users at each hop towards the target and on his prior knowledge about the target. Whereas in Milgram's experiment [133] every participant in the chain could rely on his own local information about his acquaintances to make a decision about the next user to select, the attacker here relies on global information bounded by the attributes publicly revealed by users on the path. Our navigation attack is represented by Algorithm 1, called TargetedCrawler. This generic algorithm relies on a heuristic model inspired by A^* search [83].

The TargetedCrawler's inputs are (i) the source user s, from which the attacker will start crawling, (ii) the target user t that he has to reach, (iii) a subset of the target's attributes $\mathcal{A}'_t \subseteq \mathcal{A}_t$ known a priori by the attacker, (iv) the distance functions for each attribute, and (v) the priority of the attributes. The priorities depend essentially on the OSN and on the prior knowledge about the target's attributes. For instance, we will give higher priority to profession or workplace attributes in job-oriented OSNs (such as LinkedIn), to interests in microblogging OSNs (like Twitter), or to geographical attributes for mobile OSNs. The highest- and lowest-priority attributes will be represented as A^1 and A^N, respectively. The algorithm outputs t's profile and the shortest discovered path from s to t.

The total estimated cost c_u (line 13) from the source to the target at some node u on the path is divided into (i) the cost from the source to u, $d_{\text{hop}}(s,u)$ (hop distance), and the estimated remaining cost from u to the target, $d_{\text{rem}}(u,t)$, that is expressed as

$$d_{\text{rem}}(u,t) = \begin{cases} k_h d_h(A_u^h, A_t^h) & \text{if } d_j(A_u^j, A_t^j) = 0 \ \forall j < h \\ k_1 d_1(A_u^1, A_t^1) & \text{otherwise} \end{cases} \tag{2.1}$$

where $d_h(A_u^h, A_t^h)$ is the distance function between users u and t in the attribute h (attribute with h^{th} priority). The distance functions can be represented by (i) binary

Algorithm 1: TargetedCrawler

$\mathcal{F} \leftarrow s$ % Initializing the frontier with the source user
$\mathcal{E} \leftarrow \varnothing$ % The explored set is initially empty
repeat
  if $\mathcal{F} = \varnothing$ **then**
    Failure
  else
    Select the user $u^* \in \mathcal{F}$ with the lowest estimated cost to the target t and remove it from $\mathcal{F}$
    $\mathcal{E} \leftarrow u^*$
    if $t \in \Gamma(u^*)$ **then**
      Return t's profile and the path from s to t
    else
      for all $u \in \Gamma(u^*)$ **do**
        $c_u = d_{\text{hop}}(s, u) + d_{\text{rem}}(u, t)$
        if $u \notin \mathcal{F}$ AND $u \notin \mathcal{E}$ **then**
          $\mathcal{F} \leftarrow (u, c_u)$
        else if $u \in \mathcal{F}$ AND $c_u < c_u^{\text{old}}$ **then**
          $c_u^{\text{old}} = c_u$
          Replace the former parent of u by u^*
until t reached

values (e.g., 0 or 1 for last names), (ii) real values (e.g., difference for ages, or geographical distance for locations), or (iii) integers based on hierarchical decompositions (e.g., half the tree distance for tree-based hierarchies). k_h is a normalization parameter translating the attribute distance into a hop distance. k_h should decrease with h, as the more attributes we share, the closer to each other we should be. With d_{rem}, the targeted crawler will reach a user sharing the same first-priority attribute as the target before considering the second-priority attribute, then reach a user sharing a second-priority attribute before considering the third-priority attribute, and so on. We conjecture that OSN users share certain categories of attributes more than others (depending on the OSN) and that these attributes affect the way users cluster on different OSNs. Thus, in order to increase the search efficiency, we prioritize different categories of attributes depending on the type of OSN.

2.4 Experiments

As the current largest OSN (1.1 billion users as of March 2013), Facebook is the most representative candidate for evaluating our attack. Moreover, its privacy policies are notoriously designed to encourage public disclosure: the default policy for many important user attributes is *everybody*, i.e., full public visibility.[5] We also implemented our attack in Google+ in order to validate our findings in Facebook. This OSN is now the second largest OSN after Facebook [169] and shares many privacy features with Facebook. It also reveals the users' social links by default but, contrary to Facebook, allows users to be not searchable by name.

[5]As of this writing, this is the case for the following attributes: current city, hometown, sexual orientation, friend list, relationship status, family, education, work, activities, as well as music, books, movies, and the sports users like.

2.4.1 Implementation in Facebook and Google+

Gathering Source-Target Pairs Before beginning the navigation attack, we had to collect source users from which to start and target users to be reached. We chose to select pairs of users that would act both as source and as target to further evaluate the paths' symmetry. In order to have representative and meaningful results, we wanted to avoid sampling biases as much as possible. Unfortunately, as Facebook and Google+ IDs are encoded over 64 bits, there is a very small probability that a randomly generated ID corresponds to an existing profile.

For this reason, we decided to sample on the Facebook directory to gather source and target profiles, as in [45]. The Facebook directory[6] has a tree structure, and profiles are sorted in first-name alphabetical order. The first layer of the tree is divided into Latin characters and non-Latin characters. Then, all subsequent layers are divided by alphabetical order into at most 120 subcategories until the fifth layer where we can actually select users' profiles. At each layer of the directory tree, we randomly selected one branch, until we reached the last layer, where we randomly selected one profile. Unfortunately, Google+ does not provide such a public directory. Thus, we decided to sample source and target users by relying on a random walk method. Our method starts by walking through 50 different profiles in order to reach a random profile in the network. Once we have reached this profile, we select a node with probability inversely proportional to its (bidirectional) degree to be added to the source-target set. This probability compensates the random walk bias towards high degree nodes [73]. Finally, we retain only profiles with at least two publicly accessible attributes, assuming these to be part of the attacker's prior knowledge.[7] Note that we repeat the random walk through 50 profiles for each new node that we add into the source-target set. We discuss selection bias and the representativeness of our target set in Subsection 2.4.2.

Navigating in Facebook and Google+ Because of the very limited Facebook API, we had to implement our own crawler of users' friend lists. With the standard HTTP request to access the friend list, Facebook provides only the first 60 friends of a user. Then, it dynamically provides the rest of the friends if the Web user scrolls down the friend list's page. While the user is scrolling down, his Web browser actually sends an Ajax request to get the subsequent 60 friends in the friend list. The server replies in about 2 seconds with a JSON (JavaScript Object Notation) object that contains the next 60 friends in the list. We parsed the list of user IDs of each JSON object, as well as the additional piece of information (if any) provided right below each friend's name that would be used for the navigation. We also implemented our own crawler for Google+. We could get both all outgoing and incoming social links with only two HTTP requests. Both requests returned a JSON object with the social links' profiles (names), and some attributes (including location, employer, education, profile picture) also useful for the navigation.

Several lessons can be learned from previous work on navigation in social networks: (i) Geography and occupation are the two most crucial dimensions in choosing the next hop in a chain [110]; (ii) geography tends to dominate in early stages of routing [58]; (iii) adding non-geographic dimensions once the chain has reached a point geographically

[6] http://www.facebook.com/directory

[7] This does *not* mean that a target without any publicly available attributes could not be found. We need this information here to replace the prior knowledge the attacker is assumed to have.

close to the target can make the routing more efficient [162, 170]; and (iv) seeking hubs (highly connected users) seems to be effective in some experiments [162, 23] and to have limited effect in others [58]. As Facebook and Google+ share many properties with real social networks, we incorporate these findings into our navigation attack in order to maximize its efficiency. We select location (*current city* or *hometown*) as the first-priority attribute in Algorithm 1, and education, employer/workplace, and last name as second-priority attributes. We make this choice also because of the OSN structure and design. All aforementioned attributes are those most publicly shared by the Facebook and Google+ users. Location (*current city* or *hometown*), *education* and *work* are publicly revealed by around 35%, 30%, and 25% of the Facebook users, respectively [45, 78]. In Google+, location, education, and employer are publicly shared by 26%, 27%, and 21% of the users, respectively [129]. Moreover, these attributes are directly available from the social links' JSON objects, allowing us to not crawl all friends' profiles individually, and thus dramatically decreasing the number of HTTP requests and crawling time.

We propose to rely on two different types of distance function to evaluate the similarity between two locations. The first metric is computed as half the tree distance where the tree is defined by a discrete geographical hierarchy: $d_1(A_u^1, A_t^1)$ is equal to 3, 2, 1, or 0, if user u shares a continent, a country, a region/state or a city, respectively, with the target t. $d_1(A_u^1, A_t^1) = 4$ if u and t are from different continents. The second distance metric relies on the real geographical distances between two locations and $d_1(A_u^1, A_t^1)$ is then defined as

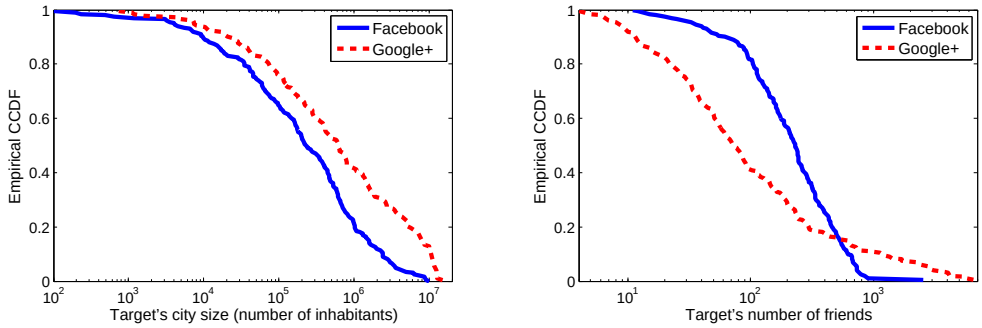
$$d_1(A_u^1, A_t^1) = \max(0, \log(d_{\text{geo}}(u, t)/\alpha)) \tag{2.2}$$

where the logarithm is base-10, d_{geo} is the great-circle distance (in km), and α is a normalization constant set to 1 km. We notice that this distance is very close to the discrete-hierarchy distance (first metric). In order to infer detailed geographical information from any location attribute, we relied on GeoNames[8], a Web service with a database containing over 10 million geographical names. More precisely, we used GeoNames (i) to find the region, country and continent associated with a city in the first distance metric and (ii) to compute the distance between two locations in the second metric. k_1 is set to 2 to get a maximal (theoretical) hop distance of around 8.

We give all non-geographical attributes (education, workplace and last name) second, thus same, priority. We make these design choices mainly because we can only access a single attribute in the Facebook users' friend lists (below each friend's name). These structural constraints, imposed by the OSN architecture, lead us to trade off some of Algorithm 1's steps against efficiency. Moreover, we make use of a binary distance function for these second-priority attributes (0 if two attributes match, 1 otherwise) because (i) we believe it is more efficient to directly select users based on whether they share the same attribute with the target once we have reached the same city, and (ii) it is particularly complex to build more elaborate distance functions for last names, employers, high schools or universities. k_2 can be set to any number strictly smaller than 2; we chose $k_2 = 1$.

For simplicity, we verify whether we have reached the target profile by checking his ID or alias, which both uniquely identify users. An attacker who is not supposed to know such identifiers will have to check the target's first and last names that, in addition to the location, should uniquely identify most of the people. In case there are multiple

[8] http://www.geonames.org/



(a) (b)

Figure 2.2: Empirical complementary cumulative distributions of (a) the targets' city sizes, and (b) the targets' degrees.

matching targets, the attacker could, for instance, just check the profile pictures of these few potential targets in order to select the correct target. Facial recognition could be further used to automatize the targets' check for targets making use of pseudonyms.[9]

2.4.2 Dataset Description

We ran our experiments on Facebook from April to November 2012, discontinuously and not too intensively. In this way, we avoided overloading the system and our crawler had a behavior similar to an energetic human user. Despite this, we attempted to reach 200 targets, collecting approximately 393k different friend lists, 197 million social links, and 138 million public user attributes. We also targeted 200 different users in Google+, during Spring 2013, collecting 398k friend lists and 139 million social links. For the Google+ crawler, we took similar precautions as for Facebook.

In both Facebook and Google+, we gathered targets in 42 different countries, spread over all continents. North America encompasses 33.5% of the targets in Facebook and 44% in Google+, Asia 26% in Facebook and 31% in Google+, Europe 18% and 15%, South America 13.5% and 8%, Africa 7.5% and 1%, and Oceania 1.5% and 1%. The continent distribution is quite close to the actual distribution of users' continents, except for North America that is a bit over-represented with respect to Europe and Asia. Regarding the countries, USA represents 26% of the targets in Facebook, followed by Indonesia, Brazil, and India, with 9.5%, 8.5%, and 8%, respectively. Almost the same sequence appears in Google+, with USA representing 38% of the targets, India 13%, Brazil 4%, and Indonesia 4%.

Regarding the targets' cities, we can notice in Figure 2.2(a) that the populations' distributions of Facebook and Google+ follow a similar shape, Google+'s targets living in cities with a bit more inhabitants in general. The average and the median city populations are equal to 870k and 233k, respectively, in Facebook, and to 2.6M and 440k, respectively, in Google+.

Regarding the targets' degrees (friends' or social links' numbers), we clearly notice a phase transition in the degree distribution (Figure 2.2(b)) in Facebook, which is very

[9]Face recognition has been shown to be very accurate and efficient for subject re-identification in OSNs [22].

Table 2.1: Success rates and numbers of crawled nodes for all continents.

	Facebook			Google+		
Continent	% success	# nodes: mean	median	% succ.	mean	med.
North Am.	71.6	1,065	467	67.1	668	272
Asia	51.9	1,061	658	49.2	565	179
Europe	86.1	513	144	53.3	348	72
South Am.	59.3	1,275	445	56.3	667	628
Africa	60	1,500	1,608	67	805	100
Oceania	66.7	2,270	553	100	92	14

similar to the one shown in [164]. Moreover, the average target has 291 friends, which is fairly close to the global average which was around 278 in April 2012 according to [82]. The targets' degree distribution is more scattered in Google+, with more targets having degrees smaller than 100 and greater than 1000. The median number of social links is equal to 71, smaller than Facebook, but its average is 424, greater than Facebook. It is hard to link these numbers with other studies, as Google+ is a recent OSN known to be evolving rapidly [129]. The geographical distance between sources and targets is quite uniformly distributed between 450 km (shortest distance) and 18,962 km (longest distance) in Facebook, and between 285 km and 15,814 km in Google+.

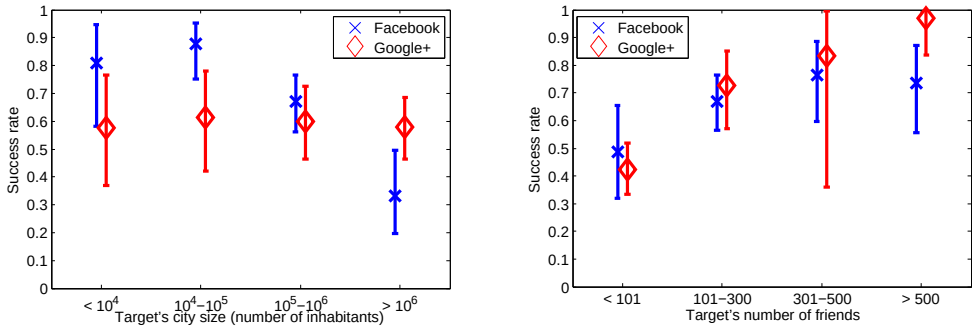
2.5 Results

In this section, we will first exhibit the results of our generic navigation attack, showing its success rate and efficiency. We will also provide some explanations for the failed cases. We will then mention how, by using some search filters, we can drastically reduce the crawling effort.

2.5.1 General Results

Our objective is *not* to launch a brute-force attack by crawling millions of nodes, which would demand a lot of resources. We rather aim to develop an algorithm that can reach a specific target in the network in a limited amount of time. For this reason, we decide to stop the attack after a certain number of crawled nodes, even if the frontier $\mathcal{F}$ is not empty. We choose a limit of 4,000 users, which already takes about 14 hours in Facebook (much slower than in Google+). We assume this is the maximum bearable time for an attacker attempting to reach someone in Facebook, and we keep the same limit with Google+, for consistency. Despite this limit, our attack successfully reaches its target in 66.5% of the cases in Facebook, and 59% of the cases in Google+. Using the Clopper-Pearson interval in order to evaluate the confidence interval for this success rate, we find that 95% of the users are reachable with a success rate in the intervals $[59.5\%, 73\%]$ and $[52\%, 66\%]$ for Facebook and Google+, respectively. The Clopper-Pearson interval is an exact method for calculating binomial confidence intervals. However, it is quite conservative, thus the intervals above might be wider than what it needs to be in order to achieve 95% confidence. Table 2.1 shows the success rates, average and median numbers of crawled nodes, for each continent.

We notice that the North American targets are reached quite successfully in both OSNs, whereas reaching Asian users are more challenging to reach. We also note that

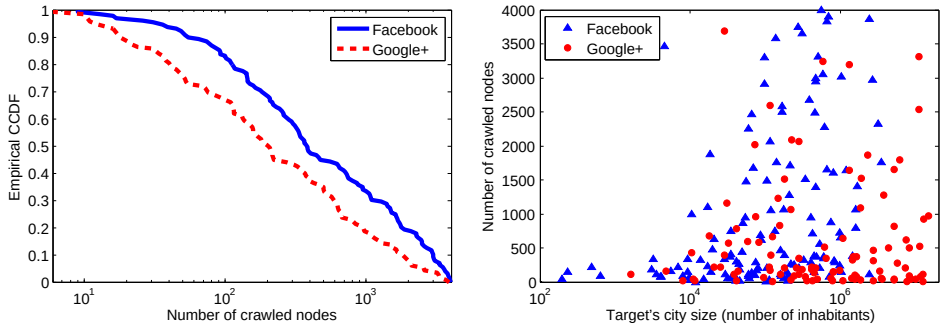


(a) (b)

Figure 2.3: Success rates with respect to (a) the target's city size, and (b) his number of friends. We made use of the Clopper-Pearson method to compute the 95% confidence intervals.

European targets are reached very successfully in Facebook but not in Google+. Figure 2.3 helps us understand these discrepancies. In particular, Figure 2.3(a) shows that the success rate drops with the size of the target's city in Facebook but not with Google+. We note in Figure 2.3(b) that the success rate increases with the target's number of friends, especially in Google+. Lower success rates in Facebook can be explained by comparing the average numbers of inhabitants for the different continents. We find that European and North American city populations have averages way below 1M (217k and 449k, respectively), whereas Asia, South America and Africa have average city sizes close to or above 1M (925k, 1.83M, and 2.46M, respectively). This lower success rate is certainly due to the fact that, in large cities, our algorithm has to crawl more nodes in order to cover all the users living in these cities. Our 4,000-node limit is certainly too low for such cities. However, this does not seem to explain the difference in success rates in Google+. This is probably due to the fact that Google+ being more recent and smaller than Facebook, there are less people publicizing the same city, thus less people to potentially crawl. What seems to have the highest impact on the success rate in Google+ is the number of friends of the targets. For instance, the median number of friends in Europe is equal to 33, where it is equal to 81 in North America. This is certainly due to the fact that the young age of Google+, and smaller adoption by European users. Note that there is no significant effect of the distance between sources and targets on the success rate. This shows that it is possible to efficiently navigate through large geographical distances in Facebook. We must also mention that source users have no effect on the success rate: all crawls successfully navigate out of the source neighborhood, and the large majority of them (92% in Facebook and 93.5% in Google+) reach the target's city.

We evaluate the nodes' efficiency by looking at the number of nodes crawled during our searches. Crawling a node in our experiment means crawling his friend list, not his personal profile. On average, 983 and 591 nodes needed to be crawled before a target could be reached, in Facebook and Google+, respectively. Half of the targets were attained in 380 and 291 or few nodes in Facebook and Google+, respectively. European targets were especially rapidly reached, after 513 and 348 nodes on average, half of the targets being found after less than 144 and 72 crawled nodes in Facebook and Google+,



(a) (b)

Figure 2.4: (a) Empirical CCDF of the number of crawled nodes in successful cases, (b) number of crawled nodes with respect to the target's city size (number of inhabitants).

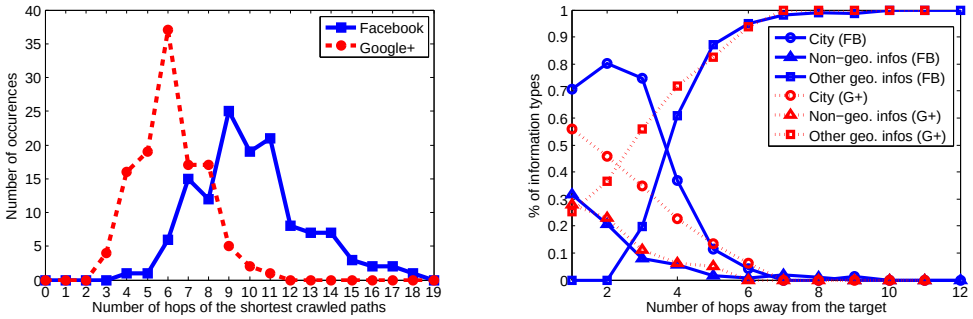
respectively. We see in Figure 2.4(b) that the number of crawled nodes is (positively) correlated to the target's city size. This is again due to the fact that more nodes will be seen in larger cities, thus reaching the target after a higher expected number of crawled nodes. It also tells us that the failures to reach European targets is not due to the city size but rather to the low number of neighbors. Moreover, for all failed and successful cases, on average 44 and 28 nodes had to be crawled before reaching a user in the target's city, and half of the searches found a user living in the target's city in less than 13 and 8 crawled nodes, in Facebook and Google+, respectively. This shows that our search algorithm makes use of long-range social links to efficiently reach the target's city, and that the most challenging part of the search is the navigation within the target's city, when we have to narrow down the search using second-priority attributes.

The target's neighborhood also has a huge impact on how easy this target can be reached. Some targets have only a few friends revealing their friend lists and who display information similar to the target's information. These targets have less chance of being reached. For instance, around 6.5% of the targets have no friends who publicly reveal their friend list *and* display information similar to the target's. Due to their privacy-cautious friends, these targets are obviously impossible to reach with our attack. Table 2.2 demonstrates the importance, for the success of the attack, of similar attributes being publicly shared by the target's friends. The difference between the median number of attributes (city, or other information) of successful and unsuccessful cases is very significant, especially in Google+. Furthermore, the median amount of attributes revealed by friends of non-reached targets is quite low. This leads us to conclude that, in addition to the size of the city, the amount of attributes revealed by the target's friends is crucial to the attack success. Whereas the influence of the city size is inherent to the OSN structure and to the 4,000-node limit that we impose, the influence of the target's friends is due to the OSN users and their privacy behaviors. Some users might also just have arrived in their current city, thus not have many friends yet in this city. They might also have education and work attributes that are not geographically correlated to their location, thus not be of great help for our attack. In order to improve the attack performance, we could target more than one cities when needed, e.g., the target's current city, hometown, and the city where he studied.

From each subgraph crawled during a successful attack, we reconstructed the shortest

Table 2.2: Number of similar attributes publicly revealed by the target's friends with public friend lists.

	Facebook				Google+			
	Success		Failure		Success		Failure	
	Average	Median	Ave.	Med.	Ave.	Med.	Ave.	Med.
City	17	12	8	3	282	45	22	6
Other infos	14.4	7	9.1	3	4	0	0.7	0



(a) (b)

Figure 2.5: (a) Histograms of the shortest discovered path lengths within the crawled subgraphs, and (b) evolutions of the information types used to navigate towards the target (information types shared by users on the shortest paths).

discovered path from the source to the target. Figure 2.5(a) illustrates the distribution of the shortest discovered path lengths. We notice that it goes from 4 to 18 hops in Facebook, most of shortest paths being between 9 and 11-hops long. This is approximately twice the distance found in [35] with the knowledge of the complete social graph. The shortest paths are between 3 and 11 hops in Google+, most of them being 6 hops long. This result is similar to the diameter obtained in [74], where 90% of the pairs were separated by a distance of 5, 6 or 7 hops.

We show in Figure 2.5(b) how the information the nodes on the shortest path (SP) display evolves. It shows that the city is especially useful 3, 2, and 1 hop(s) before the target, for both OSNs. At 4 (and more) hops from the target, other (non-local) geographical attributes are used to navigate towards the target. We also note that other types of attributes (education, work, or last name) begin to be more used 4 hops before the target (certainly once we have reached the target's city) and increase their influence while getting closer to the target. At the latest hop before the target, the city is represented in 70% of cases in Facebook and 56% in Google+, non-geographical information representing around 30% of cases in both OSNs. This shows that geographical information remains crucial, but also that other types of information can still be useful when we get close to the target, as it was already mentioned in [162]. Finally, we note that 25% of the targets in Google+ were found from a profile sharing no similar attribute with the target. These targets were reached from a user geographically close (at a median distance of 32 km) but not sharing the same location.

2.5.2 Jumping towards the Target

Facebook provides an additional feature in order to help people find their acquaintances in the network: It allows users to apply search filters on location, education or workplace. This means that, in addition to the first and last names, we can, for instance, specify the city of the searched person. We did not want to rely extensively on this feature for our navigation attack because we wanted to keep it generic and applicable to other OSNs. However, we show here that the attacker can take advantage of Facebook's search filters to facilitate his attack.

We search for the last names and the cities of the targets using the Facebook search filters, and then crawl the friend lists of the users found by the search directory. We search for last names because users sharing same last names are more likely to be relatives, thus to be friends. If more than ten users are found, we select the first ten displayed users as sources. Of course, our targets can appear in the users found by the search filters, as we chose targets that are in the Facebook directory for our experiments. Searching for the last names and the cities of our targets, we directly find the targets in 49.5% of the search results. As targets are assumed to not be in the directory, we remove them from the list of users to be crawled. At least 10 users satisfying the search criteria are found in 30% of the filtered searches, and the search requests output no user in 15% of the cases. By crawling only the friend lists of users found by our filtered search, we reach the targets with a success rate of 16.5%. This means that an attacker can find a target in only one hop (and a maximum of 10 crawled nodes) in 16.5% of the cases by relying on the Facebook search filters. It is interesting to note that 18.2% of the targets discovered in one hop are targets that were not found by our generic attack. Most of these are living in large African or South American cities. The size of these cities is probably the reason our targeted crawler did not find them in less than 4,000 crawled nodes.

2.6 Countermeasures

Countermeasures should logically be developed and implemented by the OSN operators themselves. An obvious solution, already advanced in [161], is to set the visibility policy as the intersection of visibility policies selected by all users involved in the published information. Although it is difficult to force a friend to change his privacy settings on his personal attributes, it is possible to enforce his social links' privacy policy. Choosing the intersection of both users' policies on social links would mean that a user electing to reveal his social links to his friends, or friends of friends only, would automatically enforce non-public social links for his own friends. It would prevent any curious stranger from accessing his profile by using his friends' friend lists. Another change in the privacy policies could be to automatically remove users who are not in the search directory from their peers' friend lists. OSN operators could finally also prevent anyone from publicly showing his social links, as it is the case in LinkedIn. They could at least design non-public default privacy settings on social links. Detailed formal requirements to protect multilateral privacy are presented in [80].

If the OSN operators themselves do not re-design their privacy policies, the users could also take action. The first option is to change the default privacy settings on social links to more restrictive settings. For this option though, users must collectively deviate from the default policy in order for it to be efficient. Another strategy is to

unfriend "dangerous" friends who publicly reveal their social links and other personal attributes. However, this strategy, already envisaged in a more general setting in [78], can dramatically spoil users' experiences and social lives. Finally, if more users decided to hide their personal attributes (such as city, education, ...), the attacker's ability to navigate efficiently in the social graph would decrease, thus reducing the threat presented in this chapter.

The last and most extreme countermeasure is certainly to change the full OSN architecture, and rely on a decentralized architecture with encrypted personal data and social links (e.g., [49, 96]), even though it seems too involved to be accepted by most of the OSN users.

2.7 Related Work

We present here the most closely related work on privacy threats in OSNs, showing how our work complements existing attacks. We also discuss the background on navigation in social networks.

2.7.1 Privacy Issues in Online Social Networks

Acquisti and Gross were among the first to mention the potential risks induced by information sharing in OSNs in their seminal papers [76, 21]. They study in detail the Facebook privacy settings and data visibility, and they emphasize the potential threats caused by weak privacy settings (used by most users). In [117] and [118], Krishnamurthy and Wills study what types of information are shared with whom, by default or not, and what kind of privacy settings are available for various pieces of *personally identifiable information*. They show that, among 12 OSNs, 10 publicly reveal social links by default and 1 reveals them always (i.e., without any possibility of changing the settings). 7 reveal by default the user's location and 5 always reveal it. 8 reveal the attended schools by default and 6 the employers. These statistics are relevant for our work as they show what kind of attributes are publicly revealed, and thus can be used for the navigation.

He et al. [84] were among the first to propose inference attacks based on the users' neighborhood. They make use of Bayesian inference and multi-hop inference to predict private attributes based on the friends, and friends of friends of the targeted users. The authors apply their analytical findings to a LiveJournal dataset with hypothetical attributes. In the same vein, Lindamood et al. propose to infer political affiliation (binary attribute: liberal or conservative) based on a modified Naive Bayes classifier [127]. Their results show that simply sanitizing user attributes or links is not enough to prevent inference attacks. Johnson [104] also emphasizes that social links can leak very sensitive information about a specific Facebook user, for instance whether a certain user is homosexual or not.

Zheleva and Getoor [174] propose novel inference attacks based on social links and group memberships, which they apply in four different social networks. Another work on inference of undisclosed attributes proposes to rely on any of the user's public attributes, and on any of the aggregates of his friends' attributes [115]. Finally, Chaabane et al. [45] show how music interests can be used to infer private sensitive attributes of Facebook users. Their approach does not rely on users' social links or group memberships, but only on users' attributes.

Thomas et al. [161] examine how the lack of joint privacy controls can put a user's privacy at risk. Notably, they highlight the inherent interdependent privacy risks due to friends in Facebook, and the fact that a user had no control over his friends' friend lists. They present inference techniques that, based on wall posts and friends, present improvements compared to previous work by relying only on friends to infer private attributes. Yamada et al. [173] also emphasize the impact of conflicting privacy policies on users' privacy. They propose 3 different attacks: friend-list, profile and wall-post recovery attacks. Dey et al. [54] estimate the leakage of age information in Facebook, either by relying on the target's profile directly, or by using information released by the targets' friends.

While these previous papers exploit the notion of homophily to infer hidden attributes of a user from the visible attributes of his neighbors, our work exploits the global structure of visible attributes to navigate efficiently towards a target. While the former is a purely local operation, ours exploits a macroscopic property of the social network. It complements existing work by showing how to efficiently find anyone in an OSN, necessary condition for any targeted inference attack.

Finally, Jain and Kumaraguru propose an integrated system which uses major dimensions of a user identity (profile, content and network) to search and link a user across multiple social networks [97]. Our work notably differs in the method used to search for a user. Our navigation attack does not require the targeted user to be present in multiple OSNs, and does not assume the target profile to be known in one OSN in order to find him in another.

2.7.2 Navigation in Social Networks

The seminal experiment by Milgram [133] shows that any arbitrarily selected individuals can reach any other person through a short chain of acquaintances. There generally exists a short path from any individual to another, thanks to a few long-range social links. However, knowing that short chains exist does not tell us how arbitrary pairs of strangers are able to find them. Since Milgram's experiment, there have been many theoretical and experimental papers that explain how people can find short paths, and thus navigate, in social networks [126].

Travers and Milgram ask 296 arbitrarily selected individuals in the United States to generate acquaintance chains (using postal mail) to a single target person. Out of the 296 starting chains, 64 reach the target (22% of completion rate) with a mean number of intermediaries between the sources and the target of 5.2 [162]. They also show that chains converge essentially by using geographic information; but once in the target's city, they often circulate before entering the target's circle of acquaintances. Dodds et al. propose a similar social-search experimental approach except that they rely on e-mails instead of classic postal service to reach a target [58]. This allows them to increase the number of targeted individuals (18 in 13 countries, instead of 1 target) and the number of distinct chains (24,163 instead of 296). In total, 384 out of the 24,163 chains reach their targets, showing an extremely low chain completion rate of 1.6% with an average chain length of 4.05. They show that geography clearly dominates the routing strategies of senders at early stages of the chains and is less frequently used than other characteristics (such as occupation) after the third step.

On a more theoretical side, Kleinberg develops a graphical model, d-dimensional

lattices encompassing the small-world properties, and derives several analytical results, notably showing the conditions under which a decentralized algorithm that uses only local information could efficiently (i.e., in polylogarithmic time) route messages from a source to a target [112, 113]. Considering another model, *rank-based friendship*, Kumar et al. prove that greedy routing can find a short path (of expected polylogarithmic length) from an arbitrary source to a target as long as the doubling dimension of the metric space of locations is low [120].

Watts et al. present a hierarchical model for categorical organization in social networks for message routing. They define the social distance between two people as the minimum ultrametric distance over all group hierarchies [170]. Eppstein et al. study the existence of mathematical frameworks that demonstrate the feasibility of local category-based routing in social networks [61]. They notably introduce the notion of membership dimension that characterizes the cognitive load of performing routing tasks in a given system of categories. Their results show how participants in a social network, while remembering an amount of information that is polylogarithmic in the size of the network, can efficiently route messages by using a local, greedy, category-based routing strategy. Liben-Nowell et al. study the role of geography in order to route messages in social networks and provide a theoretical model to explain path discovery [126]. To the best of our knowledge, they are the first to analyze routing in an "online" social network, namely the LiveJournal social network. However, they limit themselves to the problem of reaching the target's city. Among other results, they show that geography remains a crucial factor in online friendship and is thus very helpful when trying to reach a target. Lattanzi et al. extend this one-dimensional approach based on geographical proximity to a multidimensional space of interests relying on a model of social networks called "affiliation networks" [123].

Knowing that acquaintances' and social networks show small-world properties, we now question whether current OSNs do so as well. Mislove et al. already provided a piece of the answer to that question in 2007 [134]. The four considered OSNs exhibit power-law degree distributions, a densely connected core of high-degree nodes linking small groups of strongly clustered nodes and, as a result, short path lengths. Wilson et al. make another step in that direction, by crawling a significant portion of Facebook and showing its small-world properties [172]. A crucial step in providing evidence about the small-world characteristics of OSNs has recently been achieved with the publication of two reports by Facebook researchers on the Facebook full social graph [164, 35]. Their dataset of 721 million users follows the main small-world properties: 99.91% users belong to the largest component, the distribution of nodes degree follows a power-law distribution, and the average distance between users equals 4.7, showing that online social networks are even smaller than real-world social networks. We can thus predict that, by relying on users' attributes, most OSNs should also be navigable. However, how to efficiently navigate on them was until now an open question. Furthermore, Facebook reports considered the full social graph, with all social links, whereas the attacker assumed in this work would not have access to all those links. In this chapter, we study whether the public subgraph induced by the users' privacy settings on their social links is navigable by relying on publicly revealed attributes.

2.8 Summary

In this chapter, we have introduced a navigation privacy attack, where an external adversary attempts to find a target user by exploiting publicly visible attributes of intermediate users. We describe a search algorithm that relies on public attributes of users and distance heuristics. As most attributes (such as place of residence, age, or alma mater) tend to correlate with social proximity, they can be exploited as navigational cues while crawling the network. The problem is exacerbated by privacy policies where an OSN user who keeps his profile private remains nevertheless visible in his friends' "friend lists", leading to interdependent privacy risks.

Our search algorithm discovers more than 66% of the targeted Facebook users and 59% of the targeted Google+ users, in a median number of crawled nodes smaller than 400 in Facebook, and smaller than 300 in Google+. Moreover, the targets' cities are reached in 92%, respectively 93.5%, of the cases, in a median number of 13, respectively 8, crawled nodes, showing the efficiency of geographic navigation in Facebook and Google+. The navigation within the targets' cities, that rely on more attributes, is less efficient and successful. One important reason for the failed cases is the privacy behaviors of the target's friends: the more friends have public attributes and public social links, the more likely the target is to be found. This demonstrates the crucial role of social ties in OSNs, who can have a non-negligible impact on our own privacy. Finally, we highlight the increased risk induced by advanced search filters in OSNs.

Our results suggest that an OSN user cannot hide simply by excluding himself from a central directory or search function. This leads us to conclude that it is urgent that OSN operators implement countermeasures to thwart navigation attacks, thus to reduce interdependent privacy risks. The most obvious one is to set by default the social links (friend lists) to be non-public.

Consequences of our Work In addition to being a required prerequisite for most of the targeted attacks already proposed in the literature, our navigation attack also demonstrates that it is in most cases impossible for a user to claim that he does not have any account in a given OSN, thus jeopardizing OSN-membership privacy. This is of particular relevance when considering the Arab Spring. It is well-known that the successful protest against the Tunisian and Egyptian regimes was channeled by social media, and in particular Facebook. The security officials of those countries were apparently unprepared for such a threat and the rulers were toppled. But, meanwhile, the Syrian government seems to have learned the lesson. Several Syrian activists have indeed reported having been arrested and forced to reveal their Facebook passwords [141]. Of course, one of first reaction of an arrested activist was to claim that he did not have any Facebook account, but the police had already found his profile and were monitoring him. Considering our results, most political activists could never hide in Facebook. Our results also apply to the job applicants who were required by recruiters to allow for access to their entire profiles [156]. These individuals would also be affected by the attack shown in this chapter. Most of them could not claim that they do not have any Facebook account. This leads us to conclude that OSN-membership privacy is in jeopardy.

Part II

Interdependent Privacy in Genomics

Chapter 3

Quantifying Kin Genomic Privacy

3.1 Introduction

With the help of rapidly developing technology, DNA sequencing is becoming less expensive. As a consequence, large collections of human genomes are now available to geneticists, which dramatically increases the speed of genomic research and paves the way to personalized medicine. Furthermore, individuals can obtain the sequencing of the most significant part of their DNA (genotype) for less than $100 via direct-to-consumer genetic testing. Individuals are then using their genotypes to learn about their (genetic) predispositions to diseases, their ancestries (e.g., on 23andMe [8]), and even their (genetic) compatibilities with potential partners (e.g., on GenePartner [9]). This trend has also caused the launch of genome-sharing websites and online social networks (OSNs), in which individuals share their genomic data (e.g., OpenSNP [2] or 23andMe [8]).[1] Thus, already today, thousands of genomes are available online and this number keeps increasing.

Even though most of the genomic sequences on the Internet are anonymized, many individuals publish their genomes under their real identities (e.g., on OpenSNP). Furthermore, it has been shown that anonymization is not sufficient for protecting the real identities of the genome donors [81, 158]. The genome containing very sensitive information about ethnicity, kinship, and predisposition to diseases, its leakage/usage can lead to genetic discrimination (e.g., by employers or insurance companies) [29, 62], and even divorce [5]. Some believe that they have nothing to hide about their genetic structure, hence they might decide to give full consent for the publication of their genomes on the Internet to help genomic research. However, our DNA sequences are highly correlated to our relatives' sequences. The DNA sequences between two random human beings are 99.9% similar, and this value is even higher for closely related people due to familial inheritance. Consequently, somebody revealing his genome does not only damage his own genomic privacy, but also puts his relatives' privacy at risk [154]. Moreover, currently, a person does not need consent from his relatives to share his genome online, thus making the protection of genomic privacy even more complicated.

A recent New York Times' article [10] reports the controversy about sequencing and

[1] A survey about users' motivation for and fear about genome sharing can be found in [4].

publishing, without the permission of her family, the genome of Henrietta Lacks (died 1951). On the one hand, the family members think that her genome is private family information and it should not be published without the consent of the family. On the other hand, some scientists argue that the genomes of current family members have changed so much over time (due to gene mixing during reproduction), that nothing accurate can be told about the genomes of current family members by using Henrietta Lacks' genome. We will show in this chapter that they are, at least partially, wrong. Unfortunately, the Lacks family is only the tip of iceberg. As mentioned before, thousands of genomes are already available online, thus there are currently thousands of families facing the same threat. Once the identity of a genome donor is known, an attacker can learn about his relatives (or family tree) by using an auxiliary side channel, such as an OSN, and infer significant information about the DNA sequences of the donor's relatives. We show the feasibility of such an attack in Section 3.5.

Although the researchers took Henrietta Lacks' genome offline from SNPedia, other databases continued to publish portions of her genomic data. Unfortunately, publishing only portions of a genome does not, however, completely hide the unpublished portions; even if a person reveals only a part of his genome, other parts can be inferred using the statistical relationships between the nucleotides in his DNA. For example, James Watson, co-discoverer of DNA, made his whole DNA sequence publicly available, with the exception of one gene known as Apolipoprotein E (ApoE), one of the strongest predictors for the development of Alzheimer's disease. However, it was shown that the correlation (called *linkage disequilibrium* by geneticists) between one or multiple polymorphisms and ApoE can be used to predict the ApoE status [137]. Thus, an attacker can also use these statistical relationships (which are publicly available) to infer the DNA sequences of a donor's family members, even if the donor shares only part of his genome. It is important to note that these privacy threats not only jeopardize kin genomic privacy, but, if not properly addressed, these issues could also hamper genomic research due to untimely fear of potential misuse of genomic information.

In this chapter, we evaluate the genomic privacy of an individual threatened by his relatives revealing their genomes. Focusing on the most common genetic variant in human population, single nucleotide polymorphism (SNP), and considering the statistical relationships between the SNPs on the DNA sequence, we quantify the loss in genomic privacy of individuals when one or more of their family members' genomes are (either partially or fully) revealed. To achieve this goal, first, we design a reconstruction attack based on a well-known statistical inference technique. The computational complexity of the traditional ways of realizing such inference grows exponentially with the number of SNPs (which is on the order of tens of millions) and relatives. In order to reduce the complexity and infer the values of the unknown SNPs in linear complexity, we represent the SNPs, family relationships and the statistical relationships between SNPs on a factor graph and use the belief propagation algorithm [139, 119] for inference. Then, using various metrics, we quantify the genomic privacy of individuals and show the decrease in their level of genomic privacy caused by the genomes of their family members. We also quantify the health privacy of the individuals by considering their (genetic) predisposition to certain serious diseases. We evaluate the proposed inference attack and show its efficiency and accuracy by using real genomic data of a pedigree. More importantly, by using genomic data and familial information we collected from a public genome-sharing website and OSNs, we show that the proposed inference attack threatens not only the

Lacks family, but also many other families. We define in this chapter the quantification concepts and formalism that we will rely upon in Chapters 4 and 5.

3.2 Background

In this section, we briefly introduce the relevant genetic principles, as well as the concept of belief propagation.

3.2.1 Genomics

DNA is a double-helix structure that consists of two complementary polymer chains. Genetic information is encoded on the DNA as a sequence of nucleotides (A,T,G,C) and a human DNA includes around 3 billion nucleotide pairs. With the decreasing cost of DNA sequencing, genomic data is currently being used mainly in the following two areas: (i) clinical diagnostics, for personalized genomic medicine and genetic research (e.g., genome-wide association studies[2]), and (ii) direct-to-consumer genomics, for genetic risk estimation of various diseases or for recreational activities such as ancestry search. In the following, we briefly introduce some concepts, which we use throughout this thesis, about the human genome and reproduction.

Single Nucleotide Polymorphism

As already mentioned, human beings have 99.9% of their DNA in common. Thus, there is no need to focus on the whole DNA but rather on the most important variants. Single nucleotide polymorphism (SNP) is the most common DNA variation in human population. A SNP occurs when a nucleotide (at a specific position on the DNA) varies between individuals of a given population (as illustrated in Figure 3.1). There are approximately 50 million SNP positions currently known in the human population [11]. Recent discoveries show that the susceptibility of an individual to several diseases can be computed from his SNPs [103, 12]. For example, it has been reported that two particular SNPs (rs7412 and rs429358) on the Apolipoprotein E (ApoE) gene indicate an (increased) risk for Alzheimer's disease. SNPs carry privacy-sensitive information about individuals' health, hence we will quantify health privacy focusing on individuals' published (or inferred) SNPs and the diseases they reveal.

Two different nucleotides (called alleles) can usually be observed at a given SNP position: (i) the major allele is the most frequently observed nucleotide, and (ii) the minor allele is the rare nucleotide. From here on, we represent the major allele as B for a SNP position, and the minor allele as b (where both B and b are in $\{A, T, G, C\}$).

Furthermore, each SNP position contains two nucleotides (one inherited from the mother and one from the father, as we will discuss next). Thus, the content of a SNP position can be in one of the following states: (i) BB (*homozygous-major* genotype), if an individual receives the same major allele from both parents; (ii) Bb (*heterozygous* genotype), if he receives a different allele from each parent (one minor and one major); or (iii) bb (*homozygous-minor* genotype), if he inherits the same minor allele from both parents. For simplicity of presentation, in the rest of the thesis, we encode BB with 0,

[2]Examination of many genetic variants in different individuals to determine if any variant is associated with a trait.

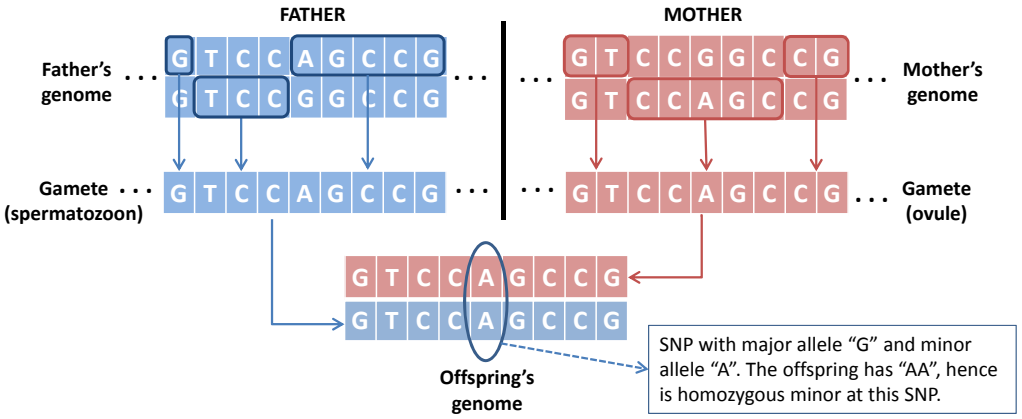


Figure 3.1: Reproduction and single nucleotide polymorphism (SNP). Each parent produces gametes that are derived from his or her genome. The offspring's genome is the combination of these two gamets. As an example, the SNP circled on the offspring's genome is homozygous-minor for the offspring but heterozygous for the parents.

		Father (*F*)		
		BB	Bb	bb
Mother (*M*)	BB	(1,0,0)	(0.5,0.5,0)	(0,1,0)
	Bb	(0.5,0.5,0)	(0.25,0.5,0.25)	(0,0.5,0.5)
	bb	(0,1,0)	(0,0.5,0.5)	(0,0,1)

Table 3.1: Mendelian inheritance probabilities $\mathcal{F}_R(\mathbf{X_M^i}, \mathbf{X_F^i}, \mathbf{X_C^i})$ for a SNP g_i, given different genotypes for the parents. The probabilities of the child's genotype is represented in parentheses. Each table entry represents $\big(P(\mathbf{X_C^i} = BB|\mathbf{X_M^i}, \mathbf{X_F^i}), P(\mathbf{X_C^i} = Bb|\mathbf{X_M^i}, \mathbf{X_F^i}), P(\mathbf{X_C^i} = bb|\mathbf{X_M^i}, \mathbf{X_F^i})\big)$.

Bb with 1, and bb with 2. Finally, each SNP g_i has a minor allele frequency (MAF), p^i_{maf}, which represents the frequency at which the minor allele b of the corresponding SNP occurs in a given population (typically, $0 < p^i_{\text{maf}} < 0.5$).

Reproduction

Mendel's First Law states that alleles are passed independently from parents to children for different meioses (the process of cell division necessary for reproduction). For each SNP position, a child inherits one allele from his mother and one from his father (as shown in Figure 3.1). Each allele of a parent is passed on to a child with equal probability of 0.5. Let $\mathcal{F}_R(\mathbf{X_M^i}, \mathbf{X_F^i}, \mathbf{X_C^i}) = P(\mathbf{X_C^i}|\mathbf{X_M^i}, \mathbf{X_F^i})$ be the function modeling the Mendelian inheritance of a SNP g_i, where M, F, and C represent mother, father, and child, respectively. We illustrate the Mendelian inheritance probabilities in Table 3.1.

Linkage Disequilibrium

Linkage disequilibrium (LD) [63] is a correlation that appears between any pair of SNP positions in the whole genome due to the population's genetic history. Because of LD, the content of a SNP position can be inferred from the contents of other SNP positions. The strength of the LD between two SNP positions is usually represented by the correlation coefficient r^2, where $r^2 = 1$ represents the strongest LD relationship.

3.2.2 Belief Propagation

Belief propagation [139, 119] is a message-passing algorithm for performing inference on graphical models (Bayesian networks, Markov random fields). It is typically used to compute marginal distributions of unobserved variables conditioned on observed ones. Computing marginal distributions is hard in general as it might require summing over an exponentially large number of terms. The belief propagation algorithm can be described in terms of operations on a factor graph, a graphical model that is represented as a bipartite graph. One of the two disjoint sets of the factor graph's vertices represents the (random) variables of interest, and the second set represents the functions that factor the joint probability distribution (or global function) based on the dependences between variables. An edge connects a variable node to a factor node if and only if the variable is an argument of the function corresponding to the factor node. The marginal distribution of an unobserved variable can be exactly computed by using the belief propagation algorithm if the factor graph has no cycles. However, the algorithm is still well-defined and often gives good approximate results for factor graphs with cycles. Belief propagation is commonly used in artificial intelligence and information theory. It has demonstrated empirical success in numerous applications including LDPC codes [140], reputation management [31], and recommender systems [30].

3.3 The Proposed Framework

In this section, we formalize our approach and present the different components that will allow us to quantify kin genomic privacy. Figure 3.2 gives an overview of the framework.

The SNPs of all relatives are represented by the random variable $\mathbf{X}$ that takes value in the set $\mathcal{X} = \{0, 1, 2\}^{n \times m}$, where n is the number of relatives in the targeted family and m is the number of SNPs in a single DNA sequence. Moreover, the hidden SNPs are represented by the random variable $\mathbf{X_H}$ (that takes value in the set $\mathcal{X}_H$), and the SNPs observed by the adversary by the random variable $\mathbf{X_O}$ (that takes value in the set $\mathcal{X}_O$). We define $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ to be the set of relatives in the targeted family (whose family tree, showing the familial connections between the relatives, is denoted as $\mathcal{T}$) and $\mathcal{G} = \{g_1, g_2, \ldots, g_m\}$ to be the set of SNPs (i.e., positions on the DNA sequence). Let $\mathbf{X}_j^i$, respectively $x_j^i \in \{0, 1, 2\}$, represent the random variable representing SNP g_i of individual r_j, respectively its value. Furthermore, we let $\mathbf{x}_i = \begin{bmatrix} x_i^1 & x_i^2 \cdots & x_i^m \end{bmatrix}$ represent the values of the SNPs of individual r_i, and $\mathbf{x} \in \mathcal{X}$ be the $n \times m$ matrix representing the

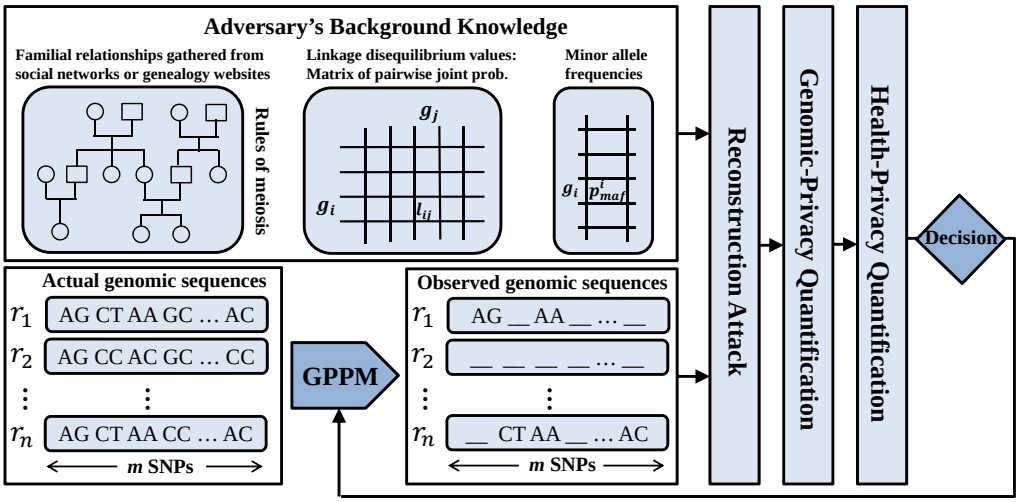


Figure 3.2: Overview of the proposed framework to quantify kin genomic privacy. Each vector $\mathbf{x}_j$ ($j \in \{1, \ldots, n\}$) includes the set of SNPs for an individual in the targeted family. Furthermore, SNP g_i of relative r_j is represented by $x_j^i \in \{0, 1, 2\}$. Once the health privacy is quantified, the family should ideally decide whether to reveal less or more of their genomic information through the genomic-privacy preserving mechanism (GPPM). The optimization of the GPPM is presented in more details in Chapter 5.

values of the SNPs of all relatives:

$$\mathbf{x} = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^m \\ x_2^1 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^m \end{bmatrix} \tag{3.1}$$

$\mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i)$ is the function representing the Mendelian inheritance probabilities (in Table 3.1), where M, F, C represent mother, father, and child, respectively. The $m \times m$ matrix L represents the pairwise linkage disequilibrium (LD) values between the SNPs in $\mathcal{G}$, that can be expressed by r^2; l_{ij} refers to the matrix entry at row i and column j. $l_{ij} > 0$ if i and j are in LD, and $l_{ij} = 0$ if these two SNPs are independent (i.e., there is no LD between them). The m-size vector $\mathbf{p}_{\text{maf}} = \left[p_{\text{maf}}^1\ p_{\text{maf}}^2\ \cdots\ p_{\text{maf}}^m\right]$ represents the minor allele probabilities (or MAF) of the SNPs in $\mathcal{G}$. Finally, note that, for any $r_k \in \mathcal{R}$, $g_i \in \mathcal{G}$, and $g_j \in \mathcal{G}$, the joint probability $P(\mathbf{X}_k^i, \mathbf{X}_k^j)$ can be derived from l_{ij}, p_{maf}^i, and p_{maf}^j.

The adversary carries out a reconstruction attack to infer the value $\mathbf{x_H} \in \mathcal{X}_H$ by relying on his background knowledge, $\mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i)$, L, $\mathbf{p}_{\text{maf}}$, and on his observation $\mathbf{x_O} \in \mathcal{X}_O$.[3] After carrying out this reconstruction attack, we evaluate genomic and health privacy of the family members based on the adversary's success and his certainty about the targeted SNPs and the predispositions to diseases they reveal.

[3] $\mathbf{x_o}$ is constructed by replacing hidden SNPs in $\mathbf{x}$ by $\perp$.

3.3.1 Adversary Model

An adversary is defined by his objective(s), attack(s), and knowledge. The objective of the adversary is to compute the values of the targeted SNPs for one or more members of a targeted family by using (i) the available genomic data of one or more family members, (ii) the familial relationships between the family members, (iii) the rules of reproduction (in Section 3.2.1), (iv) the minor allele frequencies (MAFs) of the nucleotides, and (v) the population LD values between the SNPs. We note that (i) and (ii) can be gathered online from genome-sharing websites and OSNs, and (iii), (iv), and (v) are publicly known information. Note that, in the future, the increasing possibility to accurately sequence, and to impute the actual haplotypes carried by an individual in each of the copies of the diploid genome will allow a more accurate inference of relatives' genotype than relying on population LD patterns only.

Various attacks can be launched, depending on the adversary's interest. The adversary might want to infer one particular SNP of a specific individual (targeted-SNP-targeted-relative attack) or one particular SNP of multiple relatives in the targeted family (targeted-SNP-multiple-relatives attack) by observing one or more other relatives' SNP at the same position. Furthermore, the adversary might also want to infer multiple SNPs of the same individual (multiple-SNP-targeted-relative attack) or multiple SNPs of multiple family members (multiple-SNP-multiple-relatives attack) by observing SNPs at various positions of different relatives. In this chapter, we propose an algorithm that implements the latter attack, from which any other attacks can be carried out. We formulate this attack as a statistical inference problem.

3.3.2 Inference Attack

We formulate the reconstruction attack (on determining the values of the targeted SNPs) as finding the marginal probability distributions of the random variable $\mathbf{X_H}$ representing the hidden SNPs, given the observed values $\mathbf{x_O}$, familial relationships $\mathcal{T}$, and the publicly available statistical information. We represent the marginal distribution of a SNP g_i for an individual r_j as $P(\mathbf{X}_j^i = x_j^i | \mathbf{X_O} = \mathbf{x_O})$.

These marginal probability distributions could traditionally be extracted from $P(\mathbf{X_H} = \mathbf{x_H} | \mathbf{X_O} = \mathbf{x_O}, \mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), L, \mathcal{T}, \mathbf{p}_{\text{maf}})$, which is the joint probability distribution function of of the hidden SNPs, given the available side information and the observed SNPs. Then, clearly, each marginal probability distribution could be obtained as follows:

$$P(\mathbf{X}_j^i = x_j^i | \mathbf{X_O} = \mathbf{x_O}) = \tag{3.2}$$

$$\sum_{\mathbf{x_{H'}} \in \mathcal{X}_H \setminus \mathcal{X}_j^i} P(\mathbf{X_{H'}} = \mathbf{x_{H'}}, \mathbf{X}_j^i = x_j^i | \mathbf{X_O} = \mathbf{x_O}, \mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), L, \mathcal{T}, \mathbf{p}_{\text{maf}}), \tag{3.3}$$

where $\mathbf{X_{H'}}$ is the random variable representing all SNPs except SNP g_i of relative r_j. However, the number of terms in (3.3) grows exponentially with the number of variables, making the computation infeasible considering the scale of the human genome (which includes tens of million of SNPs). In the worst case, the computation of the marginal probabilities has a complexity of $\mathrm{O}(3^{nm})$. Thus, we propose to factorize the

joint probability distribution function into products of simpler local functions, each of which depends on a subset of variables. These local functions represent the conditional dependences (due to LD and reproduction) between the different SNPs represented by $\mathbf{X}$. Then, by running the belief propagation algorithm on a factor graph, we can compute the marginal probability distributions in linear complexity (with respect to nm).

A factor graph is a bipartite graph containing two sets of nodes (corresponding to variables and factors) and edges connecting these two sets. Following [119], we form a factor graph by setting a variable node x_j^i for each random variable $\mathbf{X}_j^i$ ($g_i \in \mathcal{G}$ and $r_j \in \mathcal{R}$). We use two types of factor nodes: (i) *familial factor node*, representing the familial relationships and reproduction, and (ii) *LD factor node*, representing the LD relationships between the SNPs. We summarize the connections between the variable and factor nodes below (Figure 3.3):

- Each variable node x_j^i has its familial factor node f_j^i and they are connected. Furthermore, x_k^i ($k \neq j$) is also connected to f_j^i if k is the mother or father of j (in $\mathcal{T}$). Thus, the maximum degree of a familial factor node is 3.
- Variable nodes x_i^j and x_i^m are connected to a LD factor node $g_i^{j,m}$ if SNP g_j is in LD with SNP g_m. Since the LD relationships are pairwise between the SNPs, the degree of a LD factor node is always 2.

Given the conditional dependences given by reproduction and LD, the global distribution $P(\mathbf{X_H} = \mathbf{x_H}|\mathbf{X_O} = \mathbf{x_O}, \mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), L, \mathcal{T}, \mathbf{p}_{\text{maf}})$ can be factorized into products of several local functions, each having a subset of variables from $\mathbf{x}$ as arguments:

$$P(\mathbf{X_H} = \mathbf{x_H}|\mathbf{X_O} = \mathbf{x_O}, \mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), L, \mathcal{T}, \mathbf{p}_{\text{maf}}) =$$
$$\frac{1}{Z}\Big[\prod_{g_i \in \mathcal{G}} \prod_{r_j \in \mathcal{R}} f_j^i(x_j^i, x_{m(j)}^i, x_{f(j)}^i, \mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), \mathbf{p}_{\text{maf}})\Big] \times \Big[\prod_{r_i \in \mathcal{R}} \prod_{\substack{(j,m) \text{ s.t.} \\ l_{jm} \neq 0}} g_i^{j,m}(x_i^j, x_i^m, l_{jm})\Big], \tag{3.4}$$

where Z is the normalization constant, and $x_{m(j)}^i$, respectively $x_{f(j)}^i$, are the SNPs g_i of the mother, respectively father, of r_i (if they exist in $\mathcal{T}$).

Next, we introduce the messages between the factor and the variable nodes to compute the marginal probability distributions using belief propagation. We denote the messages from the variable nodes to the factor nodes as μ. We also denote the messages from familial factor nodes to variable nodes as λ, and from LD factor nodes to variable nodes as β. Let $X^{(\nu)} = \{x_j^{i\,(\nu)} : r_j \in \mathcal{R}, g_i \in \mathcal{G}\}$ be the collection of variables representing the values of the variable nodes at the iteration ν of the algorithm. The message $\mu_{i\to k}^{(\nu)}(x_j^{i\,(\nu)})$ denotes the probability of $x_j^{i\,(\nu)} = \ell$ ($\ell \in \{0,1,2\}$), at the ν^{th} iteration. Furthermore, $\lambda_{k\to i}^{(\nu)}(x_j^{i\,(\nu)})$ denotes the probability that $x_j^{i\,(\nu)} = \ell$, for $\ell \in \{0,1,2\}$, at the ν^{th} iteration given $x_{m(j)}^i$, $x_{f(j)}^i$, $\mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i)$, and $\mathbf{p}_{\text{maf}}$. Finally, $\beta_{k\to i}^{(\nu)}(x_j^{i\,(\nu)})$ denotes the probability that $x_j^{i\,(\nu)} = \ell$, for $\ell \in \{0,1,2\}$, at the ν^{th} iteration given the LD relationships between the SNPs.

For the clarity of presentation, we choose a simple family tree consisting of a trio (i.e., mother, father, and child) and 3 SNPs (i.e., $|\mathcal{R}| = 3$ and $|\mathcal{G}| = 3$). In Figure 3.3, we show how the trio and the SNPs are represented on a factor graph, where r_1 represents

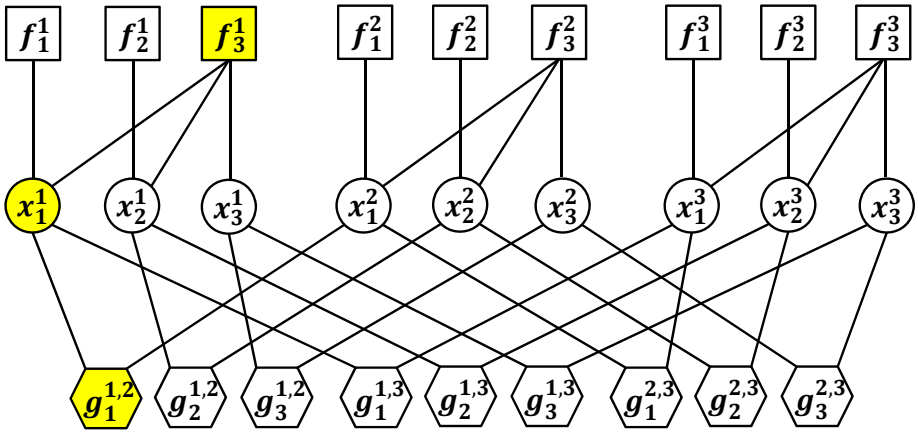


Figure 3.3: The factor graph representation of a trio (mother, father, child) and 3 SNPs per family member. The message passing described in the main text is between the nodes x_1^1, f_3^1, and $g_1^{1,2}$ highlighted in the graph.

the mother, r_2 represents the father, and r_3 represents the child. Furthermore, the 3 SNPs are g_1, g_2, and g_3. We describe the message exchange between the variable node representing the first SNP of the mother (x_1^1), the familial factor node of the child (f_3^1), and the LD factor node $g_1^{1,2}$. The belief propagation algorithm iteratively exchanges messages between the factor and the variable nodes in Figure 3.3, updating the beliefs on the values (in $\mathbf{x_H}$) of the targeted SNPs at each iteration, until convergence. We denote the variable and factor nodes x_1^1, f_3^1, and $g_1^{1,2}$ with the letters i, k, and z, respectively.

The variable nodes generate their messages (μ) and send them to their neighbors. Variable node i forms $\mu_{i\to k}^{(\nu)}(x_1^{1(\nu)})$ by multiplying all information it receives from its neighbors excluding the familial factor node k.[4] Hence, the message from variable node i to the familial factor node k at the ν^{th} iteration is given by

$$\mu_{i\to k}^{(\nu)}(x_1^{1(\nu)}) = \frac{1}{Z} \times \prod_{w\in(\sim k)} \lambda_{w\to i}^{(\nu-1)}(x_1^{1(\nu-1)}) \times \prod_{y\in\{z,g_{1,3}^1\}} \beta_{y\to i}^{(\nu-1)}(x_1^{1(\nu-1)}), \tag{3.5}$$

where Z is a normalization constant, and the notation $(\sim k)$ means all familial factor node neighbors of the variable node i, except k. This computation is repeated for every neighbor of each variable node. It is important to note that the message in (3.5) is valid if the value of x_1^1 is unobserved to the adversary. However, the value of x_1^1 can also be observed by the adversary. In this case, if $x_1^1 = \rho$ ($\rho \in \{0, 1, 2\}$), then $\mu_{i\to k}^{(\nu)}(x_1^{1(\nu)} = \rho) = 1$ and $\mu_{i\to k}^{(\nu)}(x_1^{1(\nu)}) = 0$ for other potential values of x_1^1 (regardless of the values of the messages received by the variable node i from its neighbors).

Next, the factor nodes generate their messages. The message from the familial factor node k to the variable node i at the ν^{th} iteration is formed using the principles of belief propagation as

$$\lambda_{k\to i}^{(\nu)}(x_1^{1(\nu)}) = \sum_{\{x_2^1,x_3^1\}} f_3^1(x_1^1, x_{m(1)}^1), x_{f(1)}^1) \mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), \mathbf{p}_{\text{maf}}) \prod_{y\in\{x_1^2,x_1^3\}} \mu_{y\to k}^{(\nu)}(x_1^{1(\nu)}). \tag{3.6}$$

[4]The message $\mu_{i\to z}^{(\nu)}(x_1^{1(\nu)})$ from the variable node i to the LD factor node z is constructed similarly.

Note that $f_3^1(x_1^1, x_{m(1)}^1), x_{f(1)}^1)\mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), \mathbf{p}_{\text{maf}}) \propto P(x_1^1|x_{m(1)}^1), x_{f(1)}^1)\mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), \mathbf{p}_{\text{maf}})$, and this probability is computed using Table 3.1. Furthermore, if the degree of the familial factor node is 1 for a particular SNP, then the local function corresponding to the familial factor node only depends on the MAF of the corresponding SNP. For example, the degree of f_1^1 (in Figure 3.3(c)) is 1, hence $f_1^1(x_1^1, x_{m(1)}^1), x_{f(1)}^1)\mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), \mathbf{p}_{\text{maf}}) \propto P(x_1^1|p_{\text{maf}}^1)$. The above computation must be performed for every neighbor of each familial factor node.

Similarly, the message from the LD factor node z to the variable node i at the ν^{th} iteration is formed as

$$\beta_{z\to i}^{(\nu)}(x_1^{1(\nu)}) = \sum_{x_1^2} g_1^{1,2}(x_1^1, x_1^2, l_{12}) \prod_{y\in\{x_1^2\}} \mu_{y\to k}^{(\nu)}(x_1^{1(\nu)}). \tag{3.7}$$

As before, this computation is performed for every neighbor of each LD factor node. We further note that $g_1^{1,2}(x_1^1, x_1^2, l_{1,2}) \propto P(x_1^1, x_1^2)$, which is derived from $l_{1,2}$, p_{maf}^1, and p_{maf}^2. The algorithm proceeds to the next iteration in the same way as the ν^{th} iteration.

The algorithm starts at the variable nodes. Thus, at the first iteration of the algorithm (i.e., $\nu = 1$), the variable node i sends messages to its neighboring factor nodes based on the following rules: (i) If the value of x_1^1 is hidden from the adversary, $\mu_{i\to k}^{(1)}(x_1^{1(1)}) = 1$ for all potential values of x_1^1 and, (ii) if the value of x_1^1 is observed by the adversary and $x_1^1 = \rho$ ($\rho \in \{0, 1, 2\}$), $\mu_{i\to k}^{(1)}(x_1^{1(1)} = \rho) = 1$ and $\mu_{i\to k}^{(1)}(x_1^{1(1)}) = 0$ for other potential values of x_1^1. The iterations stop when all variable nodes have converged to stable distributions. The marginal probability of each variable in $\mathcal{X}_H$ is given by multiplying all the incoming messages at each variable node representing an unobserved SNP.

3.3.3 Computational Complexity

The computational complexity of the proposed inference attack is proportional to the number of factor nodes. In our setting, there are nm familial factor nodes and a maximum of $nm(m-1)/2$ LD factor nodes. Hence, the worst-case computational complexity per iteration is $\mathrm{O}(nm^2)$. However, as each SNP is in LD with a limited number of other SNPs, the matrix L is sparse and the number of LD factor nodes grows with m rather than with $m(m-1)/2$, especially if we focus on SNPs in strong LD only. Thus, the average computational complexity per iteration is $\mathrm{O}(nm)$. Based on our experiments, we can state that the number of iterations before convergence is a small constant, between 10 and 15. Note finally that this complexity can be further reduced by using similar techniques developed for message-passing decoding of LDPC codes (e.g., working in log-domain [46]).

3.3.4 Genomic-Privacy Metrics

A crucial step towards protecting genomic privacy is to quantify the privacy loss induced by the release of genomic information. Through the inference attack, the adversary infers the targeted SNPs belonging to the members of a targeted family by using his background knowledge and observed genomic data (of the family members). The inferred information can be expressed as the posterior distribution $P(\mathbf{X_H} = \mathbf{x_H}|\mathbf{X_O} = \mathbf{x_O}, \mathcal{F}_R(\mathbf{X}_M^i, \mathbf{X}_F^i, \mathbf{X}_C^i), L, \mathcal{T}, \mathbf{p}_{\text{maf}})$. Moreover, each posterior marginal probability distri-

bution is represented as $P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O})$,[5] for all $r_j \in \mathbf{R}, g_i \in \mathbf{G}$. We propose to quantify kin genomic privacy using the following metrics: expected estimation error (incorrectness) and uncertainty.[6]

Correctness was already proposed in the context of location privacy [152]. In our scenario, correctness quantifies the adversary's success in inferring the targeted SNPs. That is, it quantifies the expected distance between the adversary's estimate on the value of a SNP, $\hat{x}_j^i$ and the true value of the corresponding SNP, x_j^i. This distance can be expressed as the expected estimation error as follows:

$$E_j^i = \sum_{\hat{x}_j^i \in \{0,1,2\}} P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O}) \left\| x_j^i - \hat{x}_j^i \right\|. \tag{3.8}$$

Privacy can also be represented as the adversary's *uncertainty* [55, 147], that is the ambiguity of $P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O})$. This uncertainty is generally considered to be maximum if the posterior distribution is uniform. This definition of uncertainty can be quantified as the (normalized) entropy of $P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O})$ as follows:

$$H_j^i = \frac{-\sum_{\hat{x}_j^i \in \{0,1,2\}} P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O}) \log P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O})}{\log(3)} := \frac{H(\mathbf{X}_j^i | \mathbf{X_O})}{\log(3)}. \tag{3.9}$$

The higher the entropy is, the higher is the uncertainty.

Finally, we propose another entropy-based metrics that quantifies the mutual dependence between the hidden genomic data that the adversary is trying to reconstruct, and the observed data. This is quantified by mutual information $I(\mathbf{X}_j^i; \mathbf{X_O}) = H(\mathbf{X}_j^i) - H(\mathbf{X}_j^i | \mathbf{X_O})$ [24]. As privacy decreases with mutual information, we propose the following (normalized) privacy metrics:

$$I_j^i = 1 - \frac{H(\mathbf{X}_j^i) - H(\mathbf{X}_j^i | \mathbf{X_O})}{H(\mathbf{X}_j^i)} = \frac{H(\mathbf{X}_j^i | \mathbf{X_O})}{H(\mathbf{X}_j^i)}. \tag{3.10}$$

The aforementioned metrics are useful for quantifying the genomic privacy of individuals. In order to quantify a more tangible privacy, we must convert these genomic-privacy metrics into health-privacy metrics. To quantify an individual's health privacy, we focus on his predisposition to different diseases. Let $\mathcal{S}_d$ be the set of SNPs that are associated with a disease d. Then, a metric quantifying the health privacy for an individual r_i regarding the disease d can be defined as follows:

$$D_i^d = \frac{1}{\sum_{k: g_k \in \mathcal{S}_d} c_k} \sum_{k \in \mathbf{S_a}} c_k G_i^k, \tag{3.11}$$

where G_i^k is the genomic privacy of SNP g_k of individual r_i, computed using (3.8), (3.9), or (3.10), and c_k is the contribution of SNP k to disease d.[7] Other health-privacy metrics based on non-linear combinations of genotypes or combinations of alleles will be defined in future work. Note that health-privacy metrics are valid at a given time, and cannot be used to evaluate future privacy provision, as genome research can change knowledge on the contribution of SNPs to diseases.

[5] We use here $\hat{x}_j^i$ to refer to the estimate of x_j^i.

[6] These metrics are not specific to the proposed inference attack; they can be used to quantify genomic privacy in general.

[7] These contributions are determined as a result of medical studies. Some SNPs might increase (or decrease) the risk for a disease more than others.

3.3.5 Genomic-Privacy Preserving Mechanism

Individuals willing to share genomic data for research or recreational purposes might be unwilling to share all their DNA sequence, and thus need to properly obfuscate the sensitive part(s) before releasing their genomic data. To do so, their DNA will go through an obfuscation process, that we call *genomic-privacy preserving mechanism* (GPPM). GPPM can be implemented using one of the following techniques: (i) hiding the SNPs, or (ii) reducing the precision or the quantity of the revealed SNPs.

Hiding all or specific SNPs can be achieved either by not releasing them or by encrypting them. Obviously, not releasing any of the SNPs would hinder genetic research, thus it is not a preferred way to protect the genomic privacy of individuals. Instead of not releasing the SNPs, the use of cryptographic algorithms to encrypt the genome is proposed. For example, Kantarcioglu *et al.* propose using homomorphic encryption on the SNPs of the individuals to perform genetic research on the encrypted SNPs [106]. However, the security of an individual's genome should be guaranteed for at least 70-100 years (i.e., during the typical lifetime of a human). As we show in this chapter, even lifelong protection is not enough, considering kin privacy implications (e.g., for offsprings). It is known that even the best of the cryptographic algorithms we use today could be broken in around 30 years. Therefore, the appropriateness of cryptographic techniques for storing and processing the genomic data has been questioned due to long-term security requirements of the genomic data.

As an alternative to the cryptographic techniques, utility can be traded off for privacy. The precision of the revealed SNPs can be reduced, for example, by revealing only one of the two alleles of a SNP. Similarly, family members' SNPs can be selectively revealed by also considering the previously revealed SNPs from the corresponding family (to keep the genomic privacy of other family members above a desired threshold): we evaluate the privacy provided by this technique in Section 3.4 by assessing the inference power of the adversary for different fractions of observed data from a targeted family. Eventually, using one of the above techniques, the GPPM will take $\mathbf{x}$ as input and output the matrix $\mathbf{x_O}$ of observed SNPs. We detail the GPPM and its optimization in Chapter 5.

3.4 Evaluation

In this section, we first evaluate the performance of the proposed inference attack, then compare the performance of the inference with and without considering the linkage disequilibrium (LD) between SNPs, and finally evaluate the entropy-based metrics with respect to the expected estimation error in quantifying the genomic privacy.

For this evaluation, we use the *CEPH/Utah Pedigree 1463* that contains the partial DNA sequences of 17 family members (4 grandparents, 2 parents, and 11 children) [59]. We note in Figure 3.4 that we only use 5 (out of 11) children for our evaluation because (i) 11 is much above the average number of children per family, and (ii) we observe that the strength of adversary's inference does not increase further (due to the children's revealed genomes) when more that 5 children's genomes are revealed. As the SNPs related to important diseases, like Alzheimer's, are not included in this dataset, we quantify health privacy in Section 3.5 by using the data collected from a genome-sharing website.

To quantify the genomic privacy of the individuals in the CEPH family, we focus on their SNPs on chromosome 1 (which is the largest chromosome). We rely on the three

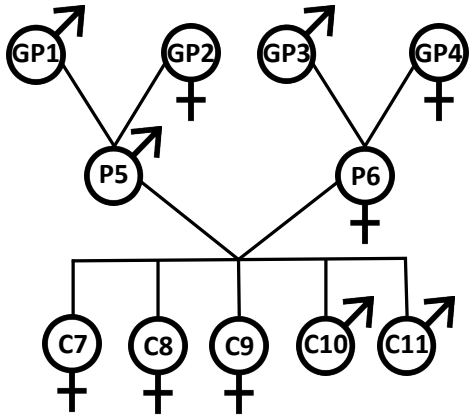


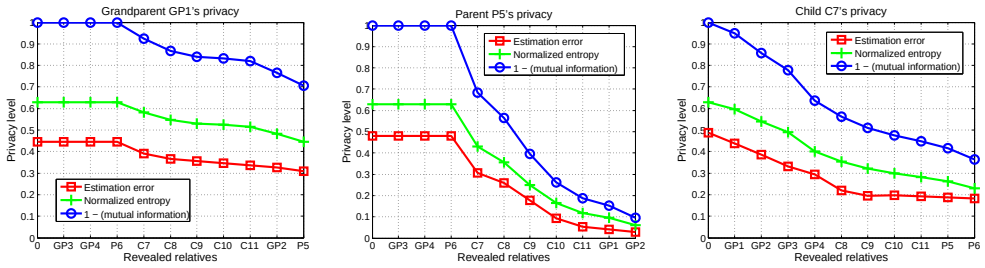
Figure 3.4: Family tree of *CEPH/Utah Pedigree 1463* consisting of the 11 family members that were considered. The symbols ♂ and ♀ represent the male and female family members, respectively.

metrics introduced in Section 3.3.4. That is, we compute the genomic privacy of each family member using the expected estimation error in (3.8), the (normalized) entropy in (3.9), and the (normalized) mutual information in (3.10) on the targeted SNPs, and we average the result based on the number of targeted SNPs for each individual. We rely on the L_1 norm to measure the distance between two SNP values in (3.8).

First, we assume that the adversary targets one family member and tries to infer his/her SNPs by using the published SNPs of other family members without considering the LD between the SNPs. We select an individual from the CEPH family and denote him as the target individual. We construct $\mathcal{G}$, the set of SNPs that we consider for evaluation, from 80k SNPs on chromosome 1. Thus, the targeted SNPs are the 80k SNPs of the target individual. Furthermore, we gradually fill the matrix of $\mathbf{x_O}$ of observed SNPs with the values of the 80k SNPs of other family members. That is, we sequentially reveal 80k SNPs (in $\mathcal{G}$) of all family members (excluding the target individual) beginning with the most distant family members from the target individual (in terms of number of hops in Figure 3.4) and we keep revealing relatives until we reach his/her closest family members.[8]

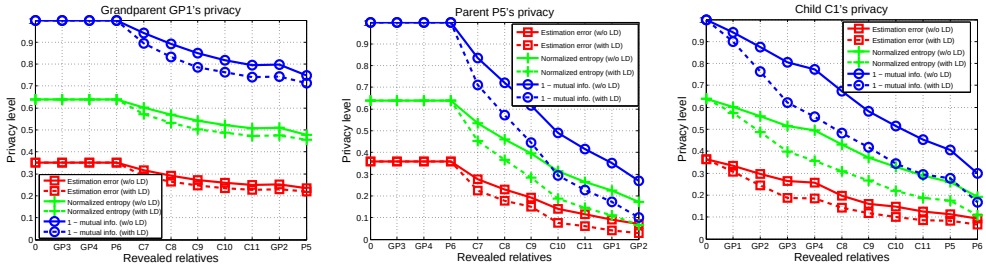
In Figure 3.5 we show the evolution of the genomic privacy of three target individuals from the CEPH family (in Figure 3.4): (i) grandparent (GP1), (ii) parent (P5), and (iii) child (C7). We note that all entropy-based metrics for each target individual start from the same values. We also observe that the parent's and the child's genomic privacy decreases considerably more than the grandparent's (the adversary's error for the grandparent's genome does not go below 0.3). Moreover, the observation of GP3, GP4 and P6's genomes has no effect on GP1 and P5's privacy as their genomes are independent (if no other relatives' genomes are observed). We observe in Figure 3.5(a) that the grandparent's genomic privacy is mostly affected by the SNPs of the first revealed children (C7, C8), and also by those of his spouse and his child (P5). We also observe (in Figure 3.5(b)) that, by revealing all family members' SNPs (expect P5), the adversary can almost reach an estimation error of 0. The target parent's genomic privacy significantly

[8]The exact sequence of the family members (whose SNPs are revealed) is indicated for each evaluation.



(a) (b) (c)

Figure 3.5: Evolution of the genomic privacy of the (a) grandparent (GP1), (b) parent (P5), and (c) child (C7). We reveal all the 80k SNPs on chromosome 1 of other family members starting from the most distant family members of the target individual (in terms of number of hops to the target individual in Figure 3.4); the x-axis represents the disclosure sequence. We note that $x = 0$ represents the prior distribution, when no genomic data is observed by the adversary.



(a) (b) (c)

Figure 3.6: Evolution of the genomic privacy of the (a) grandparent (GP1), (b) parent (P5), and (c) child (C7), with and without considering LD. For each family member, we reveal 50 randomly picked SNPs (among 100 SNPs in $\mathbf{S}$), starting from the most distant family members, and the x-axis represents the exact sequence of this disclosure. Note that $x = 0$ represents the prior distribution, when no genomic data is revealed.

decreases only with the observation of his children's and spouse's SNPs. Finally, we observe in Figure 3.5(c) that C7's genomic privacy decreases smoothly with the observation of his grandparents' SNPs, and then of his siblings'. We also observe a slight decrease of privacy once the parents' SNPs (P5 and P6) are also revealed, but the observation of parents (after the other children) does not have a significant effect on the adversary's error. It is important to note that the importance of a family member for the inference power of the adversary also depends on the sequence at which his/her SNPs are revealed in Figure 3.5. For example, in Figure 3.5(c), if the SNPs of the parents (P5 and P6) of the target child (C7) were revealed before her siblings (C8-C11), then the observation of her parents would reduce the genomic privacy of the target child more than her siblings (but the final genomic privacy would not change).

Next, we include the LD relationships and observe the change in the inference power of the adversary using the LD values. We construct $\mathcal{G}$ from 100 SNPs on chromosome 1. Among these 100 SNPs, each SNP is in LD with 5 other SNPs on average. Furthermore,

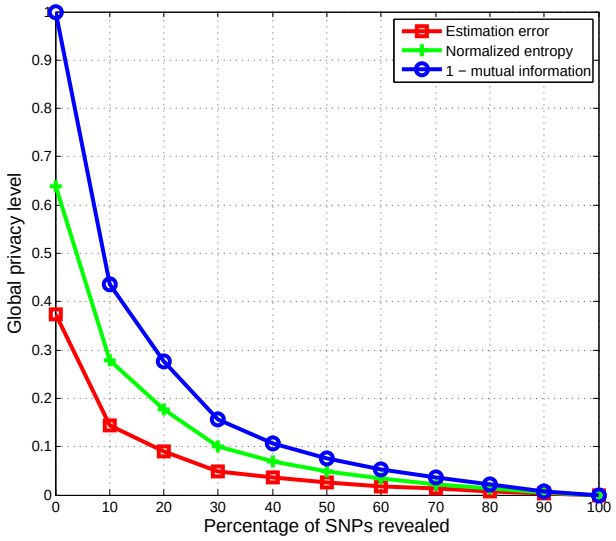


Figure 3.7: Evolution of the global privacy for the whole family by gradually revealing 10% of family's SNPs.

the strength of the LD (r^2 value in Section 3.2.1) uniformly varies between 0.5 and 1 (where $r^2 = 1$ represents the strongest LD relationship, as discussed before). We note that we only use 100 SNPs for this study as the LD values are not yet completely defined over all SNPs, and the definition of such values is still an ongoing research. As before, we define a target individual from the CEPH family, and sequentially reveal other family members' SNPs to observe the decrease in the genomic privacy of the target individual. We observe that individuals sometimes reveal different parts of their genomes (e.g., different sets of SNPs) on the Internet. Thus, we assume that for each family member (except for the target individual), the adversary observes only 50 random SNPs from $\mathcal{G}$ (instead of all the SNPs in $\mathcal{G}$), and these sets of observed SNPs are different for each family member. In Figure 3.6, we show the evolution of genomic privacy of three target individuals when the adversary also uses the LD values. We observe that LD decreases genomic privacy, especially when few individuals' genomes are revealed. As more family member's genomes are observed, LD has less impact on the genomic privacy.

We also evaluate the inference power of the adversary to infer multiple SNPs among all family members, given a subset of SNPs belonging to some family members, and also considering the LD between SNPs. That is, we evaluate the inference power of the adversary for different fractions of observed data for the family members. Using the same set of 100 SNPs, we construct $\mathbf{x_o}$ from $(\kappa \times 100 \times n)$ SNPs, randomly selected from all family members, where n is the number of family members in the family tree ($n = 11$ for this scenario), and $0 \leq \kappa \leq 1$. We assume that the SNPs that are not in $\mathbf{x_o}$ are hidden from the adversary (i.e., in $\mathbf{x_H}$), and we observe the inference power of the adversary for all the SNPs in $\mathbf{x}$, for different increasing values of κ. In Figure 3.7, we observe an exponential decrease in the global genomic privacy (privacy of all family members), showing that the observation of a small portion of the family's SNPs can have a huge impact on genomic privacy. The estimation error is decreased by around 3 by observing only the first 10% of the SNPs.

3.5 Cross-Website Attacks

In order to show that the proposed inference attack threatens not only the Lacks family, but potentially *all* families, we collected publicly available data from a genome-sharing website and familial relationships from OSNs, and evaluated the decrease in genomic and health privacy of people due to the observation of their relatives' genomic data.

3.5.1 OpenSNP and Facebook

We gathered individuals' genomic data from OpenSNP [2], a website on which people can publicly share sets of SNPs. Then, we identified the owners of some gathered genomic profiles by using their names and sometimes profile pictures. Among these identified individuals, we managed to find family relationships of 6 of them (who publicly reveal the names of some of their relatives) on Facebook.[9] We expect this number to increase in the future, as more health-related OSNs (which let people share their genomic profiles, such as 23andMe [8]) emerge. Furthermore, we anticipate that the current widely used health-related OSNs (e.g., PatientsLikeMe [13]) will let users upload and share their genomic data. We identified 29 target individuals from 6 different families, whose genomic data can be inferred using the observed SNPs of the identified individuals.

We focus on 2 individuals I_1 and I_2 out of these 6 identified individuals and evaluate the genomic and health privacy for their family members. We observed that both I_1 and I_2 publicly disclosed around 1 million of their SNPs. Furthermore, we identified the names of (i) 1 mother, 2 sons, 2 daughters, 1 grandchild, 1 aunt, 2 nieces, and 1 nephew of I_1, and (ii) 1 sibling, 1 aunt, 1 uncle, and 6 cousins of I_2 on Facebook. We compute the genomic and health privacy of these target individuals using the (normalized) entropy in (3.9) on the targeted SNPs, and normalize the result based on the number of targeted SNPs for each individual. We do not use the expected estimation error in (3.8), as we do not have the ground truth for the genomes of the target individuals. Thus, privacy is quantified as the uncertainty of the adversary in this section.

To quantify the genomic privacy of the target individuals (i.e., family members of I_1 and I_2), we first construct $\mathcal{G}$ from all SNPs on chromosome 1 (from the observed genomes of I_1 and I_2). The set of observed SNPs includes the observed SNPs of I_1 (respectively I_2) for the inference of family members of I_1 (respectively I_2). The set of targeted SNPs includes 77k SNPs for I_1's family and 79k for I_2's family for each evaluation. In Figure 3.8, we show the decrease in the genomic privacy for different family members of I_1 (aunt, niece/nephew, grandchild, mother, child) and I_2 (cousin, aunt/uncle, sibling) as a result of our proposed inference attack, first without considering the LD dependencies (similarly to previous section). We observe that, as expected, the decrease in the genomic privacy of close family members is significantly higher than that of more distant family members. However, as we have seen in Section 3.4 and we will show in the next Subsection, the observation of one (or more) additional family member(s) has often much more impact on the target's privacy than the observation of only one relative.

In Figure 3.9, we display the decrease of genomic privacy with respect to 100 SNPs of chromosome 1.[10] We first show the different privacy levels by using all 100 SNPs of the

[9] According to [79], around 12% of Facebook users publicly share at least one family member on their profiles.

[10] We consider only 100 SNPs here for the same reason as in Section 3.4.

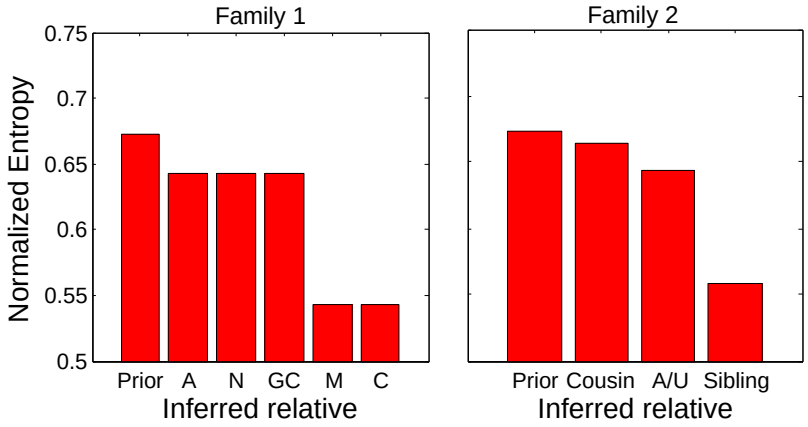


Figure 3.8: Attacker's uncertainty about all SNP values on chromosome 1 for two different families gathered on Facebook, without using LD. A stands for aunt, N for niece/nephew, GC for grandchild, M for mother, C for child, U for uncle. Same notations are used in Figures 3.9 and 3.10.

.

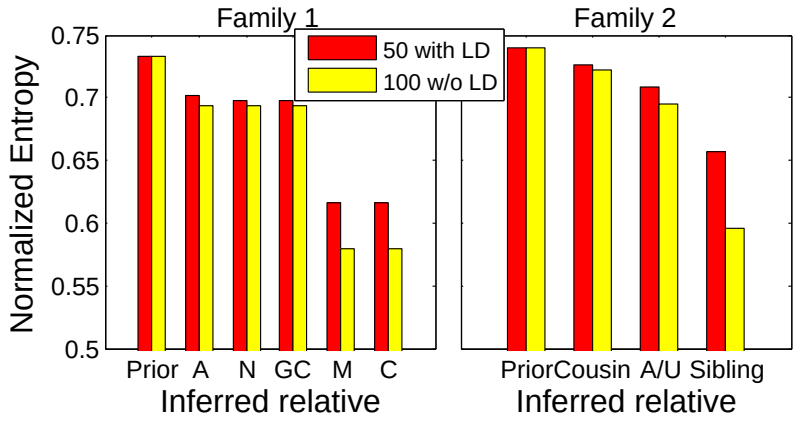


Figure 3.9: Attacker's uncertainty about values of 100 SNPs on chromosome 1 for two families gathered on Facebook, by observing (i) all 100 SNPs of the relative that reveals his/her genome, and (ii) only 50 SNPs but using LD.

observed relative (i.e., I_1 or I_2), and then show the same by using only 50 SNPs of the observed relative and LD values. We note that the use of LD decreases privacy slightly more for the first familyt than for the second family. This is because we randomly picked 50 different SNPs for both families, and those picked in the second family had weaker LD relationships with other SNPs. We finally observe that the difference between the two observation cases (50 SNPs with LD and 100 SNPs without LD) is higher for close relatives (mother, child, or sibling) than for others.

We also evaluate the health privacy of the family members of I_1 and I_2 considering their predispositions to various diseases. We first noticed that almost all important SNPs for privacy-sensitive diseases affected by genomic factors, like Alzheimer's, ischemic heart

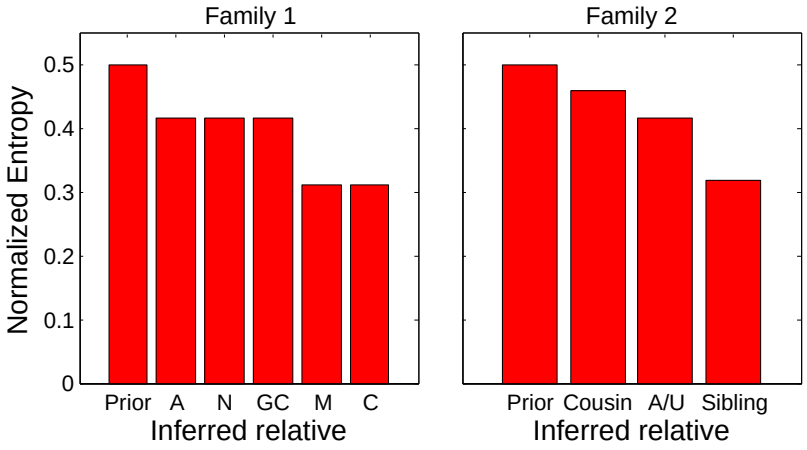


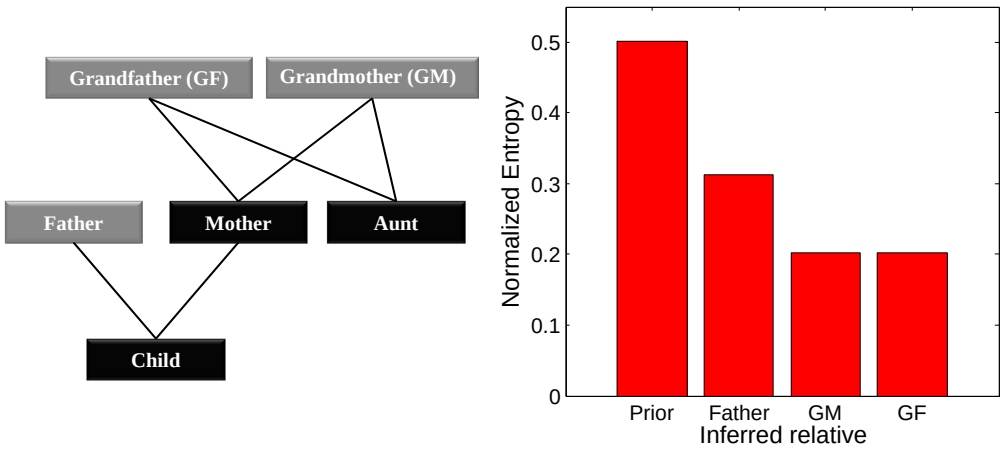
Figure 3.10: Adversary's uncertainty about Alzheimer's disease predisposition for two families gathered on Facebook.

disease, or macular degeneration, were revealed by I_1 and I_2. Due to lack of space, we focus on Alzheimer's as it is one of the most important diseases that are mainly attributable to genetic factors. Having two ApoE4 alleles (in SNPs rs7412 and rs429358 located on chromosome 19) dramatically increases an individual's probability of having Alzheimer's by the age of 80. Thus, the contents of these two SNPs carry privacy-sensitive information for individuals. We use the metrics in (3.11) to quantify the health privacy of family members for Alzheimer's disease. We assign equal weights to both associated SNPs (as their combination determines the predisposition to Alzheimer's disease). In Figure 3.10, we show the attacker's uncertainty about the predisposition to Alzheimer's disease for the family members of I_1 and I_2. We notice a decrease of around 0.2 (from 0.5 to 0.3) in uncertainty between close relatives. Clearly, the knowledge of the SNPs of more relatives would further worsen the situation.

3.5.2 OpenSNP and Genealogy Website

We gathered individuals' genomic data from OpenSNP.org here too. Then, we matched 47 OpenSNP profiles (who provided their full names) with profiles on genealogy websites (that included familial information), clearly showing us the scale of the threat. We noticed that three of the individuals identified on OpenSNP were associated to the same family (which is hereafter referred to as the target family). Furthermore, from the family tree, we obtained the names of 3 *target individuals* (only considering ancestors up to the grandparents of youngest identified individual revealing his SNPs) in the same family, as shown in Figure 3.11(a). We emphasize again that these 3 target individuals did not publicly share any genomic data and that they would possibly be against such a disclosure. We compute the health privacy of these 3 targets about their predispositions to Alzheimer's disease by using the same SNPs as in Subsection 3.5.1.

In Figure 3.11(b), we show the attacker's uncertainty about the predisposition to Alzheimer's disease for the target individuals. We notice a decrease of 40% for the father, and of 60% for both the grandmother and the grandfather, compared to their initial privacy (prior, without any information about the genomes of their relatives).



(a) (b)

Figure 3.11: Quantification of health privacy for one family (with three relatives revealing their SNPs) found in a genealogy webiste. (a) The family tree of the target family, where black means that the genomic data of the family member is publicly available on OpenSNP, and grey means it is not. (b) Adversary's uncertainty about Alzheimer's disease predisposition of family members whose genomes are not publicly available.

This demonstrates that the more genomic data is shared, the less privacy is provided to the family members.

3.6 Related Work

Stajano *et al.* were among the first to raise the issue of kin privacy in genomics and to suggest discussing questions such as; Should you be allowed to disclose your genome if other relatives do not want to? [154]. Cassa *et al.* provide a framework for measuring the risks to siblings of someone who reveals his SNPs [44]. They show that the inference error is substantially reduced when the sibling's SNPs are known, compared to when only the population frequencies are used. We push this work further, by considering any kind of family members, and LD relationship between SNPs, by proposing and evaluating different privacy metrics, and by presenting a real attack scenario using publicly available data. Our generic framework considers any observation of a family's SNPs, and the adversary's background knowledge. Re-identification attacks have also been proposed in the literature. Homer *et al.* [86] prove that de-identification is an ineffective way to protect the privacy of participants in genome-wide association studies, which is also supported by other works [167, 72, 175]. More recently, Gymrek *et al.* show how they identified DNAs of several individuals (and their families) who participated in scientific studies [81]. Finally, Sweeney et al. de-anonymized participants of the Personal Genome Project by linking their demographic information to public records such as voter lists [158].

Several algorithms for inference on graphical models have been proposed in the context of pedigree analysis. Exact inference techniques on Bayesian networks are used in order to map disease genes and construct genetic maps [67, 124, 105]. Monte Carlo methods (Gibbs sampling) were also proved to be efficient for genetic analyses in the

case of complex pedigrees [100, 160, 148]. All these methods aim to infer specific genotypes given phenotypes (like diseases). Another paper relies on Gibbs sampling in order to infer haplotypes (used in association studies) from genotype data [111]. Genotype imputation [125] is another technique used by geneticists to complete missing SNPs based upon given genotyped data. A similar approach has recently been used to infer high-density genotypes in pedigrees, by relying notably on low-resolution genotypes and identity-by-descent regions of the genome [42]. None of these contributions addresses privacy.

In contrast with these contributions, in this chapter, we propose a novel and efficient inference attack in order to reconstruct genomic data of individuals given observed genomic data of their family members and special characteristics of genomic data. Furthermore, we quantify the genomic privacy of individuals as a result of this attack using different metrics, and show the real threat by using the data collected from different websites and OSNs.

3.7 Summary

In this chapter, we have formalized the problem of quantifying kin genomic privacy. To quantify privacy, we mimic the reconstruction attack of an adversary who tries to infer hidden data based on the genomic data he gets access to and some public background knowledge. We propose an efficient inference attack that relies upon probabilistic graphical models and belief propagation. Our inference algorithm provides us with the exact posterior marginal distributions of the random variables representing the genomic data when linkage disequilibrium is not taken into account. When linkage disequilibrium is included in the graphical model, the posterior distributions are only estimations but which are, in practice, very close to the exact values. We introduce different genomic-privacy metrics that express the (in)correctness and (un)certainty of the adversary's estimation of genomic data. In order to get more tangible metrics, we also suggest to quantify health privacy, that is the privacy of individuals regarding the predisposition to certain diseases. We evaluate our approach and metrics on real genomic data gathered from eleven close relatives. Furthermore, we demonstrate the extent of the threat by matching users sharing their genomic data online with OSNs profiles where these users also reveal (some of) their relatives.

Our results notably show that, by disclosing only 10% of its genomic data, a family loses more than 50% of its global genomic privacy. This is an effect of both genetic dependencies between relatives and dependencies within each genome via linkage disequilibrium. We also show, in our cross-website attack, that the privacy regarding predisposition to Alzheimer's disease can drop by 40% due to the disclosure of a first-degree relative's genomic data related to this disease. The privacy situation even worsens if more than one relative reveals his genomic data.

Chapter 4

Non-cooperative Behavior in Genomic Privacy

4.1 Introduction

The decreasing cost in genome sequencing has dramatically increased the availability and use of genomic data in many domains such as healthcare, research, law enforcement, and recreational genomics. This availability raises many questions regarding the management (storage, sharing, etc.) and, ultimately, the privacy of genomic data. For instance, thousands of individuals are already sharing their genomic data online, either anonymously[1] or with their real identity (e.g., on OpenSNP.org), showing the willingness of some people to disclose their genomic data. In addition to this, all individuals whose DNA has been sequenced have to carefully manage their genomic data. Some may decide to store it on personal devices, others on external (potentially untrusted) servers. In both cases, guaranteeing security and privacy has a non-negligible cost. Schematically, in this work, we consider that an individual whose DNA has been sequenced must make decisions on (i) whether to share his genomic data or not, and (ii) how much to invest in securing the storage of this data. In this chapter, we analyze the strategic behaviors of members of the same family in a genomic-privacy context by using a game-theoretic approach. Game theory has been shown to be very useful for analyzing the behavior of strategic agents in information security settings [26]. In particular, interdependent security (IDS) games have been proposed for scenarios where agents make decisions that affect not only their own security risks but also those of others [122]. As we have seen in the previous chapter, the genomic data of close relatives is highly correlated, thus leading to interdependent privacy (IDP) risks. Following the IDS works, we define two IDP games between family members with different perceived benefits, costs and privacy levels: (i) the storage-security game where family members have to decide on the security investment to protect their genomic data, and (ii) the disclosure game where relatives have to choose whether to disclose or not their genomes. First, we study the interplay between two family members, who are selfish or (partially) altruistic. With

[1] Anonymization has been proven to not be an effective technique for protecting identities of the data owners in the genomic context [81, 158].

the two-player setting, we derive a closed-form expression to quantify genomic privacy of any individual given one of his relatives' genome, and compute different closed-form Nash equilibria for the two games we study. This closed-form expression enables us to compute the genomic privacy of individuals three orders of magnitude faster than with the belief propagation method (proposed in Chapter 3). Furthermore, we consider some altruistic behavior within a family. Then, we extend the two-player game to consider n family members who decide whether to secure or disclose their genomes. To efficiently compute the Nash equilibrium of the n-player game, we make use of multi-agent influence diagrams (MAIDs), an extension of Bayesian networks that enables us to include decision and utility variables. With this approach we can significantly reduce computational complexity with respect to a classic extensive-form game. Note that, compared to IDS games that rely upon theoretical models of interdependence, the indirect risks in the IDP games come from the actual familial correlations evidenced by genetics. Furthermore, we quantify genomic-privacy losses with real genomic data, which provides very tangible results.

Our results show that, if the discrepancy between the players' perceptions of the genome-sharing benefits is too high, they follow opposite strategies, creating externalities. These misaligned incentives lead to inefficient equilibria that result in a lower familial utility than when incentives are aligned. Our analysis also shows that, surprisingly, altruism does not always lead to a more efficient outcome in a genomic privacy game. Yet, such suboptimal equilibrium can be avoided if the players coordinate.

4.2 Model

Users We consider a set of n users from a family whose genotypes are sequenced. We assume that all users have the same number and set $\mathcal{G}$ of SNPs sequenced and stored on their devices. Users have to make choices regarding the investment on securing their genomic data and the sharing of this data (e.g., to help research). A user might prefer storing his genomic data on a personal, and possibly mobile, device. For instance, as suggested in [52], there are various advantages to keep a person's genome on a smartphone. It is portable, highly personal, and nowadays has very good computational and storage capabilities. Unfortunately, the number of pieces of malware in current smartphones has exploded over the last few years [155], and keeping a mobile device secure yields non-negligible costs. Alternatively, a user could decide to outsource the storage of his genomic data to an untrusted third party. Second, a user might want to publicly share his SNPs, essentially because his perceived benefits outweigh the perceived cost (loss) for his genomic privacy.[2] We assume such users typically do not invest in securing their genomes on their personal devices, as they are already publicly disclosed.

Adversary The adversary's goal is to collect and infer genomic data. His motivations for gathering individuals' genotypes can be multiple. For instance, the adversary could sell the collected genomic data to life or health insurances that would then use it to genetically discriminate potential insurees. As usually assumed in IDS games, the adversary is considered to be an exogenous, persistent threat [122]. Thus, we do not model him as a strategic agent, but rather as a probability $h(\cdot)$ of a successful breach on the targeted

[2] See, e.g., [4] to understand users' motivations for and fears about genome sharing.

system. If a user decides to publicly disclose his SNPs online, the probability of a breach is 1 as the adversary can easily access the data.

4.3 Genomic Privacy Games

The genomes of close family members are highly correlated. Thus, the individuals' behaviors regarding their genomic data will not only affect their personal genomic privacy, but also those of their relatives, thus leading to interdependent risks. Game theory enables us to model the interplay between users with dependent payoffs, with potentially conflicting interests, and it enables us to predict their behaviors. We define two interdependent privacy games between family members: (i) the (storage-)security game G_s, and the disclosure game G_d. Both G_s and G_d are defined as a triplet ($\mathcal{P}$, $\mathcal{S}$, $\mathcal{U}$), where $\mathcal{P}$ is the set of players, $\mathcal{S}$ is the set of strategies, and $\mathcal{U}$ is the set of payoff functions.

• **Players:** The set of players $\mathcal{P} = \{P_1, ..., P_n\}$ corresponds to the set of n family members having their genomes sequenced, in both games G_s and G_d.

• **Strategies:** In the game G_s, for each player P_i, the strategy $s_i \in \mathcal{S}$ represents the security investment for the storage of his genomic data. As differences between discrete and continuous models of investment appear only in some boundary cases [122, 77], we consider here the discrete model, i.e. $s_i \in \{0, 1\}$. $s_i = 1$ means "to invest in securing his own device", and $s_i = 0$ means "to not invest", with his data on his device or outsourced to an untrusted third party (that could be itself attacked). The strategy profile is then defined as $\mathbf{s} = [s_1, \cdots, s_n]^T$. In the game G_d, the strategy is represented by the decision d_i to publicly share P_i's SNPs (e.g., on OpenSNP.org) or not. As the majority of genome-sharing people currently choose to disclose nothing or their whole set of SNPs, we consider here a discrete binary model, i.e. $d_i \in \{0, 1\}$ (0 meaning "no disclosure" and 1 "full disclosure"). Note that we study is studied in detail a finer granularity of disclosure in a cooperative context in Chapter 5. A player will choose $d_i = 1$ if and only if he perceives more utility by sharing than by protecting. The strategy profile is then represented by $\mathbf{d} = [d_1, \cdots, d_n]^T$.

• **Payoff Functions:** The utility of a player is, by definition, equal to the benefit minus the cost. In our setting, the first term of the benefit, b_i^g, represents the fact that a user's genome is sequenced and available for various benefits (e.g., personalized medicine). This generic benefit can be added to the benefit b_i^d that player P_i obtains by disclosing his genomic data online in game G_d. The cost comprises the (unit) cost of a security investment for protecting his genome, c_i, and the potential loss l_i of genomic privacy. [3] The cost c_i can represent the OS updates that can lead to a non-negligible cost (renewal of the equipment) once a device becomes too old to support them.

In our genomic context, the privacy loss l_i can be precisely quantified by relying upon the expected estimation error E_i^j defined in (3.8) in Chapter 3. Giving the same weight to each SNP, we then get

$$E_i = \frac{1}{|\mathcal{G}|} \sum_{k:g_k \in \mathcal{G}} \sum_{\hat{x}_i^k \in \{0,1,2\}} P(\mathbf{X}_i^k = \hat{x}_i^k | \mathbf{X_O} = \mathbf{x_O}) \left\| x_i^k - \hat{x}_i^k \right\|_1, \tag{4.1}$$

where $\mathbf{X_O}$ represents the SNPs observed by the adversary. This observation depends on the strategies of the players in G_s and G_d. We will denote $E_{i,0}$ to be the genomic privacy

[3] Note that an expected monetary loss would be expressed as a non-decreasing function of l_i. This is left for future work.

when no SNP is observed, i.e. when $P(\mathbf{X}_i^j = \hat{x}_i^j | \mathbf{X_O} = \mathbf{x_O}) = P(\mathbf{X}_i^j = \hat{x}_i^j)$. This initial (prior) privacy level is computed by using the minor allele frequencies (MAFs) $\mathbf{p}_{\text{maf}}$ given by population statistics. In general, as $\mathbf{x_O}$ depends on the strategy profile $\mathbf{s}$ (respectively $\mathbf{d}$), E_i will be a function of $\mathbf{s}$ (respectively $\mathbf{d}$) in game G_s (respectively G_d). As assumed in several IDS games (e.g., [121]), the probability of successful breach is set to zero when a player invests in security, i.e. $h(s_i = 1) = 0$. Otherwise, $h(s_i = 0) = p_a$ with $0 < p_a \leq 1$. For the game G_d, $h(d_i = 1) = 1$ as discussed in Section 4.2, and $h(d_i = 0) = 0$.[4] In our genomic privacy game, contrarily to IDS games, the interdependence lies in the genomic-privacy loss and not in the breach probability $h(\cdot)$. The genomic-privacy loss l_i is defined as $E_{i,0} - E_i(\cdot)$, where $E_i(\cdot)$ is a function of the strategy profile $\mathbf{s} = (s_i, \mathbf{s}_{-i})$ or $\mathbf{d} = (d_i, \mathbf{d}_{-i})$. Note that the risk is non-additive: Either the adversary manages to know the player's genome directly (and the genomic privacy drops to zero), in which case the knowledge of another genome does not bring any extra information, or the adversary cannot get access to the player's genome and then there is only an indirect privacy loss. Defining $h(\mathbf{s}_{-i})$ as the probability of successful breaches into a subset of players' devices (other than P_i), the payoff function of a player P_i in G_s is

$$u_i(s_i, \mathbf{s}_{-i}) = b_i^g - \left(s_i c_i + h(s_i) E_{i,0} + (1 - h(s_i))\, h(\mathbf{s}_{-i})\, (E_{i,0} - E_i(\mathbf{s}_{-i}))\right), \tag{4.2}$$

and his payoff in the game G_d is

$$u_i(d_i, \mathbf{d}_{-i}) = b_i^g + d_i b_i^d - \left((1 - d_i) c_i + d_i E_{i,0} + (1 - d_i)\, (E_{i,0} - E_i(\mathbf{d}_{-i}))\right).^5 \tag{4.3}$$

• **Social Welfare:** We define the *social welfare* function as the sum of the payoffs of all players: $U(\mathbf{s}) = \sum_{P_i \in \mathcal{P}} u_i(\mathbf{s})$ for G_s, and $U(\mathbf{d}) = \sum_{P_i \in \mathcal{P}} u_i(\mathbf{d})$ for G_d.

• **Altruism:** Finally, we consider that family members are usually not purely selfish regarding their relatives, meaning that some altruistic factors can play a role in their decisions. Following an idea introduced in [132] for social networks, we define a familial factor $\alpha \in [0, 1]$ that conveys the fact that relatives tend to be altruistic among themselves. We raise this factor to the power $k(P_i, P_j) \in \mathbb{N}^*$ that represents the degree of kinship between players P_i and P_j.[6] $\alpha = 0$ means that players are purely selfish, whereas $\alpha = 1$ implies that they are fully altruistic with their whole family. For instance, in G_s, the altruistic player P_i will maximize the following utility (instead of (4.2)):

$$u_i^a(s_i, \mathbf{s}_{-i}) = u_i(s_i, \mathbf{s}_{-i}) + \sum_{j : P_j \in \mathcal{P}, j \neq i} \alpha^{k(P_i, P_j)} u_j(s_i, \mathbf{s}_{-i}). \tag{4.4}$$

4.4 Two-Player Games

In this section, we study the interplay between two relatives who are first purely selfish, and then who are partially altruistic depending on their degree of kinship.

[4] We assume that a player who does not share his SNPs will always invest in security in G_d.

[5] In the following sections, we will use the more concise notation $E_{i|-i}$ expressing the genomic privacy of P_i given a subset (that depends on $\mathbf{s}_{-i}$ or $\mathbf{d}_{-i}$) of other players' SNPs.

[6] $k = 1$ for first-degree relatives such as parent, child, sibling, $k = 2$ for second-degree relatives such as grandparent, grandchild, uncle, aunt, niece, and so on.

Table 4.1: Normal form of the two-player game G_s.

$P_1 \backslash P_2$	$s_2 = 1$	$s_2 = 0$
$s_1 = 1$	$(b_1^g - c_1, b_2^g - c_2)$	$(b_1^g - c_1 - p_a(E_{1,0} - E_{1\|2}), b_2^g - p_a E_{2,0})$
$s_1 = 0$	$(b_1^g - p_a E_{1,0},$ $b_2^g - c_2 - p_a(E_{2,0} - E_{2\|1})$	$(b_1^g - p_a E_{1,0} - (1-p_a)p_a(E_{1,0} - E_{1\|2}),$ $b_2^g - p_a E_{2,0} - (1-p_a)p_a(E_{2,0} - E_{2\|1}))$

4.4.1 Selfish Players

We start our analysis with the game G_s whose strategic representation is shown in Table 4.1. Assuming the cost of security investment to be the same for all players, i.e., $c_1 = c_2 = c$, we characterize all Nash equilibria.

Lemma 4.1. *For any value $c \in [0, \infty[$, there exists at least one pure Nash equilibrium (NE) in G_s. More precisely, the NE are defined by the best responses (s_1^*, s_2^*):*

$$(s_1^*, s_2^*) = \begin{cases} (1,1) & \text{if } c < \min(t_1, t_2) \\ (1,1), mNE & \text{if } \min(t_1, t_2) < c < \max(t_1, t_2) \\ (1,1), (0,0) & \text{if } \max(t_1, t_2) < c < p_a \min(t_1^0, t_2^0) \\ (0,0), mNE & \text{if } p_a \min(t_1^0, t_2^0) < c < p_a \max(t_1^0, t_2^0) \\ (0,0) & \text{if } c > p_a \max(t_1^0, t_2^0) \end{cases} \tag{4.5}$$

if $\max(t_1, t_2) < p_a \min(t_1^0, t_2^0)$, where $t_i = p_a E_{i,0} - p_a^2(E_{i,0} - E_{i|j})$, $t_i^0 = E_{i,0}$, and mNE is a mixed-strategy Nash equilibrium. If $\max(t_1, t_2) > p_a \min(t_1^0, t_2^0)$, the third case NE in (4.5) become $(0,1)$ if $t_1^0 < t_2^0$ and $(1,0)$ if $t_1^0 > t_2^0$, and $\max(t_1, t_2)$ and $p_a \min(t_1^0, t_2^0)$ are swapped in the inequality bounds on c.

The NE can be derived from the best responses of each player that are quite straightforward by taking a close look at Table 4.1. Figure 4.1 depicts how the NE evolves for different values of c. In order to obtain closed-formed Nash equilibria, we must analytically express the genomic privacy levels $E_{i,0}$ and $E_{i|j}$. In Chapter 3, we have shown that, in the general case, belief propagation on factor graphs can be used to compute the posterior marginal probability $P(\mathbf{X}_i^k | \mathbf{X_O})$ given some observed genomic data, and thus to quantify genomic privacy. In the following, we show that, if only two members are involved in the game, and no other familial genomic data is observed, we can derive a closed-form expression for $P(\mathbf{X}_i^k | \mathbf{X_O})$, thus for $E_{i,0}$ and $E_{i|j}$. As we assume that all players have the same set of SNPs $\mathcal{G}$ sequenced and potentially exposed, and that the adversary can access the whole sequence of SNPs or nothing (as he either successfully breaches the system or not), linkage disequilibrium (correlations) between the SNPs

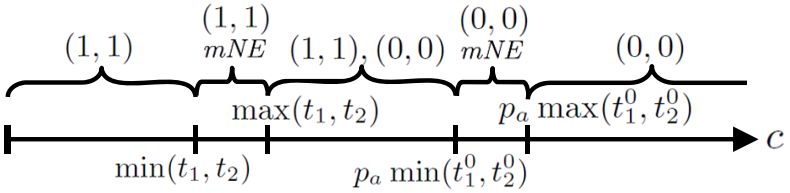


Figure 4.1: Dependence of the NE of game G_s with respect to the investment cost c.

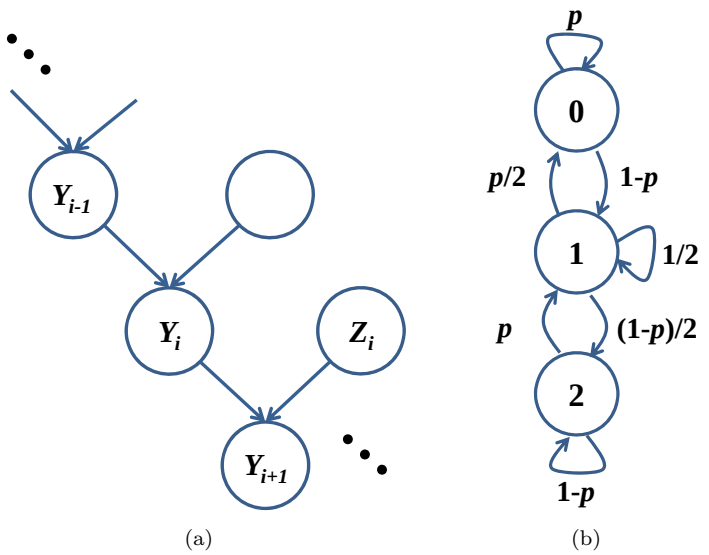


Figure 4.2: (a) Bayesian network representation of a three-generation family, and (b) Markov chain representing the probabilities of moving from one SNP value (state) to another from generation i to $i+1$ or $i-1$. Probability p is the major allele frequency of the given SNP.

does not help the adversary, and thus is not used in the computation of genomic privacy. Thus, when we want to compute the privacy at SNP g_k of player P_i, we consider only the observation at the same SNP g_k of player P_j. Each SNP can then be considered independently of other SNPs. In the following two lemmas, we focus on a single SNP, thus drop the subscript/superscript k. Assuming $\mathbf{Y}_i$ is the random variable representing a SNP of an individual at generation i in a familial branch (see Figure 4.2(a), and p is the major allele frequency of the SNP, we have the following lemma.

Lemma 4.2. *The sequence $\{\mathbf{Y}_n\}$ is a discrete stochastic process. Moreover, it is a first-order homogeneous Markov chain, i.e., the conditional probability of $\mathbf{Y}_{i+1}$ given (direct) ancestors in one of the parents' family branches is formally defined as $P(\mathbf{Y}_{i+1} = y_{i+1}|\mathbf{Y}_i = y_i, \mathbf{Y}_{i-1} = y_{i-1}, \dots) = P(\mathbf{Y}_{i+1} = y_{i+1}|\mathbf{Y}_i = y_i)$. Its transition matrix P is defined as follows:*

$$P = \begin{pmatrix} p & 1-p & 0 \\ p/2 & 1/2 & (1-p)/2 \\ 0 & p & 1-p \end{pmatrix},$$

where $p_{mn} = P(\mathbf{Y}_{i+1} = n|\mathbf{Y}_i = m)$, m and n belonging to the state space $\{0, 1, 2\}$.

Proof. Genotypes of individuals in a family can be modeled as a Bayesian network (BN), such as in Figure 4.2(a), where each node in the BN represents the SNP of a relative [124]. The two biological parents are also the two parents of each node in the BN. Thus, by definition, the SNP value given its two parents is conditionally independent of any of its ancestors. In our setting, where we focus on the ancestors in the familial branch of one parent, the same reasoning applies. This means that the SNP value $\mathbf{Y}_{i+1}$ is conditionally independent of any ancestor in the subnetwork (whose leaf node is $\mathbf{Y}_i$) given $\mathbf{Y}_i$. Thus,

$P(\mathbf{Y}_{i+1} = y_{i+1}|\mathbf{Y}_i = y_i, \mathbf{Y}_{i-1} = y_{i-1}, \dots) = P(\mathbf{Y}_{i+1} = y_{i+1}|\mathbf{Y}_i = y_i)$. Finally, the transition probability $P(\mathbf{Y}_{i+1} = y_{i+1}|\mathbf{Y}_i = y_i)$ is equal to

$$\sum_{z_i \in \{0,1,2\}} P(\mathbf{Y}_{i+1} = y_{i+1}|\mathbf{Y}_i = y_i, \mathbf{Z}_i = z_i)P(\mathbf{Z}_i = z_i), \tag{4.6}$$

where $P(\mathbf{Y}_{i+1} = y_{i+1}|\mathbf{Y}_i = y_i, \mathbf{Z}_i = z_i)$ is given by the Mendelian inheritance probabilities, and $P(\mathbf{Z}_i = z_i)$ by the major allele frequency p ($\mathbf{Z}_i$ is not observed). Equation (4.6) directly leads to the transition matrix P. □

Note that the reverse process, which is the conditional probability of $\mathbf{Y}_{i-1}$ given direct descendants $\mathbf{Y}_i$, Y_{i+1}, …, is also a first-order homogeneous Markov chain defined by the same matrix P where $p_{mn} = P(Y_{i-1} = n|Y_i = m)$. This means that going up or down the familial tree leads to the same conditional distributions. The corresponding Markov chain is shown in Figure 4.2(b).

Lemma 4.2 helps us determine the conditional probabilities of SNPs of direct ancestors or descendants given any relative's observed SNP. For instance, the conditional probability $P(\mathbf{Y}_{i+k}|\mathbf{Y}_i)$ of a relative k-degree apart from another direct relative at generation i whose SNP is observed and equal to m is given by $\pi_{i+k} = \pi_i P^k$, where π_i is a row vector that is equal to 1 in the m^{th} coordinate and 0 elsewhere. This is by definition of the Markov chain. Note also that the stationary distribution defined as the vector π such that $\pi = \pi P$, is equal to the vector of prior probabilities ($P(\mathbf{Y}_i)$), given by the major allele probability p:

$$\pi = \begin{pmatrix} p^2 & 2p(1-p) & (1-p)^2 \end{pmatrix}. \tag{4.7}$$

This follows intuition as π is defined to be equal to any of the columns of P^k when k tends to infinity. When the observed relative's generation j is really distant from the targeted individual's generation i in the family tree, the genomes of relatives have no more influence on each other. The conditional probabilities are well-defined for *direct* relatives. However, if the individual whose SNP is observed is not a relative in direct line (e.g., an uncle or a niece), the transition matrix P cannot be applied alone, and has to be combined with a matrix M whose elements m_{ab} represent the conditional probabilities $P(\mathbf{Y}_{i_1} = b|\mathbf{Y}_{i_2} = a)$ of i_1 given his sibling i_2. Defining $\mathbf{Y}^m_{i-1}$ and $\mathbf{Y}^f_{i-1}$ to be the mother and father SNP variable respectively, m_{ab} is derived as follows.

$$P(\mathbf{Y}_{i_1} = b|\mathbf{Y}_{i_2} = a) = \sum_{\substack{y^m_{i-1} \in \{0,1,2\} \\ y^f_{i-1} \in \{0,1,2\}}} P(\mathbf{Y}_{i_1} = b, \mathbf{Y}^m_{i-1} = y^m_{i-1}, \mathbf{Y}^f_{i-1} = y^f_{i-1}|\mathbf{Y}_{i_2} = a) \tag{4.8}$$

$$= \sum_{\substack{y^m_{i-1} \in \{0,1,2\} \\ y^f_{i-1} \in \{0,1,2\}}} P(\mathbf{Y}_{i_1} = b|\mathbf{Y}^m_{i-1} = y^m_{i-1}, \mathbf{Y}^f_{i-1} = y^f_{i-1}) \times \tag{4.9}$$

$$P(\mathbf{Y}^m_{i-1} = y^m_{i-1}|\mathbf{Y}^f_{i-1} = y^f_{i-1}, \mathbf{Y}_{i_2} = a)P(\mathbf{Y}^f_{i-1} = y^f_{i-1}|\mathbf{Y}_{i_2} = a), \tag{4.10}$$

where we used the chain rule to go from (4.8) to (4.10), and the fact that $P(\mathbf{Y}_{i_1}|\mathbf{Y}^m_{i-1}, \mathbf{Y}^f_{i-1}, \mathbf{Y}_{i_2}) = P(\mathbf{Y}_{i_1}|\mathbf{Y}^m_{i-1}, \mathbf{Y}^f_{i-1})$, since two siblings are conditionally independent given both their parents. $P(\mathbf{Y}_{i_1} = b|\mathbf{Y}^m_{i-1}, \mathbf{Y}^f_{i-1})$ is given by the Mendelian

inheritance probabilities, $P(\mathbf{Y}_{i-1}^f|\mathbf{Y}_{i_2}=a)$ is given by matrix P, and

$$P(\mathbf{Y}_{i-1}^m|\mathbf{Y}_{i-1}^f,\mathbf{Y}_{i_2}=a)=\frac{P(\mathbf{Y}_{i_2}=a|\mathbf{Y}_{i-1}^f,\mathbf{Y}_{i-1}^m)P(\mathbf{Y}_{i-1}^m)}{P(\mathbf{Y}_{i_2}=a|\mathbf{Y}_{i-1}^f)}, \tag{4.11}$$

using the Bayes rule and the fact that $P(\mathbf{Y}_{i-1}^m|\mathbf{Y}_{i-1}^f)=P(\mathbf{Y}_{i-1}^m)$, as two parents are independent if no child is observed. Again, one can compute every factor of (4.11) by using the inheritance probabilities, matrix P, and the major allele frequency p. Matrix M is defined as

$$M=\begin{pmatrix} p^2+pq+\frac{q^2}{4} & pq+\frac{q^2}{2} & \frac{q^2}{4} \\ \frac{p^2}{2}+\frac{pq}{4} & \frac{p^2}{2}+\frac{3}{2}pq+\frac{q^2}{2} & \frac{pq}{4}+\frac{q^2}{2} \\ \frac{p^2}{4} & \frac{p^2}{2}+pq & \frac{p^2}{4}+pq+q^2 \end{pmatrix}, \tag{4.12}$$

where $q=1-p$.

We define the 3×3 distance matrix D with elements $d_{ij}=\|i-j\|_1$ and the (column) vector $\mathbf{y}_i$ with m^{th} coordinate equal to 1 and others 0 if relative r_i's SNP has value m. We have the following lemma.

Lemma 4.3. *The genomic privacy E_i of individual r_i at any SNP is:*

$$\begin{cases} E_{i,0}=\pi D\mathbf{y}_i & \text{if no relative reveals the SNP} \\ E_{i|j}=\pi_j P^k D\mathbf{y}_i & \text{if } r_i \text{ and } r_j \text{ are direct relatives and } r_j\text{'s SNP is revealed} \\ E_{i|j}=\pi_j P^u M P^v D\mathbf{y}_i & \text{if } r_i \text{ and } r_j \text{ are not direct relatives and } r_j\text{'s SNP is revealed} \end{cases}$$

where k is the degree of kinship between r_i and r_j, u is the degree of kinship between r_j and his (direct) ancestor whose sibling is the (direct) ancestor of r_i, and v is the degree of kinship between r_i and his (direct) ancestor whose sibling is r_j's (direct) ancestor.

Proof. The genomic privacy of one SNP g of individual r_i is given by $\sum_{\hat{y}_i\in\{0,1,2\}} P(\mathbf{Y}_i=\hat{y}_i|\mathbf{Y_O})||y_i-\hat{y}_i||_1$.

(i) If no observations are made, then $P(\mathbf{Y}_i=\hat{y}_i|\mathbf{Y_O})=P(\mathbf{Y}_i=\hat{y}_i)$, the prior probability, which is given by the major allele frequency p. This is equal to π given in (4.7). The second element $||y_i-\hat{y}_i||_1$ is simply expressed in matrix format by $D\mathbf{y}_i$. Hence, $E_{i,0}=\pi D\mathbf{y}_i$.

(ii) If r_i is a k^{th}-degree direct relative of r_j, then the conditional probability distribution $P(\mathbf{Y}_i=y_i|\mathbf{Y}_j=y_j)$ is given by $\pi_{j+k}=\pi_j P^k$ from Lemma 4.2, leading to $E_{i|j}=\pi_j P^k D\mathbf{y}_i$.

(iii) In case r_i and r_j are not in direct line, we need to split the conditional probability computation into two. First, we need to compute the conditional probability of the direct ancestor a_j of r_j who is a sibling of the direct ancestor a_i of r_i. If a_j and j are u^{th}-degree relatives, $\pi_{a_j}=\pi_j P^u$. Then, as a_j and a_i are siblings, we make use of matrix M defined in (4.12) to compute the conditional probability of a_i's SNP given a_j's SNP value. Thus, $\pi_{a_i}=\pi_{a_j}M=\pi_j P^u M$. Finally, if a_i and i are v^{th}-degree relatives, we have $\pi_i=\pi_{a_i}P^v=\pi_j P^u M P^v$. Hence, we get $E_{i|j}=\pi_j P^u M P^v D\mathbf{y}_i$. □

To illustrate the third case of Lemma 4.3, let us take for example two close relatives, uncle and nephew. If r_j is the uncle of r_i, then the genomic privacy of r_i given r_j at a certain SNP is $E_{i|j}=\pi_j P^1 M P^0 D\mathbf{y}_i=\pi_j PMD\mathbf{y}_i$ whereas, if r_j is the nephew of r_i, the genomic privacy of r_i is $E_{i|j}=\pi_j MPD\mathbf{y}_i$.

We can now quantify genomic privacy for a range of SNPs and get closed-form NE.

Theorem 4.1. *For any value $c \in [0, \infty[$, the pure Nash equilibrium is:*

$$(x_1^*, x_2^*) = \begin{cases} (1,1) & \text{if } c < \max(t_1, t_2) \\ (1,1), (0,0) & \text{if } \max(t_1, t_2) < c < p_a \min(t_1^0, t_2^0) \\ (0,0) & \text{if } c > p_a \min(t_1^0, t_2^0) \end{cases} \tag{4.13}$$

if $\max(t_1, t_2) < p_a \min(t_1^0, t_2^0)$, *where* $t_i^0 = \frac{1}{|\mathcal{G}|} \sum_{l:g_l \in \mathcal{G}} \pi^l D\mathbf{y}_i^l$, $t_i = \frac{p_a}{|\mathcal{G}|}\big(\sum_{l:g_l \in \mathcal{G}}((1 - p_a)\pi^l + p_a \pi_j^l P_l^k) D\mathbf{y}_i^l\big)$ *if* r_i *and* r_j *are direct* k^{th}*-degree relatives, and* $t_i = \frac{p_a}{|\mathcal{G}|}\big(\sum_{l:g_l \in \mathcal{G}}((1 - p_a)\pi^l + p_a \pi_j^l P_l^u M P_l^v) D\mathbf{y}_i^l\big)$ *if* r_i *and* r_j *are not direct relatives,* u *and* v *as defined in Lemma 4.3. If* $\max(t_1, t_2) > p_a \min(t_1^0, t_2^0)$, *the second-case NE* $(1,1), (0,0)$ *becomes* $(0,1)$ *if* $t_1^0 < t_2^0$ *and* $(1,0)$ *if* $t_1^0 > t_2^0$, *and* $\max(t_1, t_2)$ *and* $p_a \min(t_1^0, t_2^0)$ *are swapped in the inequality bounds.*

Proof. By summing over all SNPs in $\mathcal{G}$ the genomic privacy expressions computed in Lemma 4.3 and embedding them into the NE computed in Lemma 4.1 (keeping only pure NE), after some reordering, we get the NE in (4.13), as well as the expressions t_i's and t_i^0. □

In order to make these NE more tangible, we quantify genomic privacy by relying upon real genomic data. We make use of the CEPH/Utah Pedigree 1463 that contains the partial DNA sequences of 4 grandparents, 2 parents, and 11 children [59]. Figure 4.3 represents this family with 3 children. We consider all the SNPs that are available on chromosome 1, around 82,000. Note that, thanks to our closed-form expression of $E_{i|j}$, its computation on 82,000 SNPs takes less than one second. This is at least three order of magnitude faster than the belief propagation algorithm run on the same set of SNPs (which takes more than one hour). Figure 4.4 shows the thresholds separating the three different cases of NE in Theorem 4.1 with respect to p_a and c. $(1,1)$ stands below the two (dotted) red and green curves, and $(0,0)$ stands above these two curves. Thus, we note that for most values of c and p_a, either both relatives secure their genomes (if c is smaller than around half of p_a), or both do not secure them (if c is greater than around half of p_a). This shows that players, if they have similar cost c, have aligned incentives, leading to an efficient NE. However, there are some values of c and p_a for which two pure NE $(1,1)$ and $(0,0)$ co-exist. It is between the two curves, if the (dotted) red curve lies above the green one. If the green curve lies above the dotted one,[7] then we have either $(0,1)$ if $E_{1,0} < E_{2,0}$ or $(1,0)$ if $E_{1,0} > E_{2,0}$. The discrepancy between the two curves is the highest in Figure 4.4(c), as the difference between the initial privacy levels $E_{i,0}$'s and posterior levels $E_{i|j}$ is the most significant (see Table 4.2). On the contrary, in the game between C7 and GP1, the posterior levels $E_{i|j}$ are closer to the initial ones $E_{i,0}$ (because the two players are second-degree relatives), and the $E_{i,0}$'s differ between the two players, leading to inefficient NE, like $(0,1)$, as described above.

Discussion: We conclude that, for most security cost values and probabilities of successful breach, the players follow the same strategies, even though their genomic privacy levels are slightly different. They both either secure their devices, or do not secure.

We now move to the disclosure game G_d. Table 4.3 shows the resulting payoffs for two players P_1 and P_2. The following theorem determines its NE.

[7] This happens for $p_a < 0.29$ in Figure 4.4(a) and $p_a < 0.78$ in Figure 4.4(b).

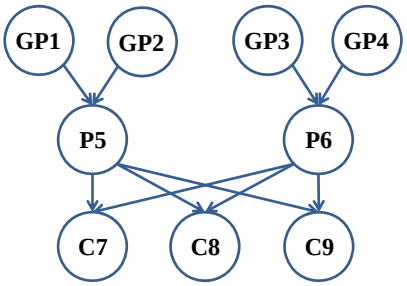
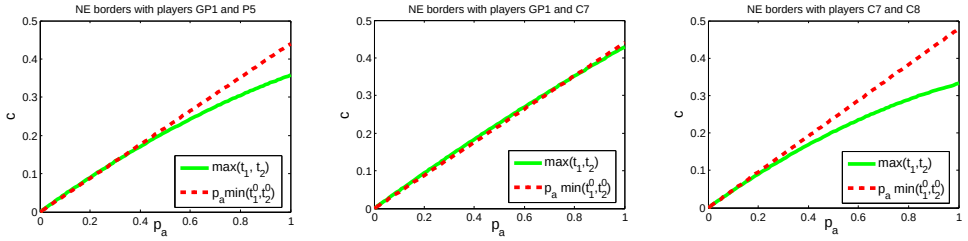


Figure 4.3: Bayesian network representation of nine relatives of the CEPH/Utah pedigree 1463.



(a) (b) (c)

Figure 4.4: Thresholds of Theorem 4.1 separating the three different pure NE cases of G_s. We show three different scenarios with two players: (a) Grandparent GP1 and parent P5, (b) GP1 and child C7, and (c) children C7 and C8.

Table 4.2: Genomic privacy levels of grandparent GP1, parent P5, children C7 and C8, from the Utah family shown in Figure 4.3.

(P_1, P_2)	$E_{1,0}$	$E_{1\|2}$	$E_{2,0}$	$E_{2\|1}$
(P5,GP1)	0.4741	0.3579	0.4402	0.3179
(C7,GP1)	0.4788	0.4296	0.4402	0.3878
(C7,C8)	0.4788	0.3310	0.4803	0.3321

Theorem 4.2. *For any value $b_1^d \in [0, \infty[$, and $b_2^d \in [0, \infty[$, the pure Nash equilibrium of game G_d is:*

$$(d_1^*, d_2^*) = \begin{cases} (0,0) & \text{if } (b_1^d < E_{1,0} - c_1) \wedge (b_2^d < E_{2|1} - c_2) \vee (b_1^d < E_{1|2} - c_1) \wedge (b_2^d < E_{2,0} - c_2) \\ (1,1),(0,0) & \text{if } (E_{1|2} - c_1 < b_1^d < E_{1,0} - c_1) \wedge (E_{2|1} - c_2 < b_2^d < E_{2,0} - c_2) \\ (1,1) & \text{if } (b_1^d > E_{1,0} - c_1) \wedge (b_2^d > E_{2|1} - c_2) \vee (b_1^d > E_{1|2} - c_1) \\ (0,1) & \text{if } (b_1^d < E_{1|2} - c_1) \wedge (b_2^d > E_{2,0} - c_2) \\ (1,0) & \text{if } (b_1^d > E_{1,0} - c_1) \wedge (b_2^d < E_{2|1} - c_2) \end{cases}$$

where $E_{i,0} = \frac{1}{|\mathcal{G}|} \sum_{l:g_l \in \mathcal{G}} \pi^l D\mathbf{y}_i^l$, $E_{i|j} = \frac{1}{|\mathcal{G}|} \sum_{l:g_l \in \mathcal{G}} \pi^l P_l^k D\mathbf{y}_i^l$ if i and j are direct k^{th}-degree relatives and, if i and j are not direct relatives, $E_{i|j} = \frac{1}{|\mathcal{G}|} \sum_{l:g_l \in \mathcal{G}} \pi^l P_l^u DMP_l^v \mathbf{y}_i^l$.

The NE can be derived from the best responses of each player that are quite straightforward by taking a close look at Table 4.3. Figure 4.5 illustrates the NE computed in

Table 4.3: Normal form of the two-player game G_d.

$P_1 \backslash P_2$	$d_2 = 0$	$d_2 = 1$
$d_1 = 0$	$(b_1^g - c_1, b_2^g - c_2)$	$(b_1^g - c_1 - (E_{1,0} - E_{1\|2}),$ $b_2^g + b_2^d - E_{2,0})$
$d_1 = 1$	$(b_1^g + b_1^d - E_{1,0},$ $b_2^g - c_2 - (E_{2,0} - E_{2\|1}))$	$(b_1^g + b_1^d - E_{1,0}, b_2^g + b_2^d - E_{2,0})$

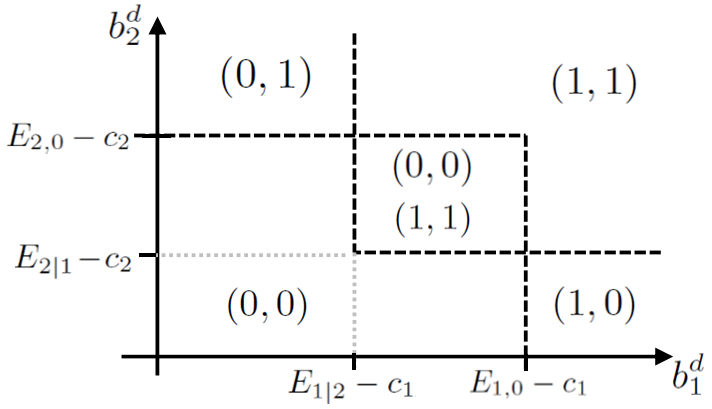


Figure 4.5: Dependence of the NE with respect to the genome-sharing benefits b_1^d and b_2^d.

Theorem 4.2. These NE depend essentially on the value of $b_i^d + c_i$ with respect to $E_{i,0}$ and $E_{i|j}$. A player P_i will disclose his genome given the other player discloses it as long as $b_i^d + c_i > E_{i|j}$ whereas, if the other player's best response is to not share, P_i will share only if $b_i^d + c_i > E_{i,0}$. Table 4.2 shows concrete values of genomic privacy $E_{1,0}$, $E_{2,0}$, $E_{1|2}$, and $E_{2|1}$, for first-degree direct relatives, second-degree direct relatives, and siblings.

Discussion: We conclude that, in G_d, if the discrepancy between the sharing benefits that players perceive is high enough between the two players, the players follow opposite strategies, one player putting the other's privacy at risk by disclosing his genome.

4.4.2 Altruistic Players

In this Subsection, we analyze how the equilibria evolve when the players are not purely selfish, but consider also their relatives' payoffs when making their decisions. Intuitively, by becoming more socially concerned, the players' decisions and their resulting NE should lead to higher social welfare. However, as we will see, social welfare does not always increase with altruism, unless some coordination between players happen.

To evaluate how the NE is affected by altruistic behavior, we focus on the game G_d. Player P_1 considers the altruistic payoff $u_1^a(d_1, d_2) = u_1(d_1, d_2) + \alpha^{k(1,2)} u_2(d_1, d_2)$ instead of merely $u_1(d_1, d_2)$. The same applies symmetrically for P_2. We define the *familial Nash equilibrium* (FNE) as a strategy profile where no player can reduce his altruistic payoff u^a by unilaterally changing his strategy given the other player's strategy. Defining $b_i = b_i^d + c_i$ for the ease of presentation, we have the following theorem.

Theorem 4.3. *For any value $b_1 \in [0,\infty[$, and $b_2 \in [0,\infty[$, the pure FNE is:*

$$(d_1^*, d_2^*) = \begin{cases} (0,0) & \text{if } (b_1 < E_{1,0} + \alpha^k(E_{2,0} - E_{2|1})) \wedge (b_2 < E_{2|1}) \vee \\ & \quad (b_1 < E_{1|2}) \wedge (b_2^d < E_{2,0} + \alpha^k(E_{1,0} - E_{1|2})) \\ (1,1),(0,0) & \text{if } (E_{1|2} < b_1 < E_{1,0} + \alpha^k(E_{2,0} - E_{2|1}) \wedge \\ & \quad (E_{2|1} < b_2 < E_{2,0} + \alpha^k(E_{1,0} - E_{1|2})) \\ (1,1) & \text{if } (b_1 > E_{1,0} + \alpha^k(E_{2,0} - E_{2|1})) \wedge (b_2 > E_{2|1}) \vee \\ & \quad (b_1 > E_{1|2}) \wedge (b_2 > E_{2,0} + \alpha^k(E_{1,0} - E_{1|2}) \\ (1,0) & \text{if } (b_1 > E_{1,0} + \alpha^k(E_{2,0} - E_{2|1})) \wedge (b_2 < E_{2|1}) \\ (0,1) & \text{if } (b_1 < E_{1|2}) \wedge (b_2 > E_{2,0} + \alpha^k(E_{1,0} - E_{1|2}) \end{cases} \tag{4.14}$$

where $E_{i,0} = \frac{1}{|\mathcal{G}|}\sum_{l:g_l \in \mathcal{G}} \pi^l D\mathbf{y}_i^l$, $E_{i|j} = \frac{1}{|\mathcal{G}|}\sum_{l:g_l \in \mathcal{G}} \pi^l P_l^k D\mathbf{y}_i^l$ *if* P_i *and* P_j *are direct* k^{th}*-degree relatives and, if* P_i *and* P_j *are not direct relatives,* $E_{i|j} = \frac{1}{|\mathcal{G}|}\sum_{l:g_l \in \mathcal{G}} \pi^l P_l^u DMP_l^v \mathbf{y}_i^l$.

The FNE can be derived from the Table of payoffs updated to u_i^a. These different NE are depicted in Figure 4.6 by circled numbers separated by (thick) dotted lines. Note the up-right shift of the borders of the $(0,0)$ FNE compared to the selfish NE (red dotted lines). This tells us that, by considering the other's player utility, the decision maker will choose to disclose his genome for a higher value of b_i than in the purely selfish scenario.
Discussion: We conclude that altruism, by internalizing externalities into players' payoffs, tends to reduce the privacy loss caused by the other player at equilibrium.

We now describe the strategies that a social planner would choose on behalf of the players in order to maximize social welfare, thus to attain the *social optimum* U^*.

Theorem 4.4. *For any value $b_1 \in [0,\infty[$, and $b_2 \in [0,\infty[$, the social optimum U^* is reached with the following strategies:*

$$(d_1^*, d_2^*) = \begin{cases} (0,0) & \text{if } (b_1+b_2 < E_{1,0}+E_{2,0}) \wedge (b_1 < E_{1,0}+E_{2,0}-E_{2|1}) \wedge (b_2 < E_{1,0}+E_{2,0}-E_{1|2}) \\ (1,0) & \text{if } (b_1 > E_{1,0} + E_{2,0} - E_{2|1}) \wedge (b_2 < E_{2|1}) \\ (0,1) & \text{if } (b_2 > E_{1,0} + E_{2,0} - E_{1|2}) \wedge (b_1 < E_{1|2}) \\ (1,1) & \text{if } (b_1 + b_2 > E_{1,0} + E_{2,0}) \wedge (b_2 > E_{2|1}) \wedge (b_1 > E_{1|2}) \end{cases} \tag{4.15}$$

where $E_{i,0} = \frac{1}{|\mathcal{G}|}\sum_{l:g_l \in \mathcal{G}} \pi^l D\mathbf{y}_i^l$, $E_{i|j} = \frac{1}{|\mathcal{G}|}\sum_{l:g_l \in \mathcal{G}} \pi^l P_l^k D\mathbf{y}_i^l$ *if* P_i *and* P_j *are direct* k^{th}*-degree relatives and, if* P_i *and* P_j *are not direct relatives,* $E_{i|j} = \frac{1}{|\mathcal{G}|}\sum_{l:g_l \in \mathcal{G}} \pi^l P_l^u DMP_l^v \mathbf{y}_i^l$.

This theorem is derived by simply summing the utilities of both players in Table 4.3 for all four strategy combinations and selecting the combination that leads to the maximum sum for any value of b_1 and b_2. The social optimum's strategies are represented schematically with respect to b_1 and b_2 by the texture of Figure 4.6. Given this social optimum $U^*(\mathbf{d})$, the Price of Anarchy (PoA), that measures how the game efficiency decreases due to selfishness, is defined as $U^*(\mathbf{d})/\min_{NE} U(\mathbf{d})$ [116]. The Price of Stability (PoS) also measures this inefficiency but considers the best NE instead of the worst one, assuming that players coordinate, thus is defined as $U^*(\mathbf{d})/\max_{NE} U(\mathbf{d})$ [28].

Following the notion of Windfall of Friendship (WoF) proposed in [132], we define the Windfall of Kinship (WoK) as the ratio between the social welfare of the worst FNE

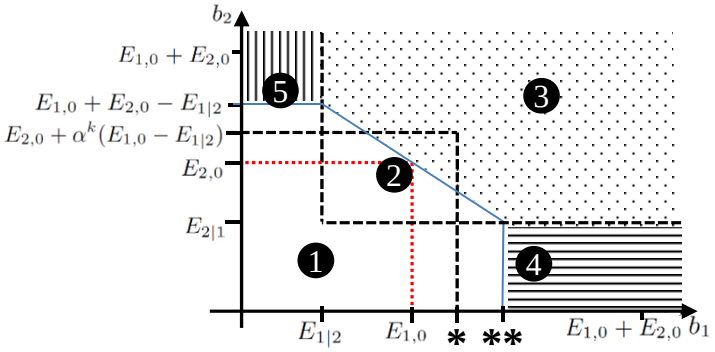


Figure 4.6: Familial NE and social optima with respect to b_1 and b_2. Circled numbers represent the five different cases of Theorem 4.3, in order, separated by (thick) dotted lines in the figure. The red (small) dotted lines represent the borders of Figure 4.5. The four different texture patterns represent the strategies of the social optimum, depicted in Theorem 4.4: white for $(0,0)$, vertical lines for $(1,0)$, horizontal lines for $(0,1)$, and dots for $(1,1)$. The single asterisk is $E_{1,0}+\alpha^k(E_{2,0}-E_{2|1})$, and the double asterisk is $E_{1,0}+E_{2,0}-E_{2|1}$.

and the social welfare of the worst NE:

$$\kappa(\alpha,k)=\frac{\min_{FNE}U(\mathbf{d})}{\min_{NE}U(\mathbf{d})} \tag{4.16}$$

Given this definition, we can state the following theorem.

Theorem 4.5. *If b_1, b_2 are such that*

$$\begin{cases} b_1+b_2>E_{1,0}+E_{2,0} \\ b_1<E_{1,0}+\alpha^k(E_{2,0}-E_{2|1}) \\ b_2<E_{2,0}+\alpha^k(E_{1,0}-E_{1|2}), \end{cases} \tag{4.17}$$

then $\kappa(\alpha,k)<1$ for any $k\geq 1$ and $0<\alpha\leq 1$.

Proof. Let us focus on the cases where NE and FNE differ. This happens essentially in the two strips between $E_{i,0}$ and $E_{i,0}+\alpha^k(E_{j,0}-E_{j|i})$ for $i=1$, $j=2$ and the contrary (see Figure 4.6). We know, from Theorem 4.4, that the social optimum in these strips is reached at $(0,0)$ except if $b_1+b_2>E_{1,0}+E_{2,0}$ where it is reached at $(1,1)$. Moreover, we know that the FNE or worse FNE is $(0,0)$ in these strips according to Theorem 4.3. However, the NE is $(1,1)$ if $(b_1^d>E_{1,0}-c_1)\wedge(b_2^d>E_{2|1}-c_2)\vee(b_1^d>E_{1|2}-c_1)\wedge(b_2^d>E_{2,0}-c_2)$ according to Theorem 4.2. Let us now compute the ratio between the social welfare at $(0,0)$ (FNE) and the social welfare at $(1,1)$ (NE):

$$\begin{aligned}\kappa&=\frac{b_1^g+b_2^g-c_1-c_2}{b_1^g+b_2^g+b_1^d+b_2^d-E_{1,0}-E_{2,0}}\\&=\frac{b_1^g+b_2^g-c_1-c_2}{b_1^g+b_2^g+b_1+b_2-c_1-c_2-E_{1,0}-E_{2,0}}.\end{aligned}$$

κ is strictly smaller than 1 if and only if $b_1+b_2>E_{1,0}+E_{2,0}$. This gives us the first condition of (4.17), the two others being given by the area we are focusing on. If

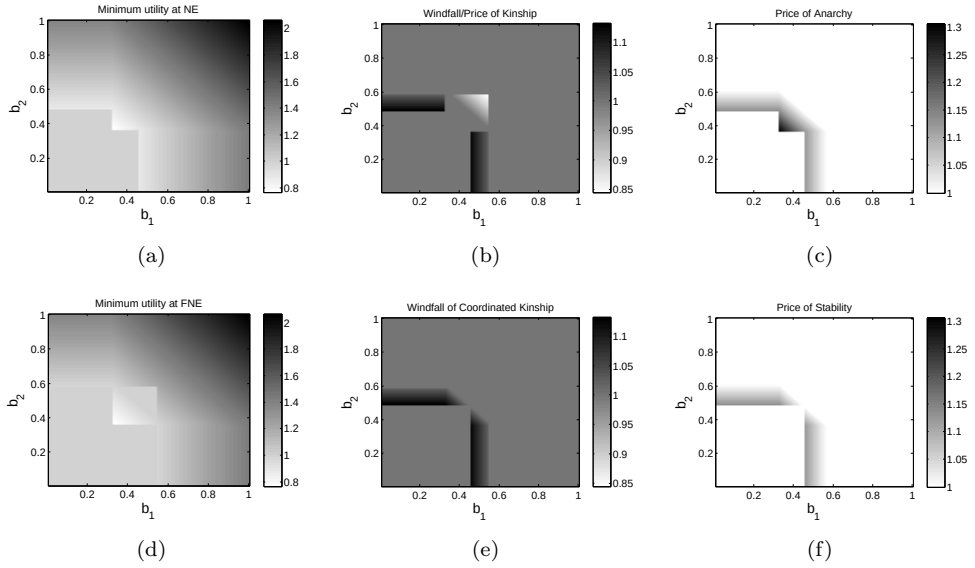


Figure 4.7: Evaluation of the (in)efficiency of the NE and FNE with respect to b_1 and b_2. (a) Minimum social welfare at NE, (b) Windfall/Price of Kinship, (c) Price of Anarchy, (d) minimum social welfare at FNE, (e) Windfall of Coordinated Kinship, and (f) Price of Stability for the game G_d with players GP1 and P5, $\alpha = 0.8$, and $b_1^g = b_2^g = 0.5$.

$b_1 + b_2 < E_{1,0} + E_{2,0}$, we know that the FNE or worse FNE cannot be improved as they are $(0, 0)$, which is the social optimum. Note that the region where $\kappa < 1$ is the small triangle with dots texture in the FNE area defined by circle 2 in Figure 4.6. □

This theorem tells us that, contrarily to intuition, altruism in a family does not necessarily lead to higher social welfare, and induces a Price of Kinship rather than a windfall if the b_i's are in the range defined in (4.17). In this range, the social optimum is to disclose their genomes for both players, but there is the possibility to end up in a "non-disclose" $(0, 0)$ FNE due to the altruistic factor, leading to a worse outcome than in the selfish NE. However, note that the WoK is always less than or equal to the PoA. Indeed, as for any $\alpha \in [0, 1]$, $k \geq 1$, $\min_{FNE} U(\mathbf{d}) \leq U^*(\mathbf{d})$, it directly follows from (4.16) that $\kappa(\alpha, k) \leq$ PoA.

If we assume that some coordination can happen between the players, we can define the Windfall of Coordinated Kinship (WoCK) as the ratio between the social welfare of the best FNE and the social welfare of the best NE:

$$\gamma(\alpha, k) = \frac{\max_{FNE} U(\mathbf{d})}{\max_{NE} U(\mathbf{d})} \tag{4.18}$$

This new definition allows us to state the following theorem.

Theorem 4.6. *For any $b_1 \in [0, \infty[$, $b_2 \in [0, \infty[$, $k \geq 1$, and $\alpha \in [0, 1]$, it holds that:*

$$1 \leq \gamma(\alpha, k) \leq PoS \leq PoA. \tag{4.19}$$

Proof. First, as $\min_{NE} U(\mathbf{d}) \leq \max_{NE} U(\mathbf{d})$, we get

$$\frac{U^*(\mathbf{d})}{\max_{NE} U(\mathbf{d})} \leq \frac{U^*(\mathbf{d})}{\min_{NE} U(\mathbf{d})}, \tag{4.20}$$

thus PoS $\leq$ PoA. Moreover, as $\max_{FNE} U(\mathbf{d}) \leq U^*(\mathbf{d})$, we get

$$\frac{\max_{FNE} U(\mathbf{d})}{\max_{NE} U(\mathbf{d})} \leq \frac{U^*(\mathbf{d})}{\max_{NE} U(\mathbf{d})}, \tag{4.21}$$

thus $\gamma(\alpha, k) \leq$ PoS. We know from Theorem 4.5 that $\kappa(\alpha, k) < 1$ in the triangle defined by (4.17). The difference between κ and γ is that the latter uses the best FNE whereas the former uses the worst FNE. In the area defined by (4.17), two FNE co-exist,$(0, 0)$ and $(1, 1)$. The worst FNE is $(0, 0)$ and the best is $(1, 1)$, which corresponds to the social optimum and the selfish NE in this area. Hence, $\max_{FNE} U(\mathbf{d}) / \max_{NE} U(\mathbf{d}) = 1$ in this triangle. For the rest of the (b_1, b_2) values where FNE and NE differ, the FNE is always equal to the social optimum U^* defined in Theorem 4.4, thus the social welfare of NE cannot be greater. It follows that $\gamma(\alpha, k) \geq 1$. □

In order to evaluate how the NE, FNE, WoK, WoCK, PoA, and PoS evolve in practice, we make use of the genomic data provided by the Utah family. We choose the two relatives GP1 and P5, and compute their genomic privacy based on their actual SNPs, as in Subsection 4.4.1. We set $\alpha = 0.8$, $b_1^g = b_2^g = 0.5$ and compute results (NE, FNE, ...) for b_1 and b_2 varying between 0 and 1, with granularity 0.01. Figure 4.7 shows the resulting graphs. First, we notice the up-right shift of $(0, 0)$ between NE and FNE that follows the borders shown in Figure 4.6. We also see that minimum social welfare is minimal in the squares standing in the middle of both Figures 4.7(a) and (4.7(d)). Looking at Figure 4.7(b), we clearly notice that the WoK is smaller than one for values of b_1 and b_2 close to 0.5, confirming Theorem 4.5. However, as soon as coordination happens between players, the ratio between the social welfare of FNE and the social welfare of NE (WoCK) becomes always greater than or equal to one, as illustrated in Figure 4.7(e). Finally, we note that PoA and PoS are always greater than or equal to one, that PoS $\leq$ PoA, and that PoS $\geq$ WoCK, confirming Theorem 4.6.

Discussion: In conclusion, if players cannot coordinate, their altruistic conservatism (or prudence), regarding the disclosure of their genomes can lead to a worse social outcome than in the purely selfish setting, as shown in Theorem 4.5 and in Figure 4.7(b).

4.5 n-Player Game

In this section, we extend the genomic privacy game to consider $n > 2$ relatives. Contrarily to the two-player framework that allowed us to derive closed-form expressions, and thus compute all pure Nash equilibria very efficiently, we now face a more challenging problem. First, in general, all players (family members) can influence other players' payoffs, thus all other players' strategies have to be taken into account when a family member optimizes his own decision. Second, privacy levels $E_{i|-i}$ cannot be expressed in closed forms if more than one other family members disclose their genomes.

In order to represent this complex game in a compact way and reduce its complexity, we rely upon *multi-agent influence diagrams* (MAIDs), introduced by Koller and Milch [114]. A MAID is an extension of the Bayesian network framework that embeds,

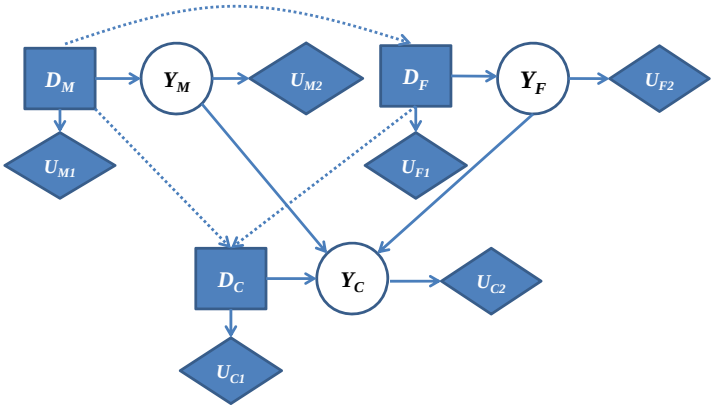


Figure 4.8: Multi-agent influence diagram representing a trio (mother, father, child) with one decision variable (square), one chance variable (circle) representing the SNP(s) of the player, and two utility variables (diamonds) per player. Full lines represent probabilistic or deterministic dependencies, whereas dotted ones represent the variables that an agent observes when he makes his decision. This figure illustrates a game with sequential moves, perfect information, and with purely selfish players.

in addition to random variables, decision and utility variables, and enables to consider multiple strategic agents, thus permitting to represent games. We define a MAID $\mathcal{M}_d$ representing the n-player genomic-privacy game G_d. We show an example of $\mathcal{M}_d$ for a trio in Figure 4.8. Note that in MAIDs, all variables are depicted by capital letters. The chance[8] variable $\mathbf{Y}_i$ is defined as $P(\mathbf{Y}_i = y_i) = 1$ (other values having probability 0) if $d_i = 1$, and $P(\mathbf{Y}_i = \hat{y}_i | \mathbf{Y_O} = \mathbf{y_O})$ if $d_i = 0$. Note that, we represent the chance variable $\mathbf{Y}_i$ for a single SNP but there actually are $|\mathcal{G}|$ chance variables that directly depend on d_i, and are independent of each other (thus they can be considered in "parallel" in the MAID). A child's SNP is probabilistically determined by his parents' genomes, as explained in Subsection 3.2.1. We also define two utility variables: $u_{i1} = b_i^g + d_i b_i^d - E_{i,0}$ that directly depends on d_i , and $u_{i2} = E_i$ that directly depends on the chance variable $\mathbf{Y}_i$. Note that E_i is zero if $d_i = 1$ (genomic privacy drops to zero) and $E_i = E_{i|-i}$ if $d_i = 0$. Then, P_i's payoff u_i is $u_{i1} + u_{i2}$.

We assume that players move (decide) sequentially and under perfect information of previous decisions made by other players. Variables observed when a decision is made are depicted by dotted directed edges. For instance, in Figure 4.8, the following decision ordering is shown: mother, father and then child. Under these assumptions, we can state the following lemma.

Lemma 4.4. *If a player $P_i \in \mathcal{P}$ moves, i.e. chooses his decision rule, at node D_i before P_j makes his own decision at node D_j, then D_i is not s-reachable from D_j.*

The concept of *s-reachability* is defined in Definition 5.3 of [114]. In a nutshell, if D_i is s-reachable from D_j, then D_i is relevant to D_j or, in other words, D_j strategically relies on D_i. The main idea of the proof is that, if a decision node D_i is observed by D_j (dotted edge in Figure 4.8), it means that the decision rule $\delta(d_j)$ at D_j will be conditioned on the instantiations of D_i. The decision rule at D_j will be defined as $\delta(d_j|d_i), \forall d_i \in \{0, 1\}$, thus

[8] In MAIDs, random variables are called chance variables.

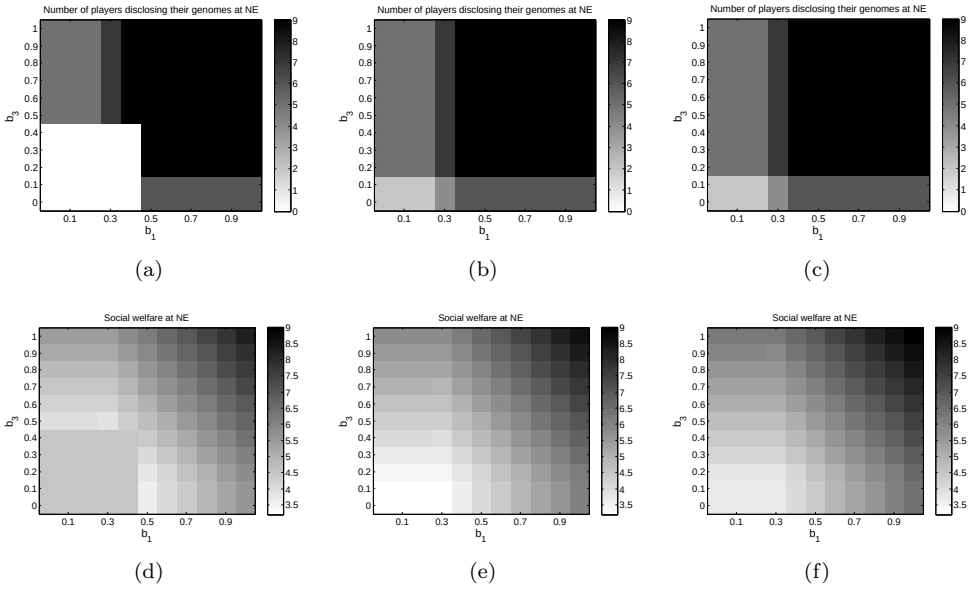


Figure 4.9: Outcome of the n-player game. Number of players disclosing their genomes (first row) and social welfare (second row) at NE in the n-player game G_d with the family members shown in Figure 4.3. We set $b_2 = 0.4$ in (a) and (d), $b_2 = 0.6$ in (b) and (e), and $b_2 = 0.8$ in (c) and (f).

this decision will not be affected by a change D_i. However, because D_j is not observed by P_i when he makes his decision, D_j will be relevant to D_i, thus s-reachable from D_i. Under perfect information, we can define, for any sequence of strategic decision among players, an acyclic relevance graph[9] using Lemma 4.4. From this acyclic relevance graph, we can construct a topological ordering of the decision nodes $D_1, ..., D_n$ such that if D_i is s-reachable from D_j, then $i < j$. In the example shown in Figure 4.8, the topological ordering is D_C, D_F, D_M. In the general case, the topological ordering is such that, if P_i chooses his decision rule before P_j, then $j < i$. Hence, the topological ordering corresponds to the reverse decision order.

Theorem 4.7. *By iteratively deriving the optimal decision rule $\delta^*(d_i|\boldsymbol{pa}_{D_i})$ for each node D_i in topological order, and every instantiation $\boldsymbol{pa}_{D_i}$ of its parents in the MAID, we obtain a strategy profile $\mathbf{d}^*$ that is a Nash equilibrium of $\mathcal{M}_d$.*

This theorem essentially follows from Algorithm 6.1 and Theorem 6.1 of [114]. Note that, in our scenario, under perfect information assumption, we do not need to define an arbitrary fully mixed strategy profile at the beginning of the algorithm. The algorithm defined by Theorem 4.7 is similar to the one defined by backward induction in extensive-form games. However, thanks to the MAID approach, we can run inference on $\mathcal{M}_d$ in order to compute the expected utilities given the decision rules of every player, and to eventually find a NE in $\mathcal{O}(|\mathcal{G}|2^n)$ instead of $\mathcal{O}(|\mathcal{G}|3^{2n})$ in the extensive-form game.

We numerically compute the NE of the n-player game G_d by considering the Utah family shown in Figure 4.3. We assume the sequence of decisions to be the following:

[9] See the definition of a relevance graph in Definition 5.4 of [114].

GP1, GP2, GP3, GP4, P5, P6, C7, C8, and C9. We skip the details of the algorithm and inference and provide here the main numerical results. We focus on 1,000 randomly chosen SNPs of chromosome 1,[10] and compute the NE and resulting social welfare of the family for varying values of b_i's. We assume $b_i = b_1$ for all grandparents, $b_i = b_2$ for all parents, and $b_i = b_3$ for all children. We make b_1 and b_3 vary between 0 and 1 with granularity 0.1, and b_2 be equal to 0.4 (first column of Figure 4.9), 0.6 (second column of Figure 4.9) and 0.8 (third column of Figure 4.9). In the first row of Figure 4.9, we see the number of players disclosing their genomes at NE. In Figure 4.9(a), because b_2 is quite small (0.4), if b_1 and b_3 are also small (≤ 0.4), then nobody has the incentive to share his genome. If b_1 or b_3 are high enough for the grandparents and the children to share their genomes, it will automatically lead the parents to do the same, because then the parents' genomic privacy will be reduced by their relatives' decision. We see this in the left strip where $b_3 \geq 0.5$ and $b_1 \leq 0.2$: 5 relatives disclose their SNPs, the 3 children and the 2 parents. By increasing b_1 to 0.3, then 2 of the 4 grandparents have the incentive to share their SNPs, considering their privacy levels. We notice that when b_2 increases to 0.6 (Figure 4.9(b)) and 0.8 (Figure 4.9(c)), then even if b_1 and b_3 are very small, the parents' best responses are to disclose their SNPs. Then, if b_1 increases to 0.3 while $b_3 \leq 0.1$ (bottom strip), then 2 grandparents have the incentive to share their SNPs (4 players thus share them), and from $b_1 \geq 0.4$ all grandparents have the incentive to disclose their genomes.

Discussion: We conclude that, in some cases, when the perceived benefits do not clearly outweigh the genomic privacy losses, some people with same perceived benefits might end up with different strategies at equilibrium.

Looking now at the social welfare values at NE, the most interesting finding is that the social welfare decreases between Figure 4.9(d) and Figure 4.9(e) for values of b_1 and b_3 smaller than 0.5, even though b_2 increases from 0.4 to 0.6. This is due to the privacy externalities that are created by the parents disclosing their SNPs whereas grandparents and children have no incentives to do the same. Hence, misaligned incentives have negative impact on the social welfare of a family. Our MAID $\mathcal{M}_d$ model can be easily adapted to take altruism into account.

We note that the proposed n-player game requires all family members to give their decisions sequentially but at a given time instant, which might not be feasible in real life considering infants, or even unborn family members.

4.6 Related Work

Interdependent risks in privacy have recently been demonstrated and explored in different settings. Due to their intrinsic social nature, online social networks (OSNs) are especially prone to indirect privacy risks. Mislove et al. evaluate the fraction of users in an OSN that would be sufficient in order to infer attributes of the remaining users [135]. Henne et al. study how OSN pictures uploaded by friends can reveal information about one's own location [85]. Dey et al. analyze the risk of age inference in OSNs, by notably relying on information posted by users' friends and friends-of-friends [54]. In the context of location privacy, Vratonjic et al. show how mobile users connecting to location-based services from the same IP address can indirectly compromise location privacy of others [166]. Olteanu

[10] As in Section 4.4, LD is not used as we assume the same set $\mathcal{G}$ of SNPs potentially shared by the players and targeted by the adversary.

et al. study how users reporting co-locations with other users (e.g., on online social networks) can decrease others' location privacy [138]. In order to precisely quantify the effect of co-location information, they propose an optimal inference algorithm and two polynomial-time approximate inference algorithms.

Acquisti et al. are among the first to propose an economic model to formalize incentives and interactions between rational agents in the context of privacy [20]. More precisely, the authors rely on a game-theoretic approach in order to study the incentives and behaviors of participants in anonymity networks. Freudiger et al. analyze the behavior of selfish mobile nodes that want to protect their location privacy by changing pseudonym and at a minimum cost [69]. Contrarily to Freudiger et al. who assumed a global attacker, Humbert et al. consider a local adversary with multiple eavesdropping stations. They study the interaction between such an adversary and mobile users deploying mix zones to protect their location privacy [92]. Shokri et al. make use of Stackelberg Bayesian games in order to model the user-adversary interplay in the context of localization attacks [153]. Biczók and Chia tackle, by using a game-theoretic framework, the issue of interdependent risks caused by agents with misaligned incentives regarding their privacy in online social networks [38]. They show how negative externalities can lead to inefficient equilibria. Their work builds upon the literature on IDS games, that is surveyed in [122]. We follow a similar approach for genomic privacy and, in addition, precisely evaluate the possible direct and indirect privacy losses with a well-defined framework and by using real data. The non-linear genetic dependencies between players in genomic privacy are also a novel compared to previous work.

4.7 Summary

In this chapter, we have studied the interplay between the members of a given family, who have to decide about whether to share their genomes and how much to invest to secure their storage. We model the interplay between the family members with different incentives and privacy levels by using a game-theoretic approach and predict their behaviors at equilibrium. First, we extensively study a two-player game between two purely selfish or partially altruistic family members. In this context, we also derive a closed-form expression to quantify genomic privacy of any individual given one his relatives' genome, which dramatically decrease the computational burden compared to the belief propagation algorithm used in Chapter 3. Then, we extend this framework to an n-player game using multi-agent influence diagrams, an extension of the Bayesian network framework that enables us to include decision and utility variables. This approach allows us to significantly reduce the computational complexity of computing the Nash equilibria with respect to a classic extensive-form game.

In the two-player setting, our results show that the players follow similar security strategies, that is either invest in security or not, if the cost of investment is the same for both players, regardless of the cost's actual value and of the probability of successful breach from the adversary. In the case of the genome-sharing game, however, if the benefits of disclosure perceived by the players are different enough, the players follow opposite strategies, one player putting the other's privacy at risk by sharing his genome. We also show that, in general, altruism tends to reduce the privacy loss caused by the other player at equilibrium. However, for some perceived sharing benefits, the altruistic prudence can, surprisingly, lead to a worse social outcome than in the purely selfish

scenario. Finally, in the n-player game, we notice that, when the perceived benefits do not clearly outweigh the genomic-privacy losses, some players with similar sharing benefits might end up with different strategies at equilibrium.

Chapter 5

Cooperative Genomic Privacy Protection

5.1 Introduction

As we have seen in Chapter 3, genomic privacy was popularized by the story of Henrietta Lacks whose cells were sequenced and whose DNA sequence was put online without the consent of her descendants [10]. After complaints from the family, essentially due to privacy concerns, Henrietta's genome was taken offline, and in 2013, the National Institutes of Health (NIH) came to an agreement with the Lacks family, which gave them some control over her genome. Even though this agreement enables the genomic researchers to use Henrietta's genome again, it also draws attention to the lack of techniques for balancing the benefits of genomic research with personal and kin genomic privacy risks. Richard Sharp, the director of biomedical ethics at the Mayo Clinic, warned that the agreement was only a "one-off solution" rather than a broad policy that addresses the tension between research and relatives' privacy, and he added that a "new policy" was absolutely needed [14].

Anonymization was the first countermeasure proposed to protect genomic privacy, but in many different studies it was proven to be inadequate [81, 158, 86]. Another protection mechanism is to add noise to aggregate statistical results (to satisfy differential privacy) [64, 102], but at the cost of reduced accuracy. The last option proposed in the literature is to rely on cryptographic techniques [36, 33]. Even though these techniques are proven to be effective for using genomic data in healthcare [33, 50], computational complexity becomes very high when it comes to conducting statistical tests on large numbers of encrypted genomes for genomic research [106].

In this chapter, we present a genomic-privacy preserving mechanism (GPPM) for reconciling people's willingness to share their genomes (e.g., to help research[1]) with privacy. Our GPPM acts at the individual data level, not at the aggregate data (or statistical) level like in [64, 102]. Focusing on the most relevant type of variants (the SNPs), we study the trade-off between the usefulness of disclosed SNPs (utility) and genomic privacy. We consider an individual who wants to share his genome, yet who

[1] http://opensnp.wordpress.com/2011/11/17/first-results-of-the-survey-on-sharing-genetic-information/

is concerned about the subsequent privacy risks for himself and his family. Thus, we design a system that maximizes the disclosure utility but does not exceed a certain level of privacy loss within a family, considering (i) kin genomic privacy, (ii) personal privacy preferences (of the family members), (iii) privacy sensitivities of the SNPs, (iv) correlations between SNPs, and (v) the research utility of the SNPs. Our GPPM can automatically evaluate the privacy risks of all the family members and decide which SNPs to disclose. To achieve this goal, it relies on probabilistic graphical models and combinatorial optimization. Our results indicate that, given the current data model, genomic privacy of an entire family can be protected while an appropriate subset of genomic data can be made available. Our contributions can be summarized as follows:

- We propose a GPPM for enabling genomic research while protecting personal and kin genomic privacy.
- Given the genomic data model, our obfuscation mechanism maximizes the utility and meets all the privacy constraints of a given family.
- Using combinatorial optimization, we first compute the optimal solution without considering correlations between SNPs, and then we extend the algorithm to address non-linear constraints induced by these correlations.

5.2 Genomic-Privacy Preserving Mechanism

In order to mitigate attribute-inference attacks and protect genomic and health privacy, the GPPM relies upon an *obfuscation mechanism*. In practice, obfuscation can be implemented by adding noise to the SNP values, by injecting fake SNP values, by reducing precision, or by simply hiding the SNP values. In this thesis, we choose SNP hiding, essentially because the genomic research community would not receive other options positively. Indeed, genetic researchers are very reluctant about adding noise or fake data, notably because of the huge investment they make to increase (sequencing) accuracy. We assume one family member, at a given time, who wants to disclose his SNPs and to guarantee a minimum privacy level for him and his family.

5.2.1 Settings

Like in the two previous chapters, we focus on one family whose members are defined by the set $\mathcal{R}$ ($|\mathcal{R}| = n$). We assume that there is only one donor r_D who makes the decision to share his genome at a given time. His relatives might have already publicly shared some of their genomic data on the Internet. r_D takes this into account when he makes his own disclosure decision. We let $\mathcal{G}$ ($|\mathcal{G}| = m$) be the set of SNPs. Its cardinality m can go up to 50 million, as this is currently the approximate number of SNPs in the human population [11]. In practice, however, people put online (e.g., on OpenSNP) up to 1 million of the most significant SNPs. We let $\mathbf{x}_D$ represent the SNPs of r_D (x^i_D is the value of SNP g_i of the donor r_D), that are all initially undisclosed (hidden). Finally, we let $\mathbf{y}_D$ represent the m-size binary decision vector of r_D, where $y^i_D = 1$ means SNP g_i will be disclosed, and $y^i_D = 0$ means SNP g_i will remain hidden. Note that the decision to disclose a SNP could also be probabilistic, thus transforming y^i_D into a continuous variable in $[0, 1]$. We leave the study of the continuous case for future work.

We express the privacy constraints of a family member both in terms of genomic and health privacy. Our framework can account for different privacy preferences for different family members, SNPs, and diseases. For all $r_j \in \mathcal{R}$, $g_i \in \mathcal{G}$, we define the privacy sensitivity as s_j^i. We can set the s_j^i's to be equal by default. Then, an individual willing to personalize his privacy preferences may further define his own privacy sensitivities regarding specific SNPs based on his privacy concerns regarding, e.g., certain phenotypes. As mentioned in Chapter 3, the most well-known example of such a scenario is the case of James Watson, co-discoverer of DNA, who made his whole DNA sequence publicly available, with the exception of one gene known as Apolipoprotein E (ApoE), one of the strongest predictors for the development of Alzheimer's disease.[2] We let the sets $\mathcal{P}_s^i \subseteq \mathcal{G}$ and $\mathcal{P}_d^i$ include the privacy-sensitive SNPs and privacy-sensitive diseases of individual r_i, respectively. We represent the tolerance to the genomic-privacy loss of individual i as $\text{Pri}(i, \mathcal{P}_s^i)$, and the tolerance to the health-privacy loss of individual r_i regarding disease $d \in \mathcal{P}_d^i$ as $\text{Pri}(i, d)$. These tolerance values represent the maximum privacy loss (after the disclosure of r_D's SNPs) that an individual would bear. By considering the privacy losses instead of the absolute privacy levels, we ensure that the donor will more likely reveal a SNP whose value is already well inferred by the attacker before donor's disclosure (e.g., by using SNPs previously shared by the donor's relatives). Note that these tolerance values can always be updated for any new family member willing to disclose his genome. Finally, the utility function is a non-decreasing function of the norm of $\mathbf{y}_D$, as the knowledge of more SNPs can only help genomic research. In a first step towards enhanced genomic privacy, we assume linear contribution of SNPs on utility. Formally, we define u_i to be the utility provided by SNP g_i. Note that, in practice, the utility of the SNPs can be determined by the research authorities and can vary based on the study.

5.2.2 Linear Optimization

Optimization Problem

The donor faces an optimization problem: How to maximize research utility while protecting his own and his relatives' genomic and health privacy. First, the objective function is formally defined as $\sum_{i:g_i \in \mathcal{G}} u_i y_D^i$. Then, privacy constraints are defined, for each individual, as the sum of privacy losses induced by the donor's disclosure over all SNPs. This sum must be capped by the respective privacy loss tolerances of all family members. Formally, for all individuals $r_j \in \mathcal{R}$ and SNPs $g_i \in \mathcal{G}$, the privacy loss induced by the disclosure of x_D^i is defined as $(E_j^i(y_D^i = 0) - E_j^i(y_D^i = 1))$. Note here that the privacy loss at a given SNP g_i for any relative is only affected by the donor's decision y_D^i regarding SNP g_i but no other SNP $g_k \neq g_i$, meaning that LD correlations are not taken into account. We make this assumption here in order to define linear constraints. We show how to extend the linear optimization problem to include LD correlations in Subsection 5.2.3. Finally, note that if an individual $r_j, j \neq D$ has already revealed his SNP g_i, the privacy loss at this SNP for g_i for r_j is zero, because $E_j^i(y_D^i = 0) = E_j^i(y_D^i = 1) = 0$.

For all $r_j \in \mathcal{R}$, $g_i \in \mathcal{G}$, the privacy weight p_j^i is defined as

$$p_j^i = s_j^i \times (E_j^i(y_D^i = 0) - E_j^i(y_D^i = 1)). \tag{5.1}$$

[2]Later researchers have used correlations in the genome to unveil Watson's predisposition to Alzheimer's [137]. In this work, we also consider such correlations.

Clearly, p_j^i at a given SNP g_i can be different for each family member r_j, depending on how close he is from the donor in the family tree, on the actual values x_j^i and x_D^i of his and the donor's SNPs, and on his sensitivity. Note that $s_j^i = 0 \; \forall j \notin \mathcal{P}_s^i$.

We can now define the linear optimization problem as

$$\begin{aligned} \underset{\mathbf{y}_D}{\text{maximize}} \quad & \sum_{i:g_i \in \mathcal{G}} u_i y_D^i \\ \text{subject to} \quad & \sum_{i:g_i \in \mathcal{P}_s^j} p_j^i y_D^i \leq \text{Pri}(j, \mathcal{P}_s^j), \forall r_j \in \mathcal{R} \\ & \sum_{k:g_k \in \mathcal{S}_d} p_j^k y_D^k \leq \text{Pri}(j, d), \forall d \in \mathcal{P}_d^j, \forall r_j \in \mathcal{R} \\ & y_D^i \in \{0,1\}, \forall g_i \in \mathcal{G}, \end{aligned} \tag{5.2}$$

where $\mathcal{S}_d$ is the set of SNPs that are associated with disease d. Note that, for the last inequality, we replace the sensitivity s_j^k in p_j^k by the contribution c_k of SNP k to disease d described in (3.11), and we embed the normalization factor $\sum_k c_k$ of (3.11) in $\text{Pri}(j, d)$.

Optimization Algorithm

Our optimization problem is very similar to the multidimensional knapsack problem [70]. We decide to follow the branch-and-bound method proposed by Shih [150], because it finds the optimal solution, represents a good trade-off between time and storage space, and allows for the extension of the algorithm to null and negative (privacy) weights. A branch-and-bound algorithm is a systematic enumeration of all candidate solutions, where large subsets of candidate solutions are pruned by using upper bounds on the quantity being optimized. A branch-and-bound method generally relies on two main rules: (i) the estimation of the upper bound at any node (state of assigned variables) in the search tree, and (ii) a choice criterion for the selection of a branching variable at the node selected for further partitioning.

In order to find (i), Shih suggests treating the C-constraint knapsack problem as C single-constraint knapsack problems with the same objective function, and then computing the value associated to the optimal fractional solution (thus relaxing $y_D^i \in \{0,1\}$ into $y_D^i \in [0,1]$) of all of these C problems separately. The fractional optimal solution is easier to solve than the integer solution, as it enables us to sort the items (SNPs), with respect to their ratios between utility and privacy weights $r_j^i = u_i / p_j^i$, from the highest to the lowest ratios, and then to select all the highest ones that can fit under the constraint, with the last SNP being partially included (based on the remaining room). Note that, in our setting, we can have different orderings of SNPs for different constraints, based on the p_j^i values of the family members. The computation of the fractional optimal solution is repeated C times, for the C different optimization problems, leading to C optimal values. Then, the upper bound at the given node is defined as the minimum among all these C values.

The node selected for the next branching is defined as the one in the search frontier whose upper bound is the highest among all nodes in the frontier, and where the solution associated with this upper bound is infeasible (some variables being different than 0 and 1, or some constraints being not satisfied). The branching variable is the one whose ratio is the smallest among all the non-zero free variables (variables not explicitly assigned to 0 or 1 at a node) in this infeasible solution. If the solution at this node is feasible (all

decision variables assigned to 0 or 1 and all constraints satisfied), then it is optimal, and the algorithm stops.

Let us mention that our optimization problem has two main differences with the multidimensional knapsack problem. First, the privacy metrics, hence weights, are expressed in real values, between 0 and 2 for E_j^i, whereas the knapsack problem assumes integer numbers only. In order to obtain integer values, we merely multiply all our privacy weights p_j^i's and tolerance values Pri(.) by 10^k, where $k \in \mathbb{N}^+$ depends on the precision we want to attain, and then round the weights to the closest greater integer and the tolerance values to the closest smaller integer. This ensures that all privacy constraints in the space of real numbers are still satisfied. Second, the privacy weight p_j^i can be equal to zero (e.g., if x_j^i has already been disclosed by r_j) or even negative (when the donor reveals a SNP whose value increases the privacy of his relative(s) at the same SNP).[3] Thus, the ratios r_j^i might not be defined or be negative. In order to resolve this issue, we give a higher ranking in the ordering of SNPs to ratios with null weights with respect to those with positive weights, and we give an even higher ranking to those with negative weights. We furthermore give higher ranking to negative weights with absolute values higher than the others. To enforce this ranking in practice in Section 5.3, we set $r_j^i = u_i/0.1$ for null p_j^i's, and $r_j^i = u_i|p_j^i|/0.01$ for negative p_j^i's. Note that, due to the requirement of integer values for weights, all other (positive) weights p_j^i belong, after the aforementioned multiplication by 10^k and rounding, to $\mathbb{N}^+$.

The output of the above optimization algorithm is an optimal solution $\mathbf{y}_D^*$ that represents the SNPs the donor could disclose and an optimal value u^* representing the maximum research utility. We represent the optimal candidate SNPs to be shared in the m-size vector $\tilde{\mathbf{x}}_D$ where $\tilde{x}_D^i = x_D^i$ if $y_D^{i*} = 1$ and $\tilde{x}_D^i = \perp$ if $y_D^{i*} = 0$. This is the output we see in state 2 of Figure 5.1. We give $\tilde{\mathbf{x}}_D$ as input to the non-linear algorithm described in Subsection 5.2.3 to eventually reach state 3.

5.2.3 Non-Linear Extension

Non-Linear Optimization Problem

The LD correlations between the SNPs are not considered in the above optimization problem in order for the constraints to remain linear. In this subsection, we propose an extension of the branch-and-bound algorithm in order to deal with non-linear constraints.

Whereas in the case without LD, the privacy loss at a given SNP g_i of individual r_j depended only on the donor's decision y_D^i regarding SNP g_i, we have now to consider all the SNPs in LD with g_i to evaluate the privacy loss at g_i. Defining $\tilde{E}_j^i$ to be the privacy level of individual r_j at SNP g_i quantified by including LD correlations, the privacy loss at SNP g_i of individual r_j induced by the disclosure of $\tilde{x}_D$ is equal to $(\tilde{E}_j^i(\mathbf{y}_D = \mathbf{0}) - \tilde{E}_j^i(\mathbf{y}_D^* D))$. This leads to the following updated privacy weights

$$\tilde{p}_j^i = s_j^i \times (\tilde{E}_j^i(\mathbf{y}_D = \mathbf{0}) - \tilde{E}_j^i(\mathbf{y}_D^*)). \tag{5.3}$$

Note that now the argument of $\tilde{E}_j^i$ is the entire vector $\mathbf{y}_D$ and not only y_D^i, because of LD. The optimization problem in (5.2) is reformulated as a non-linear optimization

[3] For example, assume a child to be homozygous-major at a given SNP and his father to be heterozygous. Then, the estimation error for the child's SNP, thus the child's privacy at this SNP, increases when the father's SNP is observed by the attacker (compared to the case when it is unknown, when only the MAF is used, and this MAF is close to 0).

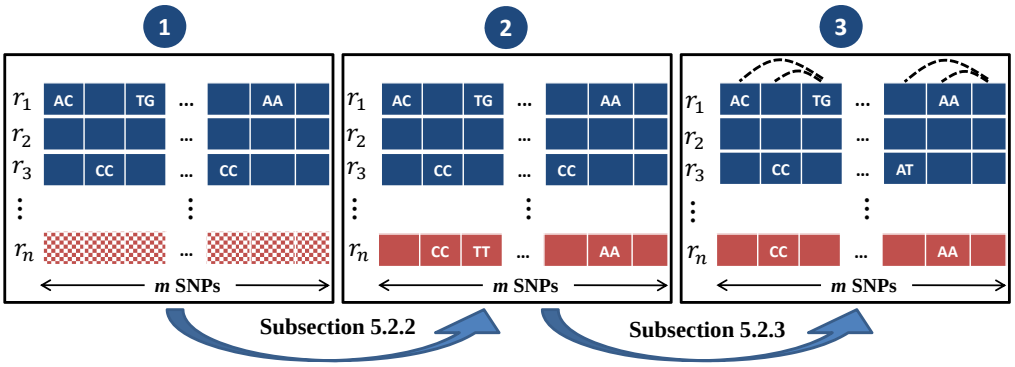


Figure 5.1: Main steps of the optimization algorithm. Without loss of generality, the donor r_D is assumed to be the n-th member of the family, thus $D = n$. First, the donor selects a subset of candidate SNPs to be shared using the optimization algorithm of Subsection 5.2.2, and then reveals less or more SNPs depending on the updated privacy weights computed with LD by relying upon the fine-tuning step of Subsection 5.2.3.

problem:

$$\begin{aligned} \underset{\mathbf{y}_D}{\text{maximize}} \quad & \sum_{i:g_i \in \mathcal{G}} u_i y_D^i \\ \text{subject to} \quad & \sum_{i:g_i \in \mathcal{P}_s^j} \tilde{p}_j^i(\mathbf{y}_D) \leq \text{Pri}(j, \mathcal{P}_s^j), \forall r_j \in \mathcal{R} \\ & \sum_{k:g_k \in \mathcal{S}_d} \tilde{p}_j^k(\mathbf{y}_D) \leq \text{Pri}(j, d), \forall d \in \mathcal{P}_d^j, \forall r_i \in \mathcal{R} \\ & y_D^i \in \{0,1\}, \forall g_i \in \mathcal{G}. \end{aligned} \tag{5.4}$$

Instead of solving this very complex optimization problem, we rely on the optimal solution $\mathbf{y}_D^*$ computed in Subsection 5.2.2, embed it into (5.4), and check whether the privacy constraints are still met with the updated privacy weights $\tilde{p}_j^i$'s. Let us first study the case when no SNP has been disclosed by any relative before the donor's decision.[4] If $\mathcal{X}_O = \varnothing$, then

$$\sum_{i:g_i \in \mathcal{P}_s^j} \tilde{E}_j^i(\mathbf{y}_D = \mathbf{0}) = \sum_{i:g_i \in \mathcal{P}_s^j} E_j^i(\mathbf{y}_D = \mathbf{0}) \tag{5.5}$$

and, because of LD correlations,

$$\sum_{i:g_i \in \mathcal{P}_s^j} \tilde{E}_j^i(\mathbf{y}_D^*) \leq \sum_{i:g_i \in \mathcal{P}_s^j} E_j^i(\mathbf{y}_D^*). \tag{5.6}$$

Embedding (5.5) and (5.6) in (5.1) and (5.3), we get

$$\sum_{i:g_i \in \mathcal{P}_s^j} \tilde{p}_j^i(\mathbf{y}_D^*) \geq \sum_{i:g_i \in \mathcal{P}_s^j} p_j^i y_D^i, \tag{5.7}$$

meaning that, for the same value of $\text{Pri}(j, \mathcal{P}_s^j)$ in (5.2) and (5.4), the privacy constraint of family member i in (5.4) will be violated with high likelihood once LD is taken into

[4] Without loss of generality, we focus on the genomic-privacy constraints in the following.

account. If $\mathcal{X}_O \neq \varnothing$, then two scenarios can happen. If

$$\underbrace{\sum_{i:g_i \in \mathcal{P}_s^j} E_j^i(\mathbf{y}_D^*) - \tilde{E}_j^i(\mathbf{y}_D^*)}_{\substack{\text{Privacy difference using LD or not} \\ \text{after } \tilde{\mathbf{x}}_D \text{ is revealed}}} \geq \underbrace{\sum_{i:g_i \in \mathcal{P}_s^j} E_j^i(\mathbf{y}_D = \mathbf{0}) - \tilde{E}_j^i(\mathbf{y}_D = \mathbf{0})}_{\substack{\text{Privacy difference using LD or not} \\ \text{before } \tilde{\mathbf{x}}_D \text{ is revealed}}},$$

then we get the same inequality (5.7), leading to the same consequences of constraint violation. If , on the contrary,

$$\sum_{i:g_i \in \mathcal{P}_s^j} E_j^i(\mathbf{y}^{*D}) - \tilde{E}_j^i(\mathbf{y}^{*D}) < \sum_{i:g_i \in \mathcal{P}_s^j} E_j^i(\mathbf{y}_D = \mathbf{0}) - \tilde{E}_j^i(\mathbf{y}_D = \mathbf{0}), \tag{5.8}$$

then we get

$$\sum_{i:g_i \in \mathcal{P}_s^j} \tilde{p}_j^i(\mathbf{y}_D^*) < \sum_{i:g_i \in \mathcal{P}_s^j} p_j^i y_D^i, \tag{5.9}$$

which might allow the donor to reveal more of his SNPs without violating any of his relatives' privacy constraints. At a first glance, Inequality (5.9) looks counterintuitive. However, in order to understand it, let us look at Inequality (5.8), which states that the difference in privacy levels if LD is used or not is smaller after the observation of a subset of the donor's SNPs $\tilde{\mathbf{x}}_D$. This means that, by revealing his own SNPs, the donor reduces the importance of using LD correlations to correctly infer some of the SNPs of his relatives. For instance, let us assume the donor to be the father of a child r_j whose mother has already revealed SNP g_i, in LD with another SNP g_k revealed by the child. Furthermore, assume that the father, mother, and child are homozygous major at SNPs g_i and g_k. Now, before the father reveals his SNP g_i (with value $x_D^i = 0$), there is some uncertainty about the child's SNP g_j (with value $x_j^i = 0$); but by observing SNP g_k of the child (with value $x_j^k = 0$), the attacker improves his estimation if he uses LD correlation and thus reduces his estimation error, meaning $\tilde{E}_j^i(\mathbf{y}_D = \mathbf{0}) < E_j^i(\mathbf{y}_D = \mathbf{0})$. However, once the father decides to reveal his homozygous major SNP g_i ($y_D^{i*} = 1$), the attacker is certain that the child's SNP g_j is homozygous major (because both mother and father SNPs are homozygous major and revealed), regardless if LD is used or not, i.e. $E_j^i(\mathbf{y}_D^*) = \tilde{E}_j^i(\mathbf{y}_D^*) = 0$. Thus, we have $E_j^i(\mathbf{y}_D^*) - \tilde{E}_j^i(\mathbf{y}_D^*) < E_j^i(\mathbf{y}_D = \mathbf{0}) - \tilde{E}_j^i(\mathbf{y}_D = \mathbf{0})$, leading by extension to Inequality (5.8).

Fine-Tuning Algorithm

Let us first describe how we proceed if one or multiple constraints are violated once LD correlations are considered in the privacy quantification. In this case, we first select the privacy constraint that is not met anymore with the highest difference between $\text{Pri}(j, \mathbf{P}_s^j)$ (or $\text{Pri}(j, d)$) and the newly computed privacy losses. Focusing on the set of genomic-privacy constraints, we thus select the constraint of the family member k, where

$$k = \underset{j:r_j \in \mathcal{R}}{\arg\max} \{ \sum_{i:g_i \in \mathcal{P}_s^j} \tilde{p}_j^i(\mathbf{y}_D^*) - \text{Pri}(j, \mathcal{P}_s^j) \}. \tag{5.10}$$

We want then to hide some SNPs g_i in $\tilde{\mathbf{x}}_D$ (i.e. where $y_D^{*i} = 1$) in order that the constraint of relative r_k is satisfied again. For all the SNPs whose value is revealed in $\tilde{\mathbf{x}}_D$, we compute a global privacy weight δ_k^i for SNP g_i of r_k that includes the privacy loss

induced by SNP g_i on the SNPs $g_l \in \mathcal{L}$ in LD with g_i. We compute this global privacy weight at SNP g_i for individual r_k as

$$\begin{aligned}\delta_k^i &= \bar{p}_k^i + \sum_{l:g_l \in \mathcal{L}} \bar{p}_k^l \\ &= s_k^i(\tilde{E}_k^i(\mathbf{y}_D = \mathbf{0}) - \tilde{E}_k^i(\mathbf{y}_D^*)) + \sum_{l:g_l \in \mathcal{L}} s_k^l(\tilde{E}_k^l(\mathbf{y}_D = \mathbf{0}) - \tilde{E}_k^l(\mathbf{y}_D^*)).\end{aligned} \tag{5.11}$$

Then, we compute the ratios of each SNP g_i (revealed in $\tilde{\mathbf{x}}_D$) for individual r_k as $\bar{r}_k^i = \delta_k^i/u_i$. The SNPs with the highest ratios represent those where LD correlations cause the highest decrease in the genomic privacy of family member r_k and/or provide low utility to the optimal solution $\mathbf{y}_D^*$ computed in Subsection 5.2.2. Thus, these should be removed first from the set of SNPs to be shared in order to meet the privacy constraint of individual r_k again, and to cause the smallest decrease in utility.

To see whether the privacy constraint is met for the family member r_k, we iteratively remove such SNPs (starting with the one with the highest ratio) in $\tilde{\mathbf{x}}_D$ and, after each removal, we input the new solution to the quantification box. We repeat this until all the privacy constraints are satisfied for all family members in $\mathcal{R}$. Finally, the SNPs left in $\tilde{\mathbf{x}}_D$ after the final iteration are publicly shared. This case is illustrated in state 3 of Figure 5.1.

In the case where including LD correlations in the privacy quantification actually decreases privacy losses, the privacy constraints are still met and can even enable for potential new SNPs to be included in $\tilde{\mathbf{x}}_D$. In this case, we select the genomic-privacy constraint where the remaining room between the genomic-privacy constraint and the newly computed privacy loss is the smallest, i.e. we select the constraint of the family member r_k, where

$$k = \underset{j:r_j \in \mathcal{R}}{\arg\min}\{\mathrm{Pri}(j, \mathcal{P}_s^j) - \sum_{i:g_i \in \mathcal{P}_s^j} \tilde{p}_j^i(\mathbf{y}_D^*)\}. \tag{5.12}$$

For all SNPs *not* revealed in $\tilde{\mathbf{x}}_D$ (i.e., where $y_D^{i*} = 0$), we compute the privacy decrease led by LD for r_k compared to the privacy level computed without LD. We compute this privacy difference at a SNP g_i for individual r_k as

$$\Delta_k^i = E_k^i(y_D^{i*} = 0) - \tilde{E}_k^i(\mathbf{y}_D^*), \tag{5.13}$$

where $E_k^i(y_D^{i*} = 0)$ is the privacy value at SNP g_i for individual r_k after the linear optimization (without considering LD), and $\tilde{E}_k^i(\mathbf{y}_D^*)$ is the privacy quantified using LD. Then, we compute the ratios of each SNP g_i (not revealed yet) for individual r_k as $\bar{r}_k^i = (u_i\Delta_k^i)/s_k^i$. The SNPs with highest ratios represent those where LD correlations cause the most significant decrease in the genomic privacy of family member r_k, and/or provide high utility. Thus, these SNPs are the first ones that should be revealed and included in $\tilde{\mathbf{x}}_D$, in order to have the smallest difference in privacy loss, thus still meeting r_k's privacy constraint and providing maximal utility increase.

We iteratively include new SNPs in $\tilde{\mathbf{x}}_D$ and input the new solution to the quantification box to check whether all the privacy constraints are still met for all family members. We repeat this step until one privacy constraint is violated again, and we publicly share the last vector $\tilde{\mathbf{x}}_D$ to have satisfied all constraints. In the next section, we briefly show experimentally how close this fine-tuning algorithm is to the maximum found with exhaustive search.

5.3 Evaluation

In this section, we evaluate the effectiveness of our optimization algorithm for protecting individual and kin privacy. We study the balance between maximum achievable utility and the privacy of each individual in a family. The results show the total utility we can obtain for different genomic-privacy guarantees.

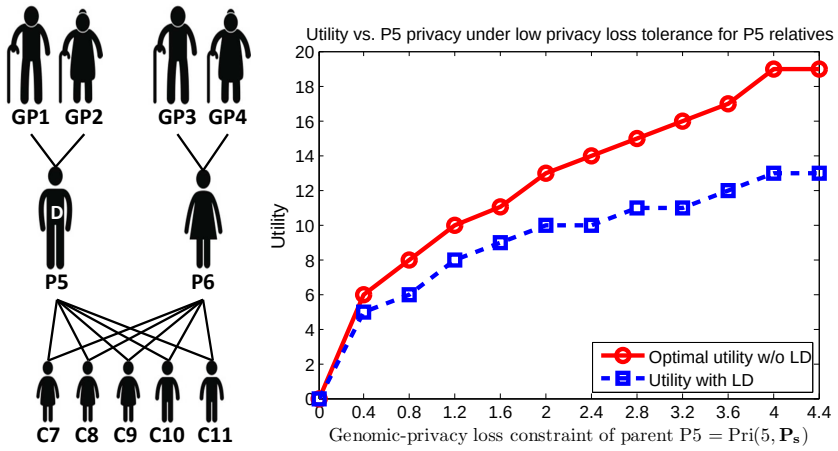
We make use of the CEPH/Utah Pedigree 1463 [59]. It includes the partial DNA sequences of 17 family members: 4 grandparents, 2 parents, and 11 children. In order to remain at a representative scale, we keep only 5 randomly chosen children out of 11, as in Chapter 3. Figure 5.2(a) presents the pedigree structure. We focus on 50 SNPs of chromosome 1 and assume one genomic-privacy constraint, including all the 50 SNPs for each family member. Thus, we have a total of 11 privacy constraints, which represents more constraints than other generic experiments in the optimization literature that included up to 5 or 7 constraints [70]. Considering LD strengths between $r^2 = 0.5$ (medium LD) and $r^2 = 1$ (strongest LD), each SNP is in LD with around 4.5 other SNPs, on average. We set a precision of 0.01 in our privacy weights and tolerance values, thus multiplying these real-valued elements by 10^2, and rounding them, as explained in Subsection 5.2.2. Parent P5 is assumed to be the donor in all scenarios presented in this section. In our evaluations, for the sake of simplicity, we assume each SNP is equally useful for the genomic research, i.e., $u_i = 1$ for all SNPs. We also assume the privacy sensitivities are equal, for all SNPs and individuals, i.e., $s_j^i = s$. Equal values of sensitivities for all SNPs would typically be the default setting if, for example, family members do not want to bother setting their privacy sensitivities themselves. Other distributions over the utility or sensitivity values should not alter the algorithm's performances significantly.

5.3.1 No Previous Disclosure by the Family

As of today most people have not publicly revealed their genome, we first analyze the case where no family member has shared any of his SNPs before the donor makes his decision. In other words, we assume that, initially, $\mathcal{X}_O = \varnothing$. We analyze the tension between utility and privacy for different values of parent P5's privacy constraint. Figure 5.2(b) shows the increase in the utility caused by the higher privacy loss tolerance of P5. Because a low tolerance to privacy loss is assumed for all the other relatives in the family in this case, the utility (computed without LD) cannot go beyond 19, even if P5's constraint increases beyond 4. We also notice that, once the LD is included in the privacy quantification, the utility decreases, reaching a maximum value of 13 instead of 19. This is because LD increases the privacy loss incurred when P5 reveals his SNPs, thus reducing the total number of SNPs parent P5 can reveal without violating the family's privacy constraints.

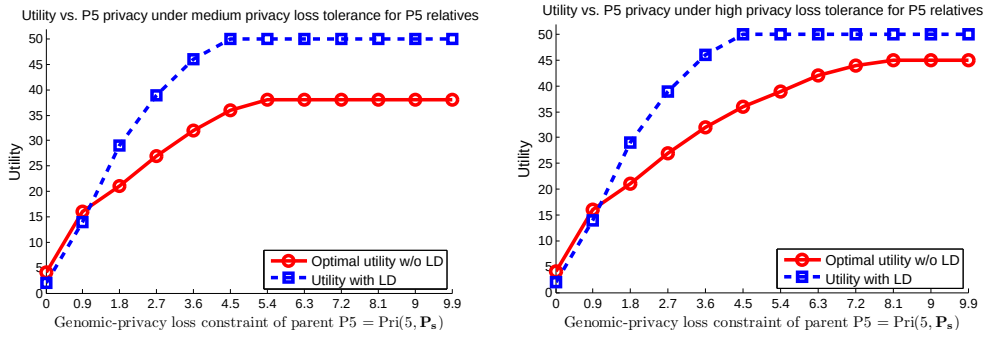
5.3.2 Previous Disclosure by Part of the Family

We want to mimic the situation where some of the family members have already revealed some of their SNPs. We simulate this by randomly selecting (with probability 0.5) some of the family members (except P5, who is the donor) who reveal a subset of their SNPs. Then, for the members who are selected to reveal their SNPs, we select, uniformly at random, some of their 50 SNPs to reveal. In the scenario we focus on, this leads to the following SNPs being revealed before the donor's decision: 8 (different) SNPs revealed



(a) (b)

Figure 5.2: Evaluation of the proposed solution on a real Utah pedigree. (a) Genealogical tree, (b) Utility versus privacy under low tolerance to privacy loss for all relatives except parent P5, and varying values of privacy constraints $\text{Pri}(5, \mathcal{P}_s^5)$ for parent P5 (x-axis). Here, $\mathcal{X}_O = \varnothing$, meaning that no relative has revealed any SNP before P5. Low tolerance is defined as 1/4 of the total privacy loss that a relative would incur if all 50 SNPs of P5 were revealed. Results are shown up to $\text{Pri}(5, \mathcal{P}_s^5) = 4.4$ even if P5's privacy constraint can go beyond because, from $\text{Pri}(5, \mathcal{P}_s^5) = 4$, the utility stops increasing (capped by other relatives' low tolerance).

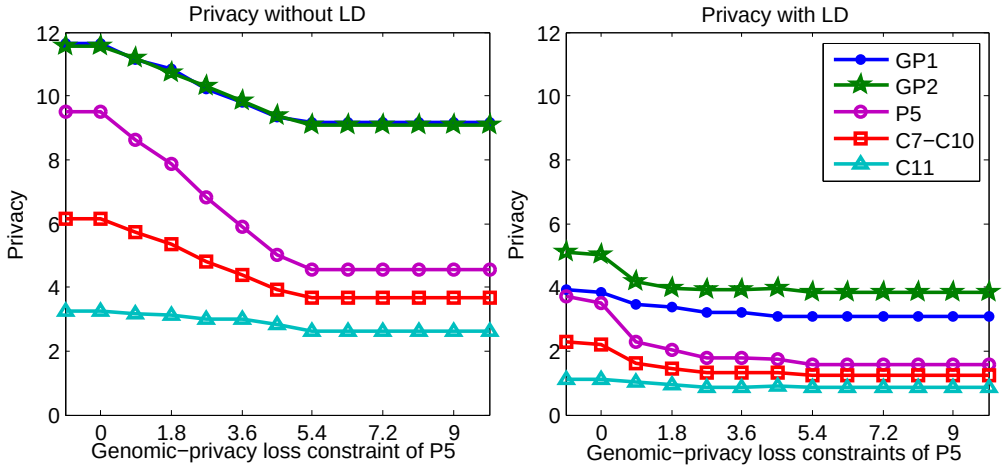


(a) (b)

Figure 5.3: Utility versus privacy under (a) medium, (b) high tolerance to privacy loss for all relatives except parent P5, and varying values of privacy constraints for parent P5 (x-axis). Medium, respectively high, tolerance is defined as around half, respectively 3/4, of the total privacy loss that a relative would incur if all 50 SNPs of P5 were revealed. The x-axis represents the privacy loss constraint of P5, from no privacy loss (strongest constraint) to 9.9 privacy loss (i.e., around 0.2 privacy loss per SNP, which is a weak constraint).

by GP1 and GP2; 35 SNPs revealed by GP3; 42 revealed by GP4; 0 by P6; 0 SNP by C7, C8, C9, C10; and 30 by C11.

We analyze the relation between utility and privacy for different genomic-privacy constraint values, for each of the eleven individuals, $\text{Pri}(i, \mathcal{P}_s^i)$. Figure 5.3(a) and 5.3(b)



(a) (b)

Figure 5.4: Genomic privacy of all family members given the genomic-privacy constraint of P5, under the same setting as in Figure 5.3(a) i.e. under medium privacy loss constraints for P5 relatives: (a) privacy computed without LD, and (b) privacy computed with LD, before the fine-tuning phase. We do not show the privacy levels of GP3, GP4 and P6 as these remain constant. Note the large discrepancy in absolute privacy values and privacy losses between Figure 5.4(a) and 5.4(b). Also notice that GP1 privacy curve is hidden by GP2 privacy curve in Figure 5.4(a) (they have same privacy levels w/o LD).

illustrate the utility gain with respect to different privacy loss tolerance levels for the donor (P5). The two figures differ essentially in terms of the genomic-privacy constraints of the rest of the family members. In Figure 5.3(a), the tolerance is medium; more precisely, the privacy constraint for each individual in the pedigree (except P5) is set to half of the maximum privacy loss that would be incurred by that individual if the donor revealed all his SNPs. In Figure 5.3(a), the tolerance is higher, set to 3/4 of the maximum privacy loss.

We first focus on the utility computed using our branch-and-bound algorithm (case w/o LD). In Figure 5.3(a), we observe that the utility does not increase beyond 38 when we increase the genomic-privacy loss constraint of the donor more than 5.4. From this point, the increased privacy tolerance of the donor does not enable him to reveal more SNPs, because he is constrained by the rest of the family's privacy requirements. In Figure 5.3(b), we note that the utility keeps increasing with the privacy loss constraint of P5 because his relatives are more tolerant regarding their own privacy losses.

Looking at the utility induced once we include the LD correlations in the privacy quantification, we notice some increase in the utility. In other words, including LD enables the donor to reveal more SNPs than without LD. Utility in both curves reaches 50 SNPs after a 4.5 privacy loss constraint for the donor. This can be explained by the fact that, when LD is considered, we use Equation (5.3) (privacy loss with LD) instead of Equation (5.1) (privacy loss without LD) to compute the privacy weights for each SNP in each constraint. And the privacy loss in Equation (5.3) is actually smaller than in Equation (5.1) in this scenario, essentially because LD already decreases significantly the relatives' privacy before the donor reveals any of his own SNPs. This is very visible

in Figure 5.4(a) and 5.4(b). In Figure 5.4(a), we show the privacy levels for any family member when LD is not included in the privacy quantification. Figure 5.4(b) shows the privacy levels when LD correlations are also used in the privacy quantification.

First, we notice that in both figures, it is P5's privacy level that decreases the most, as he is the one who actually reveals new SNPs in the process. Other relatives' privacy is only damaged due to familial correlations. At the origin of the x-axis (i.e., on the y-axis), we see the privacy levels before the donor makes a decision, i.e., before the optimization algorithm. We notice that, here again, privacy without LD is much higher than privacy once LD is used to infer the SNPs. This is because some relatives have already revealed part of their genomic data. This is the reason, once P5 reveals his own SNPs, the privacy loss is much smaller in Figure 5.4(b) than in Figure 5.4(a). As a consequence, the donor (P5) can reveal more SNPs while still meeting his family's privacy constraints, thus leading to the utility increase displayed in Figure 5.3(a) and 5.3(b). We conclude that the values of the privacy-loss constraints have to be carefully determined by the family members or the genetic counsellors, based on family members' privacy expectations and on whether LD is included or not in the initial inference and privacy quantification. In our case, in order to make use of the linear optimization framework, we defined the privacy loss constraints based on the privacy levels computed without LD.

Finally, we compared the optimal solutions computed with exhaustive search over a subset of 10 SNPs whose privacy weights were computed with LD, with the solutions derived from our optimization algorithm presented in Figure 5.1. In the various scenarios we tested, the exhaustive search method could never find higher utility values than our fine-tuning algorithm. In all scenarios, our fine-tuning algorithm reached the maximum utility. Thus, even though we do not have any formal demonstration that the fine-tuning step is optimal, we are confident that in general it provides a very good approximation of the optimum.

5.3.3 Computational Complexity

As expected, the highest computation time is on average induced by the branch-and-bound algorithm (Subsection 5.2.2), due to the high complexity of the multidimensional knapsack problem. The non-linear extension (Subsection 5.2.3) is by design very efficient, as it relies on previous optimal computations and it updates a minimal set of decision variables, trading-off exact optimality for computational efficiency. This last part only requires quantifying privacy levels twice at the beginning (in the quantification box), to get the $\tilde{E}_j^i(\mathbf{y}_D = \mathbf{0})$'s and $\tilde{E}_j^i(\mathbf{y}_D^*)$'s, and then quantifying once per update on a decision variable y_D^{i*}.[5]

The multidimensional knapsack problem is NP-complete and admits no fully polynomial-time approximation scheme. From our experiments, we notice that the complexity of the branch-and-bound algorithm highly differs for different settings, e.g., different privacy-loss tolerance values or privacy weights. With 50 SNPs, the vast majority of the solutions were found in less than one second. However, the algorithm did not scale well for more than 50 decision variables. The positive side is that this whole process has to be undertaken only once by the donor and can be run offline. Furthermore, we considered one privacy constraint for each family member, thus eleven constraints

[5]Note that the computational complexity of one quantification step is $\mathcal{O}(nm)$ (shown in Subsection 3.3.3).

in total. In practice, some relatives would certainly not care much about their genomic privacy, hence some constraints could be relaxed, thus enabling us to consider more SNPs in the optimization problem. Also, an advantage of the branch-and-bound algorithm is that it can be parallelized and distributed using a computer cluster. The algorithm's running time then scales linearly with the number of machines and cores [41]. Another way to reduce the complexity is to cluster subsets of SNPs together (based on the diseases they are associated with, or based on the LD correlations between them), thus trading-off the granularity of the obfuscation mechanism for computational efficiency. Note that our optimization problem can easily be adapted to deal with clusters of SNPs: We can simply define the privacy weight of one cluster as the sum of the privacy losses over the SNPs in this cluster. Finally, instead of using an exact optimization method, heuristic approaches [70] could be used to approximate the optimal solution and improve computational efficiency.

5.4 Related Work

Building upon [86], Sankararaman *et al.* provide quantitative guidelines for researchers willing to make a certain number of SNPs publicly available in GWAS, without revealing the presence of a single individual within a study group [145]. Fienberg *et al.* [64] propose using differential privacy to protect the identities of participants in scientific study. In the same vein, Johnson and Shmatikov [102] propose privacy-preserving algorithms for computing various statistics related to the SNPs, while guaranteeing differential privacy. However, differential privacy reduces the accuracy of research results and is aimed to be applied on aggregate results. In our work, we focus on protecting individual genomic data.

Some works also focus on protecting the privacy of genomic data and on preserving utility in medical tests such as (i) searching of a particular pattern in the DNA sequence [163, 39], (ii) comparing the similarity of DNA sequences [101, 40, 36, 52, 53, 109], and (iii) performing statistical analysis on several DNA sequences [106]. Furthermore, Ayday *et al.* propose privacy-preserving schemes for medical tests and personalized medicine methods that use patients' genomic data [34]. For privacy-preserving clinical genomics, a group of researchers proposes to outsource some costly computations to a public cloud or semi-trusted service provider [168, 47]. Finally, Ayday *et al.* propose techniques for privacy-preserving management of raw genomes [32]. All aforementioned works make use of cryptographic protocols to protect the privacy of genomic data. In this thesis, we propose a non-cryptographic approach for protecting genomic privacy, which has the advantage to be computationally more efficient when making research on genomic data.

Finally, Calmon and Fawaz propose an inference framework for evaluating privacy risks under utility constraints in a generic setting [60]. Their goal is to minimize information leakage subject to certain utility constraints. They show that their optimization problem can be cast as a modified rate-distortion problem. They eventually compare their framework with differential privacy.

5.5 Summary

In this chapter, we convey the importance of building mechanisms for preserving genomic privacy. Such privacy goes beyond the protection of genomic information of the individual to the consideration of the interests of family members. These might be unwilling to allow predictions of their own genomic data based on the leakage of information from one or several individuals of the kin. The approach presented here searches for balance between accuracy (utility) of genomic data and privacy by relying on graphical models and combinatorial optimization. We take into account the fact that different parts of the genome can have different utilities in medical research, and different levels of sensitivity for individuals. Our genomic-privacy preserving mechanism makes use of obfuscation to meet privacy requirements of family members and maximize utility. We also present an extension of the optimization algorithm to cope with non-linear constraints induced by linkage disequilibrium. We implement both linear and non-linear optimization algorithms and evaluate their computational complexities.

Part III

Positive Interdependence and Incentives

Chapter 6

Optimizing Incentives for Cooperative Privacy Protection

6.1 Introduction

Over the last two centuries, non-governmental currencies, known as *scrip*, have been issued by private companies or local communities for many different purposes. For instance, to pay employees in isolated mining or logging camps, company scrip was used in lieu of regular money. More recently, community-issued scrip, such as the Detroit Community Scrip, has been issued in order to restore economic confidence, and help consumers make ends meet [17]. In the last decade, scrip systems have been proposed in order to thwart free riding in online environments (e.g., file sharing or resource sharing [95, 165]). The free-rider problem is particularly serious in peer-to-peer (P2P) networks such as BitTorrent, LimeWire or Gnutella, in which most users (85 percent) do not share any files [88].

Although scrip systems can help ensure fairness and prevent free riding, such systems are exposed to similar behaviors as in real-world economies that lead to the same monetary issues. The Capitol Hill Baby Sitting Co-Op [157], a concrete scrip system created by a group of parents working on Capitol Hill, faced a recession and a monetary crash due to its monetary policy. Several researchers further studied the dynamics of scrip systems, based on these issues [71, 107, 108]. Among other results, they show that agents following threshold strategies lead to a nontrivial Nash equilibrium. They show the impact of the amount of scrip in circulation on the efficiency of the system. In particular, they show that efficiency (social welfare) increases with the average amount of scrip per agent, until some point where the system experiences a monetary crash. At that point, no agent is willing to work anymore and social welfare falls to zero. Finally, they consider different "irrational" behaviors, such as altruists and hoarders, and identify the impact of sybils and collusion on scrip systems.

The original scrip system assumes one transaction at a time, where one agent provides a service to another and gets paid one dollar[1] for it (one-to-one exchange) [71]. Previous work has brought a number of relevant results. However, there is an urgent need to

[1] We refer to the unit of scrip as the dollar.

extend the one-to-one scrip system to a system involving more than one dollar and two agents at a time in order to tackle new challenges led by modern IT systems, such as fostering cooperation in *privacy-enhancing applications*.

Privacy-enhancing technologies, such as anonymity networks [56, 51, 68, 143], provide valuable privacy benefits for Internet users. Among other benefits, anonymity networks can prevent price discrimination in e-commerce by concealing IP addresses. They are also used by journalists or human rights activists to circumvent censorship in dictatorial countries. For instance, there was a dramatic increase of Tunisian Tor [16] users during the Jasmine Revolution in January 2011 [15].

Many privacy-preserving mechanisms require cooperation among multiple users in order to achieve a good level of privacy. However, cooperation is not free, and its inherent cost often prevents users from collaborating. For example, in anonymity systems, running a relay node costs a non-negligible amount of bandwidth and processing power. Back in 2003, Acquisti et al. already highlighted the need of incentives to offer and use anonymity services [20]. Whereas the use of anonymity networks has increased since then, the number of relays is still much lower than the number of clients, and the client-to-relay ratio keeps growing. In 2009, there were 1,500 Tor relays for approximately 100,000 simultaneously active Tor clients [128], in June 2011, 2,500 relays for 300,000 to 400,000 clients, and today, in December 2014, there are 10,500 relays (including bridges) for 2.25 million Tor clients [15].

Among other incentives for acting as a relay in anonymity networks, several schemes propose to make use of micropayments to reward users relaying others' anonymous traffic [27, 48, 65] . These previous works have mainly contributed to the design of anonymous and secure micropayments. However, they did not evaluate the monetary issues that could appear in such systems. Assuming an anonymous circuit requires the cooperation of n relays, each client has to own (at least) n dollars in order to reward each of these n relays. In order to earn enough scrip to afford such a relaying service, each client will then have to serve — relay anonymous traffic — for other users in the anonymity network.

This leads us to define and study the one-to-n scrip system: one agent requests n other agents to fulfill a service and pays each of them one dollar. This scheme also better complies with current file sharing systems, such as BitTorrent, where an agent downloads multiple equal-size *chunks* from different neighboring peers of the torrent. In order to download an entire file and get any utility from it, an agent needs n peers who volunteer to upload their chunks. Thus, he must reward n agents with n dollars.

In this chapter, we develop and study a new analytical model for scrip systems enabling a much wider range of applications. First, we precisely characterize the distribution of scrip in the one-to-n scrip system at equilibrium as a function of n and of the fraction of agents of each type. Second, we prove that, under certain assumptions, there exists a nontrivial Nash equilibrium where all agents play threshold strategies. We study the effect of n on the agents' strategies and the consequent equilibrium and prove that agents' thresholds increase with n. Third, we evaluate the efficiency (social welfare) of the one-to-n scrip system and notice that it tends to decrease when n increases. We show that a system designer can increase the scrip supply in order to offset the loss of efficiency caused by a larger n. This works well up to a point beyond which the system experiences a monetary crash. We show that this critical upper bound increases with n. Finally, we present how our one-to-n scrip system can help to improve fairness and

Table 6.1: List of symbols.

Symbol	Definition
N	Number of agents within the system
$\mathcal{T}$	Set of agents' types
$\mathbf{f}$	Distribution of types
f_t	Fraction of agents of type t
W	Total amount of scrip in the system
m	Average amount of scrip per agent
n	Number of volunteers per request
b_t	Utility an agent gains for having a request satisfied
c_t	Cost of an agent when satisfying one request
δ_t	Rate at which an agent discounts his utility
α_t	Request rate
β_t	Probability that an agent is able to satisfy a request
γ_t	Likelihood to be chosen when an agent volunteers
k_t	Agent's threshold
$\mathbf{k}$	Vector of size $\lvert\mathcal{T}\rvert$ encompassing all k_t's
s_{k_t}	Threshold strategy with threshold equal to k_t
$\mathbf{s_k}$	Strategy profile with agents' thresholds defined by $\mathbf{k}$
$\mathcal{W}$	State space describing the wealth of every agent
$\mathcal{X}$	Markov chain defined on $\mathcal{W}$
$\mathcal{A}$	Set of agents who can afford a service
$\mathcal{V}$	Set of agents who volunteer
M_i^t	Fraction of agents of type t with i dollars
p_u	Probability of earning one dollar
p_d	Probability of having a request satisfied
μ	Fraction of agents at their threshold

efficiency in two privacy-enhancing applications. In particular, we evaluate the amount of scrip that should be allocated into the Tor network to optimize its performance.

6.2 Model

In this work, we consider a scrip system with N agents who interact with each other. We consider a population of agents with different preferences and characteristics. Each agent has a type $t \in \mathcal{T}$, where $\mathcal{T}$ is a finite set of types. The distribution of types is described by $\mathbf{f}$, where the element f_t represents the fraction of agents with type t. The type t of an agent is described by the tuple $t = (b_t, c_t, \delta_t, \alpha_t, \beta_t, \gamma_t)$, whose variables are defined in the rest of this section and in Table 6.1.

At each time slot, one agent is selected proportionally to his request rate α_t to ask for a service. If this agent has at least $\$n$, he can afford a service and request other agents to fulfill this service. In order to have his request fully satisfied, n agents must be able and willing to collaborate. If there are less than n agents able and willing to volunteer, the request cannot be fulfilled, even partially, and the requester gains no utility. The service has to be satisfied in an "atomic" way. An agent is able to satisfy a service with probability β_t, and willing to volunteer depending on his strategy. Moreover, an agent volunteering to provide service is chosen to fulfill another agent's request with likelihood γ_t.

When a service is performed, meaning that n agents fulfill the request of another agent, the requester (of type t_1) obtains some benefit $b_{t_1}(n)$ that is, in most cases, non-decreasing with n (see Section 6.5 for further details on the privacy gain). Each volunteer of type t_2 bears a utility cost c_{t_2} representing, for instance, the usage of bandwidth and processing power in anonymity networks. Thus, when n agents of same type t_2 collaborate with another agent (of type t_1) and satisfy his request, the whole cost is equal to nc_{t_2}, and the system's utility gain is $b_{t_1} - nc_{t_2}$. We assume that $b_{t_1} - nc_{t_2} > 0$, such that social welfare increases when a service is satisfied. The system would otherwise not be viable.

Regarding the monetary reward, an agent providing a service is paid some fixed amount of scrip that we assume is equivalent to \$1. Consequently, a service requester must spend \$$n$ to obtain a service. If the chosen agent does not have enough scrip, no transaction can take place in that time slot and social welfare stagnates. We model the system as an infinite extensive-form game where the total utility of an agent over time is the discounted sum of utilities at each time slot. The total discounted utility of agent i (of type t) is then $U_i = \sum_{\tau=0}^{\infty} \delta_t^{\tau} u_i(\tau)$, where δ_t represents the rate at which an agent of type t discounts utility.

As in the one-to-one scrip system, we assume that prices do not change over time, which allows the agents to know the future monetary cost of their service requests. As the first step towards an extended scrip system, we will consider a payoff-heterogenous population, i.e. b_t, c_t or δ_t might vary but $\alpha_t = \alpha$, $\beta_t = \beta$ and $\gamma_t = \gamma$, for all t. Differences in these parameters should not fundamentally change the game-theoretic results. The one-to-n scrip system can be fully described by ($\mathcal{T}$, $\mathbf{f}$, N, m, n), where m is the average amount of scrip.

6.3 Analytical Results

In this section, we prove the existence of Nash equilibrium when agents make use of threshold strategies. We also show the effect of n on the system, its equilibrium and the agents' strategies. We begin this section by describing the distribution of scrip, which will help us analyze the strategic behaviors of agents, as well as the resulting social welfare in Section 6.4.

6.3.1 Distribution of Scrip

Before analyzing the best strategies and the resulting equilibrium, it is crucial to examine what happens in the system if every agent adopts a predefined category of strategies, called *threshold strategies*. Such a class of strategies is easy to explain. If an agent has too little scrip, he will be willing to work in order to afford service requests later in time, until he reaches a point at which he will feel "wealthy" enough. This threshold represents how much scrip an agent wants to save up for future requests. Let s_k be the strategy where an agent volunteers when he has strictly less than k dollars and defects otherwise. With this definition, s_0 represents the strategy where an agent never volunteers, and s_∞ the strategy where he always volunteers. As threshold strategies depend on the agents' types, we write k_t to represent the threshold adopted by agents of type t. Vector $\mathbf{k}$ encompasses all the k_t's, for all types t, and $\mathbf{s_k}$ is the corresponding strategy profile.

In our analysis, we assume that $W = mN < \sum_t f_t k_t N$, meaning that the total amount of scrip is not too high in order that the system analysis remains interesting. If

$W \geq \sum_t f_t k_t N$, the system would converge to a state where each agent has reached his threshold, and thus does not want to volunteer anymore. We also assume that $m \geq n$. Otherwise, the system would converge to a state where no agent can afford a service, i.e. where all agents own less than n dollars. These two requirements seem reasonable because a system designer should ensure that (i) there are enough scrip in the system such that exchanges can happen, and (ii) there is not too much scrip in order to prevent procrastination and to encourage cooperation among agents.

Let $\mathcal{X}$ be a Markov chain over the state space $\mathcal{W}$ that describes the amount of scrip each agent owns. Each state of the Markov chain can be described by a vector $\mathbf{x}$, where x_i represents the amount of scrip agent i owns in state $\mathcal{W}_\mathbf{x}$. These states must satisfy some constraints: (i) $\sum_{i=1}^{N} x(i) = W$, and (ii) $0 \leq x(j) \leq k_t$, for all agents j with type t.[2] Thus, even if the Markov chain has a significant number of states (when N is large), their number is finite. If the Markov chain is in a state $\mathcal{W}_\mathbf{x}$, and agent j has a request satisfied by n agents i_1, i_2, ..., i_n, the Markov chain moves to another state, $\mathcal{W}_\mathbf{y}$, where

$$\begin{cases} y(j) = x(j) - n & \\ y(i_\ell) = x(i_\ell) + 1, & \text{for } \ell = 1, ..., n, \\ y(.) = x(.), & \text{for all other agents.} \end{cases} \tag{6.1}$$

We can already notice that, contrarily to the original scrip system, the aforementioned Markov chain is neither reversible nor symmetric, notably because no single transaction can restore the chain back to its previous state. Nevertheless, if there are at least $n + 2$ agents within the scrip system, there exists a limit distribution, as stated in the following lemma.

Lemma 6.1. *If there are at least $n + 2$ agents in the system, then $\mathcal{X}$ is finite, aperiodic and irreducible and has a limit distribution.*

Proof. $\mathcal{X}$ is aperiodic. Assume that there are (at least) $n + 2$ agents i_1, i_2, ..., i_{n+2}. Suppose $\mathcal{X}$ is in a state $\mathcal{W}_\mathbf{x}$ where at least one agent has $\$n$ or more and the others have less than their threshold amount of scrip. There must exist such a state by our assumption that m is interesting (i.e. neither too small nor too high). There exists a cycle of length $n + 1$ from state $\mathcal{W}_\mathbf{x}$ to itself: i_2, i_3, ..., i_{n+1} volunteer for i_1, then i_1, i_3, ..., i_{n+1} volunteer for i_2, and so on until i_1, i_2, ..., i_n volunteer for i_{n+1}. There is also a cycle of length $n+2$: i_2, i_3, ..., i_{n+1} volunteer for i_1, then i_1, i_3, i_4, ..., i_n, i_{n+2} volunteer for i_2, then i_1, i_2, i_4, ..., i_{n-1}, i_{n+1}, i_{n+2} volunteer for i_3, and so on until i_2, i_3, ..., i_{n+1} for i_{n+2}.

$\mathcal{X}$ is irreducible. Indeed, a Markov chain is said to be irreducible if all states communicate, or, in other words, if it is possible to reach any state from any other state. For any pair of states i and j of the Markov chain $\mathcal{X}$, we can show that the probability of going from i to j in a finite number of steps is strictly greater than 0, proving that any state is reachable from any other one.

Finally, as the number of states $\mathcal{W}$ is finite, $\mathcal{X}$ is also finite, and thus a limit distribution exists, and it is independent of the state in which the system starts [144]. □

[2] For simplicity, we assume that no one's amount of scrip exceeds their threshold.

We can express the transition probabilities for all pairs of states i and j, $i \neq j$ that are directly reachable from each other[3] as

$$P_{ij} = \frac{1}{|\mathcal{A}|} \cdot \frac{1}{\binom{|\mathcal{V}|-I}{n}}, \tag{6.2}$$

where $\mathcal{A}$ is the set of agents who can afford a service, i.e. who have at least $\$n$, in state i, and $\mathcal{V}$ is the set of volunteers, i.e. agents who have not reached their threshold amount of scrip, in state i too, and I is 1 if the agent requesting the service has an amount of scrip that is under his threshold, and 0 otherwise (because an agent cannot satisfy his own request). The transition probabilities depend on the values $|\mathcal{A}|$ and $|\mathcal{V}|$ that vary among the different states. Thus, the limit distribution is not uniform, even when $n = 1$. Instead of computing this limit distribution, we will focus on the corresponding distribution of scrip, because we are not interested in who has how much scrip, rather in the fraction of people that have a given amount of scrip.

For each state $\mathcal{W}$ of the Markov chain $\mathcal{X}$, there is a distribution of scrip M that describes the fraction of agents for each possible amount of scrip. More precisely, M_i^t represents the fraction of agents of type t who own $\$i$.[4] For instance, if there is only one type of agent and we are in a state $\mathcal{W}$ where money is uniformly distributed ($x(j) = m$ $\forall j$), then $M_m^t = 1$, and all other M_i^t are equal to zero. The distribution of scrip must satisfy two constraints:

$$\sum_t \sum_{i=0}^{k_t} i M_i^t = m \tag{6.3}$$

$$\sum_{i=0}^{k_t} M_i^t = f_t \tag{6.4}$$

First, the average amount of money is equal to m, and second, the fraction of agents playing s_{k_t} is equal to f_t (fraction of agents of type t). One can show that, if N is large, there exists a particular distribution M^* such that, with high probability, the Markov chain $\mathcal{X}$ will almost always be in a state $\mathcal{W}_{\mathbf{x}}$ such that the related distribution of scrip $M^{\mathbf{x}}$ is close to M^*. This kind of convergence around the most likely distribution is known as a *concentration phenomenon* in statistical mechanics [99]. According to Lemma 6.1, we can state that M^* exists. Before characterizing M^*, let us define two matrices B and C of size $(n+1) \times (n+1)$:

$$B = \left[\begin{array}{ccc|c} 1 & 0 & \cdots\ 0 & -\theta_n \\ \hline & & & 0 \\ & \mathbb{I}_n & & \vdots \\ & & & 0 \end{array}\right] \quad C = \left[\begin{array}{ccc|c} 1+\theta_n & 0 & \cdots\ 0 & -\theta_n \\ \hline & & & 0 \\ & \mathbb{I}_n & & \vdots \\ & & & 0 \end{array}\right]$$

where $\mathbb{I}_n$ is the identity matrix of size n, $\theta_n = \frac{1}{\lambda n}$, λ chosen to ensure that (6.3) is satisfied with the distribution M^* defined in the following theorem.

Theorem 6.1. *Given a payoff-heterogenous population, the distribution of scrip in a one-to-n scrip system will converge to*

$$(M^*)_i^t = \frac{f_t \pi_i^t}{\sum_{j=0}^{k_t} \pi_j^t} \tag{6.5}$$

[3] $P_{ij} = 0$ if i and j do not directly communicate with each other.

[4] M_i represents the fraction of agents who own $\$i$, regardless of their type.

where the π_i^t's are defined in the following way:

$$\begin{cases} \mathbf{e}_i^t = B^{n-1-i}C^{k_t-2n+1}\mathbf{v}\pi_{k_t}^t, & \text{if } i \in [0, n-2]; \\ \mathbf{e}_i^t = C^{k_t-n-i}\mathbf{v}\pi_{k_t}^t, & \text{if } i \in [n-1, k_t-n-1]; \\ \pi_i^t = \theta_n(1+\theta_n)^{k_t-i-1}\pi_{k_t}^t, & \text{if } i \in [k_t-n, k_t-1]. \end{cases}$$

$\mathbf{e}_i^t$ and $\mathbf{v}$ are vectors of size $(n+1)$ defined as:

$$\mathbf{e}_i^t = \begin{bmatrix} \pi_i^t \\ \vdots \\ \pi_{i+n}^t \end{bmatrix} \quad \mathbf{v} = \begin{bmatrix} \theta_n(1+\theta_n)^{n-1} \\ \vdots \\ \theta_n(1+\theta_n) \\ \theta_n \\ 1 \end{bmatrix}$$

Proof. Let us focus on one type t and then generalize for all types. Knowing that agents of type t have k_t+1 possible states of wealth (i.e., their amount of scrip can go from 0 to k_t), we can define a Markov chain Y over k_t+1 states that describes the amount of scrip an agent of type t can own. When the Markov chain is in some state, it can either move one state up, or move n states down, or stay in the same state. The probability of moving one state up is

$$P(Y_{i+1}|Y_i) = \frac{n}{|\mathcal{V}|} \tag{6.6}$$

and the probability of moving n states down is

$$P(Y_{i-n}|Y_i) = \frac{1}{|\mathcal{A}|} \tag{6.7}$$

where $\mathcal{A}$ is the set of agents who can afford a service and $\mathcal{V}$ is the set of volunteers.

There is one state from which the Markov chain cannot go up (the state where the agent has k_t+1 dollars), and some states from which Y cannot go down (the states where the agent has less than n dollars). From (6.6) and (6.7), we can express the balance equations for all states:

$$\begin{cases} \frac{1}{|\mathcal{A}|}\pi_i = \frac{n}{|\mathcal{V}|}\pi_{i-1}, & \text{if } i = k_t; \\ \left(\frac{1}{|\mathcal{A}|} + \frac{n}{|\mathcal{V}|}\right)\pi_i = \frac{n}{|\mathcal{V}|}\pi_{i-1}, & \text{if } i \in [k_t-n+1, k_t-1]; \\ \left(\frac{1}{|\mathcal{A}|} + \frac{n}{|\mathcal{V}|}\right)\pi_i = \frac{n}{|\mathcal{V}|}\pi_{i-1} + \frac{1}{|\mathcal{A}|}\pi_{i+n}, & \text{if } i \in [n, k_t-n]; \\ \frac{n}{|\mathcal{V}|}\pi_i = \frac{n}{|\mathcal{V}|}\pi_{i-1} + \frac{1}{|\mathcal{A}|}\pi_{i+n}, & \text{if } i \in [1, n-1]; \\ \frac{n}{|\mathcal{V}|}\pi_i = \frac{1}{|\mathcal{A}|}\pi_{i+n}, & \text{if } i = 0. \end{cases}$$

By multiplying everything by $\frac{|\mathcal{V}|}{n}$ and setting $\lambda = \frac{|\mathcal{A}|}{|\mathcal{V}|}$ (the ratio between $|\mathcal{A}|$ and $|\mathcal{V}|$ is constrained by Equ. (6.3)), we get

$$\begin{cases} \frac{1}{\lambda n}\pi_i = \pi_{i-1}, & \text{if } i = k_t; \\ \left(\frac{1}{\lambda n} + 1\right)\pi_i = \pi_{i-1}, & \text{if } i \in [k_t-n+1, k_t-1]; \\ \left(\frac{1}{\lambda n} + 1\right)\pi_i = \pi_{i-1} + \frac{1}{\lambda n}\pi_{i+n}, & \text{if } i \in [n, k_t-n]; \\ \pi_i = \pi_{i-1} + \frac{1}{\lambda n}\pi_{i+n}, & \text{if } i \in [1, n-1]; \\ \pi_i = \frac{1}{\lambda n}\pi_{i+n}, & \text{if } i = 0. \end{cases}$$

We then set $\theta_n = \frac{1}{\lambda n}$ and get the following recursions that fully describe the Markov chain distribution:

$$\begin{cases} \pi_i = \theta_n \pi_{i+1}, & \text{if } i = k_t - 1; \\ \pi_i = (1+\theta_n)\,\pi_{i+1}, & \text{if } i \in [k_t - n, k_t - 2]; \\ \pi_i = (1+\theta_n)\,\pi_{i+1} - \theta_n \pi_{i+n+1}, & \text{if } i \in [n-1, k_t - n - 1]; \\ \pi_i = \pi_{i+1} - \theta_n \pi_{i+n}, & \text{if } i \in [0, n-2]. \end{cases}$$

We can then express the last $(n+1)$ π_i's (but π_{k_t}) with respect to π_{k_t}:

$$\pi_i = \theta_n (1+\theta_n)^{k_t - i - 1} \pi_{k_t} \ \forall i \in [k_t - n, k_t - 1]. \tag{6.8}$$

From these $n+1$ values, we can build the vector $\mathbf{v}$ that will be used for the calculation of all other probabilities:

$$\mathbf{v} = \begin{bmatrix} \theta_n (1+\theta_n)^{n-1} \\ \vdots \\ \theta_n (1+\theta_n) \\ \theta_n \\ 1 \end{bmatrix} \tag{6.9}$$

Then, we can write

$$\begin{bmatrix} \pi_{k_t - n} \\ \vdots \\ \pi_{k_t - 1} \\ \pi_{k_t} \end{bmatrix} = \mathbf{v} \pi_{k_t} \tag{6.10}$$

As $\forall i \in [n-1, k_t - n - 1]$, $\pi_i = (1+\theta_n)\,\pi_{i+1} - \theta_n \pi_{i+n+1}$, we can build a matrix C of size $(n+1) \times (n+1)$ that will be used for computing these probabilities:

$$C = \left[\begin{array}{cccc|c} 1+\theta_n & 0 & \cdots & 0 & -\theta_n \\ \hline & & & & 0 \\ & & \mathbb{I}_n & & \vdots \\ & & & & 0 \end{array}\right] \tag{6.11}$$

where $\mathbb{I}_n$ is the identity matrix of size n. We can then express, for instance, the (non-normalized) probabilities from state $k_t - n - 1$ to state $k_t - 1$ in the following vectorial form:

$$\begin{bmatrix} \pi_{k_t - n - 1} \\ \vdots \\ \pi_{k_t - 1} \end{bmatrix} = C \begin{bmatrix} \pi_{k_t - n} \\ \vdots \\ \pi_{k_t} \end{bmatrix} = C \mathbf{v} \pi_{k_t} \tag{6.12}$$

By induction, we get the general form:

$$\begin{bmatrix} \pi_{k_t - n - j} \\ \vdots \\ \pi_{k_t - j} \end{bmatrix} = C^j \begin{bmatrix} \pi_{k_t - n} \\ \vdots \\ \pi_{k_t} \end{bmatrix} = C^j \mathbf{v} \pi_{k_t}. \tag{6.13}$$

Thus, we can compute π_i, $\forall i \in [n-1, k_t - n - 1]$:

$$\begin{bmatrix} \pi_i \\ \vdots \\ \pi_{i+n} \end{bmatrix} = C^{k_t - n - i} \mathbf{v} \pi_{k_t} \tag{6.14}$$

Finally, as $\forall i \in [0, n-2]$, $\pi_i = \pi_{i+1} - \theta_n \pi_{i+n+1}$, we build a matrix B of size $(n+1)\times(n+1)$ that will help computing the remaining probabilities:

$$B = \left[\begin{array}{cccc|c} 1 & 0 & \cdots & 0 & -\theta_n \\ \hline & & & & 0 \\ & & \mathbb{I}_n & & \vdots \\ & & & & 0 \end{array}\right] \tag{6.15}$$

We can then express the non-normalized probabilities from state $n-2$ to $2n-2$:

$$\begin{bmatrix} \pi_{n-2} \\ \vdots \\ \pi_{2n-2} \end{bmatrix} = B \begin{bmatrix} \pi_{n-1} \\ \vdots \\ \pi_{2n-1} \end{bmatrix} = BC^{k_t-2n+1}\mathbf{v}\pi_{k_t} \tag{6.16}$$

By induction again, we get the general form:

$$\begin{bmatrix} \pi_{n-1-j} \\ \vdots \\ \pi_{2n-1-j} \end{bmatrix} = B^j \begin{bmatrix} \pi_{n-1} \\ \vdots \\ \pi_{2n-1} \end{bmatrix} = B^j C^{k_t-2n+1}\mathbf{v}\pi_{k_t}. \tag{6.17}$$

Hence, we can compute π_i, $\forall i \in [0, n-2]$,

$$\begin{bmatrix} \pi_i \\ \vdots \\ \pi_{i+n} \end{bmatrix} = B^{n-1-j} C^{k_t-2n+1}\mathbf{v}\pi_{k_t} \tag{6.18}$$

By defining $\mathbf{e}_i = \begin{bmatrix} \pi_i & \cdots & \pi_{i+n} \end{bmatrix}^T$, we get

$$\begin{cases} \mathbf{e}_i = B^{n-1-i}C^{k_t-2n+1}\mathbf{v}\pi_{k_t}, & \text{if } i \in [0, n-2]; \\ \mathbf{e}_i = C^{k_t-n-i}\mathbf{v}\pi_{k_t}, & \text{if } i \in [n-1, k_t-n-1]; \\ \pi_i = \theta_n(1+\theta_n)^{k_t-i-1}\pi_{k_t}, & \text{if } i \in [k_t-n, k_t-1]. \end{cases} \tag{6.19}$$

There just remains to normalize the π_i's to get the distribution of scrip:

$$(M^*)_i = \frac{\pi_i}{\sum_{j=0}^{k_t} \pi_j}. \tag{6.20}$$

By multiplying by the fraction of agents of each type, we get the complete characterization of the distribution of scrip:

$$(M^*)_i^t = \frac{f_t \pi_i^t}{\sum_{j=0}^{k_t} \pi_j^t}.$$

□

We have run simulations of the one-to-n scrip system in order to evaluate how close a real-system limit distribution was to the theoretical limit distribution found in Theorem 6.1. Figure 6.1 illustrates the distribution of scrip in a one-to-n scrip system with 1000 agents with same type, $m = 10$, $k_t = 30$, $n = 5$ and (b_t, c_t, δ_t, α_t, β_t, γ_t) = (1, 0.05, 0.95, 1, 1, 1). The dark bars show the theoretical distribution, whereas the light ones show the averaged distribution of scrip after 10,000 steps in the simulated model. Both

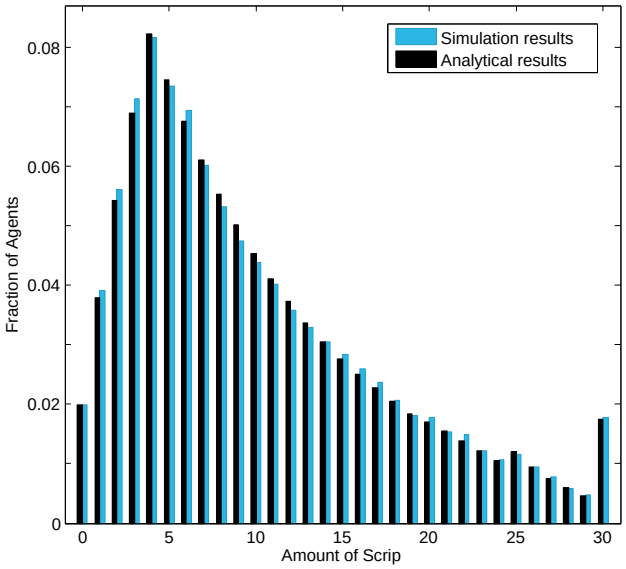


Figure 6.1: Distribution of scrip with $n = 5$ and $k_t = 30$ for 1000 agents. Dark (black) bars represent the theoretical distribution obtained in Theorem 6.1, whereas simulation results (after $10,000$ iterations) are shown in light (blue).

distributions have very similar shapes. This allows us to believe that a real system would converge to some point close to the theoretical limit distribution. Back to the example depicted in Figure 6.1, we notice that both distributions increase until their peaks (at $n - 1 = 4$), and then decrease until a very small peak (at $k_t - n = 25$). We notice a concentration of agents who have reached their threshold (at $k_t = 30$). By doing more simulations with various values of n, we have noticed the maximum of the curves always stands at $(n-1)$ if m remains smaller than half of k_t. This clearly shows how n influences the distribution of scrip.

6.3.2 Game Results: Strategies and Equilibria

In this section, we first analytically verify whether there exist an ϵ-best reply and a consequent nontrivial ϵ-Nash equilibrium in the one-to-n scrip system. Then, we evaluate the effect of n on the agents' strategies and on the Nash equilibrium. In particular, we show to what extent n influences the threshold vector $\mathbf{k}$. These results will help us measure the social welfare in the next section.

Note that δ_t has to be sufficiently large for all types t in order to reach a nontrivial Nash equilibrium where all agents follow a threshold strategy. If the discount factor is so small that it discounts too much future utility, all that matters is present utility and there is no incentive to volunteer now for future benefit. In this case, the only Nash equilibrium (*trivial* one) is to always defect for all agents. Thus, let us assume that $\delta_t > \delta^*, \forall t$. Moreover, all nontrivial Nash equilibria in threshold strategies will be of the form $\mathbf{s_k}$ with $k_t \geq n, \forall t$. Indeed, there is no incentive for a rational agent to volunteer up to $k_t < n$ and then defect, because, in this case, the agent would never be able to afford any service.

In order to analyze the game, we consider a single agent i of type t, from whom point

of view the system can be modeled as a Markov decision process (MDP). If N is large and n reasonably small with respect to N, what agent i does has essentially no effect on the behavior of the system and no great impact on the scrip distribution. We will later see that finding the best reply of agent i to the other agents' strategies is equivalent to finding an optimal policy for his MDP.

Assuming that the distribution of scrip is close to M^* (defined in Theorem 6.1) and all other agents have fixed their thresholds according to $\mathbf{k}$, we can compute two crucial probabilities for the optimal decision of agent i:

(i) p_u, which is the probability of earning a dollar:

$$\frac{|\mathcal{A}| - I}{N}\frac{n}{|\mathcal{V}|} = \left(1 - \sum_t \sum_{j=0}^{n-1} (M^*)_j^t\right) \frac{n}{1 - \sum_t (M^*)_{k_t}^t}$$

(ii) p_d, which is the probability of agent i having a request satisfied, or equivalently, of spending n dollars:

$$\frac{1}{N} Pr\left(|\mathcal{V}| \geq n\right) \cong \frac{1}{N}$$

p_u is the product of two probabilities: (i) the probability that some agent other than i who has n dollars is chosen to make a request, and (ii) the probability that i is the agent chosen to satisfy it. Whereas the first probability decreases a little with n, the second increases linearly with n, and thus p_u increases almost linearly with n. p_d is the probability of agent i will have a request satisfied, which can be approximated to the probability that agent i will be chosen to make a request.[5] This probability only depends on N. However, n will influence the repercussion of p_d because if the agent is chosen to make a request, he will then spend $\$n$.

It follows from [142] that there exists an optimal policy for the MDP of agent i that is a threshold policy. This threshold, k_t, depends on p_u, p_d, b_t, c_t, δ_t, and n. We will prove later the effect of n on k_t. Note here that k_t must be a multiple of n. Indeed, supposing that an agent should decide between a threshold k_t (multiple of n) and a threshold k_t+1, he would choose $k_t + 1$ only if the extra dollar would give him the opportunity to make one more request than with k_t, and gain more benefit in the future. As the agent needs n dollars to pay for a service, the extra dollar will be worth nothing, and eventually wasted. The cost c_t led by this extra dollar will not be compensated by a shorter expected time to make a request, assuming that δ_t is large enough and c_t is non-negligible.

Furthermore, if every other agent is playing a threshold strategy, for all m and $\epsilon > 0$, there exists an optimal threshold policy that is an ϵ-best reply to the strategy profile $\mathbf{s_k}$. This is valid only for $\delta_t > \delta^*$, large N, and n reasonably small with respect to N. Moreover, considering ϵ-best reply formalizes the fact that the optimal policy of the MDP and the best reply are not exactly the same. Indeed, both p_u and p_d are related to agent i's MDP and they slightly differ from the corresponding probabilities of the game. They are only close with high probability, and after some amount of time. For instance, remember that we consider distribution M^* in the MDP, whereas the actual distribution in the game will be close but still different.

Before proving that a nontrivial ϵ-Nash equilibrium exists, we must show that the best reply function is non-decreasing in $\mathbf{k}$. Let $BR_m^t(\mathbf{s_k})$ be the best reply of an agent of type

[5] It is almost sure that n agents will be willing and able to volunteer under our initial assumption that n is reasonably small with respect to N. See Formula (6.23) for more details.

t given an average amount of money equal to m and the strategy profile $\mathbf{s_k}$. $BR^t_m(\mathbf{s_k})$ is non-decreasing in $\mathbf{k}$. First, it can be shown that if $\mathbf{k'} \geq \mathbf{k}$ (i.e., $k'_t \geq k_t$, $\forall t$), then $\sum_{j=0}^{n-1} (M^*)^{t'}_j \geq \sum_{j=0}^{n-1} (M^*)^t_j$ and $(M^*)^{t'}_{k'_t} \leq (M^*)^t_{k_t}$ for all types t. This means that, by increasing the threshold vector, more agents will not be able to afford a service, and fewer agents will reach their threshold. Therefore, with $\mathbf{k'}$, there will be fewer opportunities to earn money and more agents willing to volunteer for those opportunities, meaning that agents will earn money less often. Thus, agents will run out of money sooner. Hence, the utility of earning more scrip will increase, and as a result so well the best reply. We can now prove the existence of a nontrivial Nash equilibrium.

Theorem 6.2. *For $\delta_t > \delta^*$, large N and n reasonably small with respect to N, there exists a nontrivial ϵ-Nash equilibrium where all agents of type t play s_{k_t} for some $k_t = l_t n$, $l_t \in \mathbb{N}$.*

Proof. As the best reply function BR is non-decreasing, Tarski's fixed point theorem ensures that there exist a least and a greatest fixed point [159] that are equilibria. The least fixed point is the trivial equilibrium, and the greatest one can be reached by starting with s_∞ for all agents and using best-reply dynamics [107]. Moreover, if $\delta_t > \delta^*$, there exists a strategy profile $\mathbf{k}$ such that $BR(\mathbf{k}) \geq \mathbf{k}$. Monotonicity ensures that the greatest fixed point $\mathbf{k}^*$ is greater or equal to $\mathbf{k}$, and thus gives a nontrivial equilibrium. Note that n affects the nontrivial ϵ-Nash equilibrium. The higher n is, the further the MDP will be from the actual game. However, we can finely tune ϵ to cope with higher values of n. Moreover, as stated before, the best reply, for all types of agent, is a multiple of n. □

The natural question that arises from the above theorem is: To what extent does n influence k_t, for all types t? We already know that, $\forall t$, k_t must be a multiple of n. In fact, $\mathbf{k}$ increases with $b_t(n)$, thus with n as proved in the following theorem.

Theorem 6.3. *For given values of m, c_t, α_t, β_t, γ_t, and $\delta_t > \delta^*$ for all t, the threshold vector $\mathbf{k}$ is increasing in n. More precisely, if $b_t = b_t(n)$,*

$$\mathbf{k} \sim \Omega(b_t(n)) \tag{6.21}$$

Proof. Let us focus on the threshold $k_t = k$ of a particular agent and generalize it to the threshold vector $\mathbf{k}$. k is defined as the maximum value such that

$$c_t \leq E[\delta_t^{j(k,p_u,p_d)}]b_t \tag{6.22}$$

holds, where $j(k, p_u, p_d)$ is a random variable whose value is the first round in which an agent starting with k dollars, using strategy s_k, has less than n dollars. The expectation is simply the discounted factor that will affect the agent's benefit at round j. First, we know that p_u increases almost linearly with n. Moreover, p_d is independent of n but the effect of being chosen to make a request is linear to n, as the agent will spend n dollars in that case. Thus, the effects of p_u and p_d on $j(k, p_u, p_d)$ approximately compensate each other. Assuming that b_t generally increases with n, the right part of (6.22) will increase with n if k remains unchanged. As c_t is fixed, the increase in b_t allows for the decrease of $E[\delta_t^{j(k,p_u,p_d)}]$ in front of b_t and still satisfy the inequality. As $j(k, p_u, p_d)$ increases in k (the higher the threshold is, the more money we have and the later we go under $\$n$) and $E[\delta_t^j]$ decreases in j, $E[\delta_t^{j(k,p_u,p_d)}]$ decreases in k. Moreover, as δ_t is close to one, $E[\delta_t^{j(k,p_u,p_d)}]$ decreases in $o(j(k, p_u, p_d))$, and so in $o(k)$. Thus, k can be increased with $b_t(n)$, more precisely in $\Omega(b_t(n))$. □

Our results in this subsection show the existence of a nontrivial equilibrium under certain conditions, as well as some properties of this equilibrium. In the next section, we focus on the social welfare and the optimal amount of scrip in the system.

6.4 Social Welfare

In this section, we investigate how much scrip should be allocated in the one-to-n scrip system in order to optimize its performance, and thus social welfare.

A natural question arises when the system is at equilibrium: How good is it? Consider a single transaction involving only agents of type t. If a request is satisfied, social welfare increases by $b_t - nc_t > 0$. If no request is satisfied then no utility is gained. For a utility gain to happen, two events are required: (i) the agent chosen to make a request must have \$$n$, which occurs with probability $1 - \sum_{i=0}^{n-1} M_i$, and (ii) there must be n volunteers able and willing to satisfy the request. If μ is the fraction of agents at their threshold (i.e., the agents who do not want to volunteer), the probability of having at least n volunteers able to satisfy a request is

$$\begin{aligned} P(|\mathcal{V}| \geq n) &= 1 - P(|\mathcal{V}| < n) = 1 - \sum_{i=0}^{n-1} \beta_t^i (1-\beta_t)^{(1-\mu)N-i} \\ &= 1 - (1-\beta_t)^{(1-\mu)N} \cdot \frac{1 - \left(\frac{\beta_t}{1-\beta_t}\right)^n}{1 - \frac{\beta_t}{1-\beta_t}}. \end{aligned} \tag{6.23}$$

Expression $\frac{1-\left(\frac{\beta_t}{1-\beta_t}\right)^n}{1-\frac{\beta_t}{1-\beta_t}}$ goes to 1 if β_t is close to 0 or $n = 1$. This expression grows until infinity if β_t approaches 1. However, this factor is negligible with respect to $(1-\beta_t)^{(1-\tau)N}$ if n is small with respect to N, which is always the case by assumption. As $(1-\beta_t)^{(1-\tau)N}$ converges to 0 for large N or β_t close to 1, the probability of finding n volunteers can be approximated by 1.

The total expected social welfare over all time is then

$$(1 - \sum_{i=0}^{n-1} M_i) \frac{b_t - nc_t}{1 - \delta_t}. \tag{6.24}$$

First of all, social welfare is maximized by minimizing the fraction of agents with less than n dollars. We can make $\sum_{i=0}^{n-1} M_i$ decrease by adding more scrip in the system. Indeed, if N is fixed, by increasing W, and thus m, the number of "poor" agents decreases. Thus, social welfare increases in m. However, social welfare does not increase to infinity with m and, beyond a certain average amount of money m^*, the only Nash equilibrium reached by the one-to-n scrip system is the trivial one, where no agent volunteers. We now evaluate the influence of n on the social welfare.

Theorem 6.4. *For given values of b_t, c_t, δ_t, and $m < m^*$, social welfare of a one-to-n scrip system is decreasing in n.*

Proof. In $(1 - \sum_{i=0}^{n-1} (M^*)_i)(b_t - nc_t)/(1-\delta_t)$, two factors depend on n. First, $(1 - \sum_{i=0}^{n-1} (M^*)_i)$ decreases in n. Indeed, from Theorem 6.1, we can compute that, if $n' > n$, $\sum_{i=0}^{n'-1} (M^*)'_i > \sum_{i=0}^{n-1} (M^*)_i$. Actually this sum increases approximately linearly with n. Second, $(b_t - nc_t)$ clearly decreases in n if b_t remains constant. Consequently, the whole expression decreases in n, and thus social welfare. □

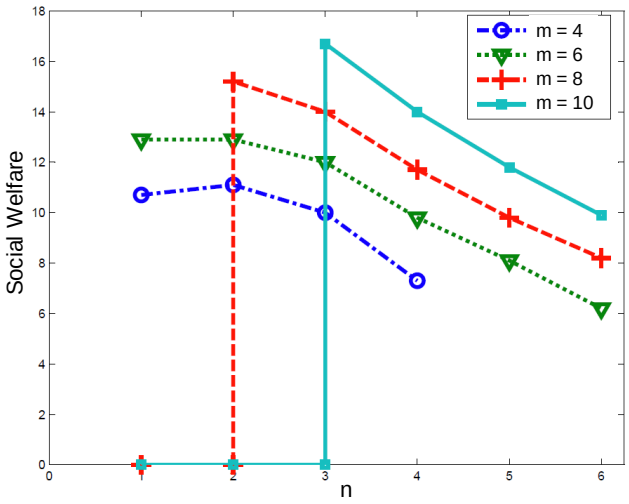


Figure 6.2: Social welfare for various average amounts of scrip m and various n. When m is too large with respect to n, social welfare falls to 0 (monetary crash).

Figure 6.2 shows social welfare with respect to n and m, with the same population used in Figure 6.1. The only change is that now the benefit varies with n: $b_t(1) = 0.7$, $b_t(2) = 0.9$ and $b_t(n) = 1$, $\forall n > 2$. We notice that social welfare tends to decrease with n. The only scenario where it increases slightly is when $m = 4$ and n moves from 1 to 2. In this case, the increase in benefit is greater than the loss in cost and the loss due to agents that cannot afford a service. Note that social welfare falls to 0 when the average amount of money is too high with respect to n (e.g., when $m = 10$ and $n = 1$ or 2).

The fact that social welfare generally decreases with n seems surprising at first sight. Indeed, the more volunteers helping you, the higher the social welfare should be. Thus, the result is counterintuitive. There are two possible explanations for that. First, we must keep in mind that the n volunteers are not optional at all; without them no benefit can be obtained. Moreover, the cost of volunteering c_t does not decrease if more agents volunteer. The cost for each agent remains the same, regardless of n, thus the total cost for the system increases linearly with n. On the contrary, the benefit b_t does not usually increase so much with higher n. We can solve the first issue, or at least decrease its negative impact, by increasing the amount of scrip in the system. Indeed, in Theorem 6.4, we assume a fixed average amount of scrip, whereas a system requiring a higher number of volunteers per request will certainly need more scrip in circulation. This intuition is formalized by the following corollary.

Corollary 6.1. *Assuming all other parameters are fixed, for a certain n, social welfare increases in m. It increases up to a certain average amount of scrip, m_n^*, beyond which there only exists the trivial Nash equilibrium (monetary crash). Furthermore, m_n^* increases in n.*

Proof. The threshold vector $\mathbf{k}$ decreases when m increases, due to best-reply dynamics. Moreover, from the definition of M^* in Theorem 6.1, we can prove that $\sum_{i=0}^{n-1} (M^*)_i$ decreases if $\mathbf{k}$ decreases. Thus, $1 - \sum_{i=0}^{n-1} (M^*)_i$ and social welfare increase if m is increasing. Furthermore, from Theorem 6.3, we know that the threshold vector at equilibrium

$\mathbf{k}$ increases with n. Thus, the threshold vector $\mathbf{k}$ will still decrease when m is increased but will reach zero (trivial equilibrium) beyond higher m with larger n. In other words, the system will bear a higher average amount of money before crashing when n increases. Hence, m_n^* increases in n. □

Figure 6.2 depicts the positive effect of higher m on the social welfare. It also shows that scrip systems with higher n support higher average amount of scrip. For instance, when $m = 10$, the system crashes with $n = 1$ or $n = 2$ but not with $n \geq 3$. The ratio m_n^*/n must not go over a certain value that will be formally defined in future work. The fact that m_n^* is increasing in n can be well explained. When n increases, the agents feel less wealthy if they keep the same threshold values. Indeed, knowing that they then need more dollars to afford a single request, they will certainly be willing to save more dollars for future requests. Thus, if n increases, the agents will stop volunteering later, and thus the system will experience a monetary crash beyond a higher m_n^*. Indeed, a monetary crash appears when agents feel so rich that they are not willing to volunteer anymore. Increasing n prevents such behavior.

6.5 Applications in Privacy Protection

In this section, we present two privacy-enhancing applications where a one-to-n scrip system can help improve fairness and efficiency: (i) anonymity networks [56, 51, 68, 143], and (ii) privacy-preserving data aggregation in participatory sensing [149]. Of course, our one-to-n scrip system could also apply to other privacy-enhancing systems where cooperation is needed, such as mix zones [37, 69] or collaborative sharing of information [151, 146] in location privacy. We focus on the two aforementioned examples because (i) anonymity networks are currently used by hundred of thousands of people to communicate and browse the Web anonymously, and (ii) participatory sensing could provide great benefits to society if there are enough mobile users participating in it, which would be possible only if the privacy of participants is ensured. In both examples, the more users involved in the privacy-preserving system, the higher privacy level the system reaches. Thus, it is absolutely crucial to have as many users as possible. Moreover, it is of the utmost importance that users help each other, i.e. volunteer for each other, in order to preserve the participants' privacy.

6.5.1 Anonymity Networks

Anonymity networks intend to prevent the Internet traffic of individuals from being tracked by governments or websites. As Tor [56] is the most popular anonymity network, we will focus on it for the rest of this section, even though the one-to-n scrip system can be applied to any other anonymity system.

The Tor network is based on onion routing, a design that creates a private network pathway by incrementally building a circuit of encrypted connections through relays (onion routers) on the network. Data packets are repeatedly encrypted (using the relays' public keys) and sent through multiple relays. Then, each relay removes a layer of encryption using its private key (it peels one layer of the onion) to uncover the address of the next relay on the path, and sends the packet to this relay where the same operation is repeated. In this way, no relay ever knows the complete path that a packet has taken. In order to prevent traffic linkability, users must renew their circuits over time. The

Tor project website states that, currently, one circuit can be used for ten minutes [16]. The circuit's path length, i.e. the number of relays in the circuit, is a key parameter in Tor's deployment. As suggested in [56], using one or two hops only would allow for colluding relays to know too easily both the source and destination packets. Thus, the authors recommend to always choose at least three relays per circuit. In the current implementation, Tor selects exactly three relays for each circuit [16].

The lack of relays remains one of the main challenges in anonymity networks [57]. There are currently (December 2014) around 10,500 Tor relays (including bridges) for 2.25 million clients [15]. The corresponding client-to-relay ratio is not likely to decrease if the Tor network does not provide incentives for users to relay others' traffic. Acquisti et al. were the first to formalize the economics of anonymity and propose incentives to encourage users to serve for others [20]. The original Tor proposal already mentioned the need of incentives for a long-term scalable development of such an anonymity system [56]. In the last few years, various incentive mechanisms have been proposed. The first category of incentives is based on differentiated service for Tor users running a relay [136, 98]. The second category proposes to foster participation in traffic relaying by rewarding volunteers with anonymous micropayments [48, 27, 98]. Our idea is that users should reward their Tor relays with the micropayments earned when relaying others' traffic, everybody being involved in the relaying work such as in a P2P anonymity network [143, 68]. In this way, the anonymity workload will be shared among all the users benefiting from the system, thus ensuring fairness and preventing free-riding. Figueiredo et al. provided an anonymous payment-based incentive for such networks [65]. Note that all of the aforementioned micropayment-based incentive mechanisms proposed techniques to distribute coins (scrip) in an anonymous way, in order that privacy gains provided by anonymity networks are not jeopardized. Figure 6.3 depicts an example of this approach. It would ideally reduce the client-to-relay ratio to 1:1, i.e. all Tor clients would eventually run a relay.[6] In order to analyze and evaluate optimal incentives to provide to the anonymity network, we can rely on the one-to-n scrip system.

In current implementation of Tor, n is equal to 3 . This means that, in our scrip model, a Tor user will have to pay \$3 whenever he wants to create a new circuit in the Tor network. It is difficult to evaluate whether the anonymity benefit would increase with a larger n. We know that, with $n = 1$ or 2, the system would be too vulnerable to insider attacks. However, would the level of privacy really increase with n greater than 3? We will consider b_t constant for $n \geq 3$. Different types of benefits can encompass the fact that some users value anonymity more than others. The cost c_t of traffic relaying represents the bandwidth and power consumption used to forward Tor traffic. Different costs can represent various bandwidth or power capabilities of the relays. Assuming that all relays are of type t, the total cost of one request is equal to nc_t. We notice that the cost induced by one anonymity service request is increasing linearly with n, whereas the anonymity benefit remains more or less constant as soon as it reaches an acceptable value for n. Hence, the system designer should keep n small to keep the relays' costs acceptable, and thus optimize social welfare. It is certainly a reason why Tor designers have chosen to set $n = 3$.

Concerning the other scrip system parameters, α_t, β_t, γ_t, they can model different

[6] In order to not discriminate Tor clients that cannot run a relay due to censorship, such as in China [18], we will make some exceptions and let such users benefit from Tor service for free. Indeed, denying anonymity to clients in censored regions would go against Tor's praiseworthy aim.

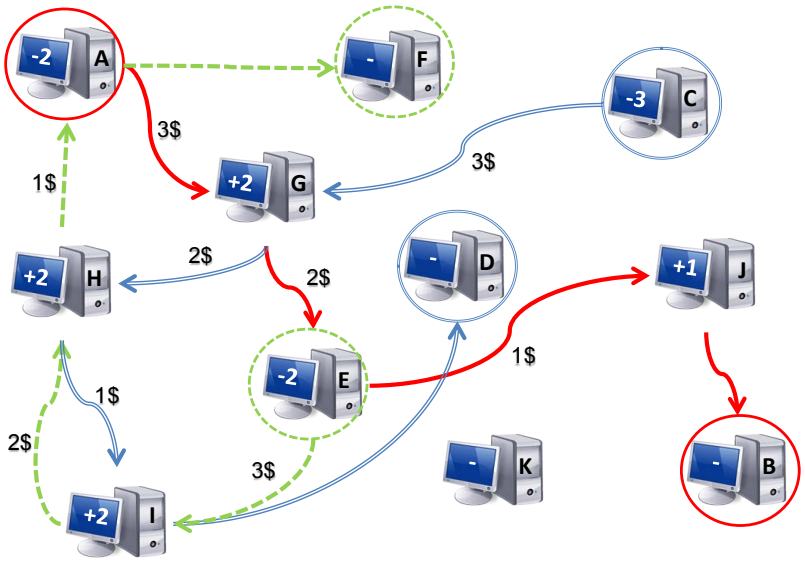


Figure 6.3: Example of an anonymity network with 3 circuits (AGEJB, EIHAF and CGHID) and 3 relays per circuit. Each user includes \$3 for each circuit he wants to build for anonymous communication. Then, each peer that accepts to relay traffic is rewarded with \$1 and can use it for his own relay requests.

behaviors and characteristics of the Tor users. First, α_t represents the request rate. Users surfing the Web more often in an anonymous way will request more service from Tor relays, thus α_t will increase. Second, a Tor agent might not be able to satisfy a traffic relaying request, which is encompassed in the value β_t. Finally, some Tor relays can have higher bandwidth than others, thus a higher quality of service when relaying traffic, or be well-known and more used than others. This could be represented by γ_t, which is the likelihood that an agent is chosen when he volunteers.

We have run simulations of an anonymity network with $N = 300,000$, $n = 3$ and the same homogenous population used in Figure 6.1, i.e. $b_t = 1$, $c_t = 0.05$, $\delta_t = 0.95$, and $\alpha_t = \beta_t = \gamma_t = 1$. Apart from N and n, the simulation parameters are not easy to determine and we plan to further investigate these in future work. Under these settings, the social welfare is maximized at $m = 10 < m_3^*$. With this average amount of money, there is only 2.5% of agents who cannot afford a service (i.e., with less than \$3). We conclude that a system designer should allocate $m \cdot N = 3$ million dollars within an anonymity networks of 300,000 users in order to optimize its efficiency.[7]

6.5.2 Privacy in Participatory Sensing

Participatory sensing is an example of novel mobile systems that leverage new capabilities in computation, communication, storage and sensing of mobile devices [43]. In participatory sensing, mobile users share sensing information, possibly including personal and/or location data, with service providers. However, the emergence of such people-centric systems leads to many issues, among which privacy is one of the most critical. Mobile

[7] Figure 6.2 provides more results on the social welfare for different values of n and m.

users would certainly be willing to share sensing data, e.g. to help monitoring urban air pollution [87], but not at any cost to their privacy.

Shi et al. have recently proposed a concrete privacy-preserving data aggregation scheme for participatory sensing [149]. In this privacy-preserving mechanism, mobile nodes rely on their nearby peers to "hide" their data from the aggregation server (or service provider) that could be malicious (or at least curious), thus requiring cooperation from mobile nodes in their vicinity. Figure 6.4 depicts an example of this privacy-preserving scheme. First, each sensing node[8] slices its data into $n+1$ pieces, sends n encrypted pieces to neighbor nodes and retains the last piece. Second, the mobile nodes receiving pieces of data from other nodes aggregate all received pieces of data before transmitting them towards the aggregation server. Assuming that a fraction R of mobile nodes are malicious and can collude with the aggregation server, the normalized level of privacy is proportional to

$$P = \max\{1 - R^n - R^{|S|-1}, 0\} \tag{6.25}$$

where $|S|$ is the number of sensing nodes, which is equal to N if we assume that all users participating in the privacy-preserving scheme are also sensing nodes. Thus, the level of privacy increases with N, but also with n. However, this privacy-enhancing technique induces significant communication and computation overhead that also increases with n. As battery consumption is, with privacy, one of the main concerns of mobile users in participatory sensing, these communication and computation costs might prevent participants from volunteering to cover other nodes' data, thus threatening the whole privacy-preserving system. In order to foster cooperation and prevent free-riding, we propose to reward with scrip the mobile nodes that volunteer, and to rely on the one-to-n scrip system to optimize the efficiency of the monetary incentive.

First, contrarily to Tor networks, the value n is not at all defined in the initial proposal [149]. The system designer can tune this value to increase the privacy level provided by the mechanism, at the cost of communication overhead. Therefore, we do not attach any fixed value to n. Note that n should remain reasonably small with respect to the number of mobile nodes in the system in order for our theoretical results to apply. This will certainly be the case as the sensing nodes requesting help from others also suffer from too high communication overhead when they send their slices to too many neighbors. Thus, they will cap the number of "cover nodes" by themselves.

The benefit b_t that a sensing node (of type t) gains when a request is satisfied is related to the privacy utility it gains. As Equation (6.25) shows, b_t is dependent on n. Furthermore, as R is smaller or equal than one, $b_t(n) \propto P$ increases with n. Moreover, different types of benefits can encompass the fact that some agents are more privacy cautious and sensitive than others. The cost of volunteering is equal to c_t for all nodes of type t. This cost represents the communication and computation overheads that lead to higher battery consumption. The type of c_t can represent the fact that some users are less willing to consume their battery or merely that their battery has a shorter lifetime. In conclusion, we clearly notice that the cost of one privacy-preserving request is increasing linearly with n, whereas the privacy benefit is increasing with n, but less than linearly. Hence, even if the requester gets higher payoff if he can send more data slices to more neighbors, the overall utility of the system, social welfare, is decreased.

[8] A sensing node refers to any agent who uses his mobile device to sense his environment and submits sensing data.

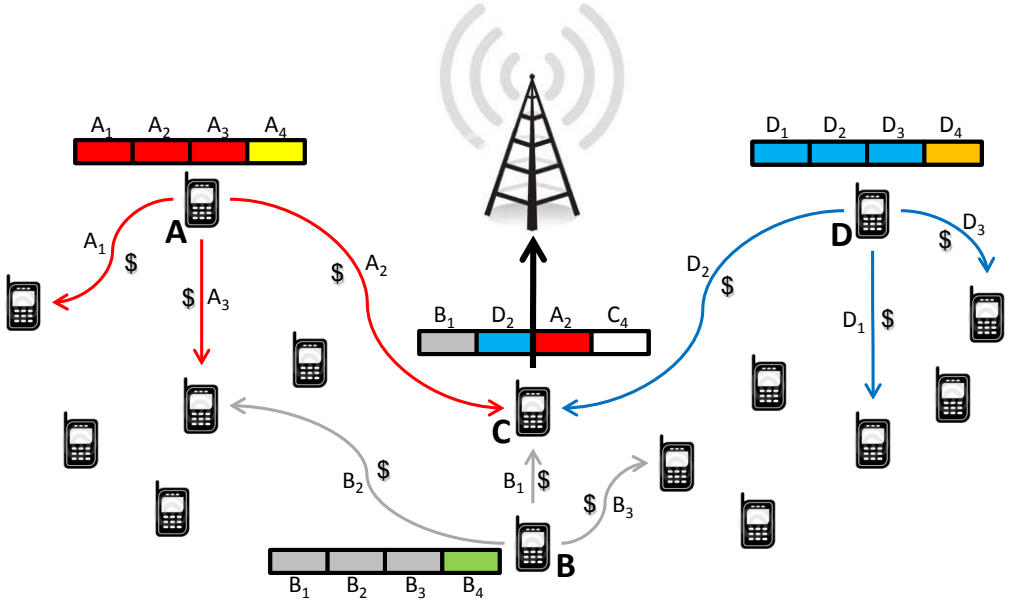


Figure 6.4: Example of privacy-preserving data aggregation in participatory sensing. Each mobile node sends data slices to neighbors in order to mix them. For encouraging cooperation, each node includes a fixed amount of scrip in all data slices.

The sensing nodes can have different amount of sensing data to submit to the aggregation server. This can be well described by the request rate α_t. Indeed, if nodes are collecting and submitting more data, they will request help of nearby peers more often. Furthermore, an agent might be unable to satisfy a request. For instance, its device can run out of battery or he can have a call at the same time. This can be encompassed in β_t. Finally, a node can be asked for covering others' data slices more often than others. For example, an agent can spend more time than another in a neighborhood with higher density of mobile sensing nodes. This difference can be represented by the likelihood that an agent is chosen when he volunteers, γ_t. As a concluding remark, we must mention that the number of data slices n a sensing node can send is also dependent on the density of the nodes in its vicinity. Thus, the optimal choice of n does not only depend on the nodes' privacy sensitivity, but also on the network density constraints.

We have also run experiments for participatory sensing systems, with various values of N and n. For $N = 1000$ and $n = 6$ and the same type of agents than for the previous application, social welfare is maximized with $m = 16$. This value is very close to m_6^* over which the system crashes. This average amount of scrip counterbalances the large value of n very well. It leads to almost the same percentage of agents who cannot afford a service (agents with less than \$6) than in Tor example with $n = 3$ (around 2.5%). Hence, in this scenario, a system designer should allocate $m \cdot N = 16,000$ dollars within the system to optimize its efficiency.

6.6 Summary

In this chapter, we have proposed the first scrip system model that is able to tackle economic systems where one agent needs multiple volunteers simultaneously in order to have his request satisfied. For the novel one-to-n scrip system, we have proved that decisions agents make, based on threshold strategies, lead to ϵ-Nash equilibrium. Assuming that all agents of the system use threshold strategies, we have shown that the limit distribution towards which our scrip system will converge highly depends on n. Simulations of the one-to-n scrip system confirm this convergence. We have studied the effect of n on all results, notably on the agents' strategies, on the social welfare and on the maximum amount of scrip that the system can handle before crashing. We have proved that, at equilibrium, the agents increase their thresholds if n increases. However, in this case, social welfare decreases, which can be partially resolved by adding more scrip in the system. This is possible because the maximum average amount of scrip that the system can bear before it crashes increases with n. Finally, we have shown that our upgraded scrip system can help improve fairness and efficiency in two privacy-enhancing applications where cooperative volunteers are required. We have notably evaluated the average amount of scrip per agent that should be allocated into the Tor network to optimize its performance and fairness.

In future work, it would be interesting to investigate agents' possible strategies other than thresholds. Furthermore, non-standard behaviors such as altruism or hoarding could be studied. These behaviors should not necessarily be considered as irrational: (i) altruists can benefit from providing help to others, and (ii) hoarders may get some utility from owning more scrip. Finally, newcomers and their effect on the amount of scrip in circulation could be evaluated. As a consequence, variable prices could also be considered in the model.

Chapter 7

Conclusion

In this thesis, we have studied the impact on privacy of one of the fundamental characteristics of humanity: interdependence between individuals. We evaluate how other people can compromise our privacy, in both the social and biological dimensions of our identity. We complete this thesis by studying scenarios where others can play a beneficial role in privacy. This is typically the case in today's most popular privacy-enhancing tool, Tor, used by millions of people around the world.

In Part I, we have shown that an external adversary can efficiently find most users in an OSN by exploiting information publicly disclosed by other OSN users. Our navigation attack discovers two thirds of the targeted users in Facebook, and 59% of targeted users in Google+, in a median number of crawled users smaller than 400, respectively 300. This suggests that it is almost impossible for any participant in an online social network to "hide in the crowd" by excluding his name from the central directory. One main reason for failed cases is the privacy behaviors of the targets' friends: the fewer friends who have public attributes and public social links, the less likely the target is to be found. This demonstrates the crucial role of social ties for privacy in OSNs. Another finding of our research is that OSN-membership privacy cannot be ensured because of other people's weak privacy settings.

In Part II, we first formally quantify genomic privacy of individuals in a family by relying upon Bayesian inference. In order to efficiently compute the posterior distributions resulting from the observation of some genomic data, we make use of probabilistic graphical models and belief propagation. In order to express genomic privacy, we propose different metrics widely used and recognized by the privacy research community that represent: (i) the (in)correctness of the adversary's estimation, and (ii) the (un)certainty in the adversary's estimation. In order to obtain more tangible metrics, we go one step further by quantifying health privacy of individuals. We evaluate our quantification framework on real genomic data, and show the scale of the threat by matching genomic profiles with OSN profiles publicly available online. Our results notably show that, by disclosing only 10% of its genomic data, a family can lose more than 50% of its global genomic privacy. We evaluate the interplay between family members with non-cooperative behaviors regarding the privacy of their genomic data. We derive closed-form expressions to measure genomic privacy, as well as closed-form expressions of the Nash equilibria, in the two-player context. We rely upon multi-agent influence diagrams in order to tackle

the computational complexity of finding a Nash equilibrium in the more general n-player setting. Our results notably show that misaligned perceived benefits in genome sharing can create externalities that negatively affect genomic privacy. In the n-player scenario, we notice that, when the perceived benefits do not clearly outweigh the genomic-privacy losses, some players with similar sharing benefits might end up with different strategies at equilibrium. Moreover, we have demonstrated that, although altruism tends to reduce the indirect genomic-privacy losses at equilibrium, it can lead to a social outcome worse than with a purely selfish behavior for specific values of sharing benefits. Finally, we have proposed, in Chapter 5, an obfuscation mechanism for balancing the utility of genome sharing with the privacy expectations of family members. Our privacy-preserving mechanism takes into account the fact that different parts of the genome can have different utilities, and different levels of sensitivity for individuals.

In Part III, we have studied cooperative privacy-enhancing technologies where other users are actually needed to provide any privacy provision, thus showing that interdependence can also be beneficial for privacy. In particular, we have shown how monetary incentives can be used and their amount optimized in order to foster cooperation between users and to improve fairness and efficiency of the privacy-enhancing systems. We have proved that decisions based on threshold strategies lead to a stable equilibrium. We have also derived the maximum amount of money that the system can bear before crashing, which happens to be the optimal amount of money for maximizing the efficiency of the system. We have studied the impact of various parameters on our analytical results, notably the effect of the number of cooperative agents needed in order to achieve privacy for others. Finally, we have demonstrated the practical utility of our findings by deriving the average amount of money that should be allocated into the Tor network in order to optimize its performance and fairness.

In conclusion, we have clearly established the negative effect of interdependence on privacy at both social and biological levels. This demonstrates that privacy is almost impossible to achieve individually and independently of other people. We hope this thesis will help raise awareness about interdependent privacy risks and encourage people to take more into account other people's privacy needs when making their own privacy decisions. This is probably the only way privacy can be enhanced for the time being. We also propose countermeasures and mechanisms for mitigating the negative impact of others' behavior on privacy. We emphasize, however, that these privacy-preserving mechanisms require some degree of cooperation between interdependent individuals. As cooperation is a strong assumption in a society where self-interest has been raised to the status of way of life, we should consider that online privacy is vanishing, and urgently begin designing new mechanisms for the upcoming post-privacy era. Such mechanisms should, in particular, prevent the misuse of personal data and enforce fair decision-making, in order to eventually build a more equitable society.

Bibliography

[1] http://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html. [cited at p. 2]

[2] http://opensnp.org/. [cited at p. 2, 29, 44]

[3] http://www.personalgenomes.org/. [cited at p. 2]

[4] http://opensnp.wordpress.com/2011/11/17/first-results-of-the-survey-on-sharing-genetic-information/. [cited at p. 2, 29, 50]

[5] http://www.vox.com/2014/9/9/5975653/with-genetic-testing-i-gave-my-parents-the-gift-of-divorce-23andme. [cited at p. 2, 29]

[6] http://blog.23andme.com/ancestry/the-uniqueness-of-ashkenazi-jewish-ancestry-is-important-for-health/. [cited at p. 2]

[7] http://gizmodo.com/google-has-most-of-your-email-even-if-you-dont-use-gma-1577324127. [cited at p. 3]

[8] https://www.23andme.com/welcome/. [cited at p. 29, 44]

[9] https://www.genepartner.com/. [cited at p. 29]

[10] http://www.nytimes.com/2013/03/24/opinion/sunday/the-immortal-life-of-henrietta-lacks-the-sequel.html?pagewanted=all. [cited at p. 29, 69]

[11] http://www.ncbi.nlm.nih.gov/projects/SNP/. Visited on 6-Feb-2013. [cited at p. 31, 70]

[12] http://www.eupedia.com/genetics/medical_dna_test.shtml. Visited on 9-Aug-2013. [cited at p. 31]

[13] http://www.patientslikeme.com/. Visited on 9-Aug-2013. [cited at p. 44]

[14] http://www.nytimes.com/2013/08/08/science/after-decades-of-research-henrietta-lacks-family-is-asked-for-consent.html?pagewanted=all. [cited at p. 69]

[15] Tor metrics portal. http://metrics.torproject.org. [cited at p. 86, 100]

[16] Tor project. https://www.torproject.org. [cited at p. 86, 100]

[17] Communities print their own currency to keep cash flowing. USA Today, April 2009. [cited at p. 85]

[18] Tor partially blocked in China, October 2009. https://blog.torproject.org/blog/tor-partially-blocked-china. [cited at p. 100]

[19] A. Acquisti. An experiment in hiring discrimination via online social networks. Berkeley, April 2012. [cited at p. 9]

[20] A. Acquisti, R. Dingledine, and P. Syverson. On the economics of anonymity. In *Financial Cryptography*, pages 84–102. Springer, 2003. [cited at p. 67, 86, 100]

[21] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006. [cited at p. 22]

[22] A. Acquisti, R. Gross, and F. Stutzman. Faces of facebook: Privacy in the age of augmented reality. *BlackHat USA*, 2011. [cited at p. 16]

[23] L.A. Adamic, R.M. Lukose, A.R. Puniyani, and B.A. Huberman. Search in power-law networks. *Physical review E*, 64, 2001. [cited at p. 15]

[24] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255. ACM, 2001. [cited at p. 39]

[25] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6, 2012. [cited at p. 10]

[26] R. Anderson and T. Moore. The economics of information security. *Science*, 314(5799):610–613, 2006. [cited at p. 49]

[27] E. Androulaki, M. Raykova, S. Srivatsan, A. Stavrou, and S. Bellovin. Par: Payment for anonymous routing. In *Privacy Enhancing Technologies*, 2008. [cited at p. 86, 100]

[28] E. Anshelevich, A. Dasgupta, J. Kleinberg, E. Tardos, T. Wexler, and T. Roughgarden. The price of stability for network design with fair cost allocation. *SIAM Journal on Computing*, 38(4):1602–1623, 2008. [cited at p. 60]

[29] E. Ayday, E. De Cristofaro, G. Tsudik, and J. P. Hubaux. The chills and thrills of whole genome sequencing. *arXiv:1306.1264*, 2013. [cited at p. 29]

[30] E. Ayday, A. Einolghozati, and F. Fekri. BPRS: Belief propagation based iterative recommender system. *IEEE ISIT*, 2012. [cited at p. 33]

[31] E. Ayday and F. Fekri. Belief propagation based iterative trust and reputation management. *IEEE Transactions on Dependable and Secure Computing*, 9(3), 2012. [cited at p. 33]

[32] E. Ayday, J. L. Raisaro, U. Hengartner, A. Molyneaux, and J. P. Hubaux. Privacy-preserving processing of raw genomic data. *DPM 2013*, 2013. [cited at p. 81]

[33] E. Ayday, J. L. Raisaro, and J. P. Hubaux. Protecting and evaluating genomic privacy in medical tests and personalized medicine. *WPES '13: Proceedings of ACM Workshop on Privacy in the Electronic Society*, 2013. [cited at p. 69]

[34] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J. P. Hubaux. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. *HealthTech*, 2013. [cited at p. 81]

[35] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference*, 2011. [cited at p. 20, 24]

[36] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. *CCS*, 2011. [cited at p. 69, 81]

[37] Alastair R. Beresford and Frank Stajano. Mix zones: User privacy in location-aware services. In *PerSec*, 2004. [cited at p. 99]

[38] G. Biczók and P. H. Chia. Interdependent privacy: Let me share your data. In *Financial Cryptography and Data Security*, pages 338–353. Springer, 2013. [cited at p. 67]

[39] M. Blanton and M. Aliasgari. Secure outsourcing of DNA searching via finite automata. *DBSec*, 2010. [cited at p. 81]

[40] F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls. Privacy-preserving matching of DNA profiles. Technical report, 2008. [cited at p. 81]

[41] M. Budiu, D. Delling, and R. F. Werneck. DryadOpt: Branch-and-bound on distributed data-parallel execution engines. In *Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International*, pages 1278–1289. IEEE, 2011. [cited at p. 81]

[42] J. T. Burdick, W.-M. Chen, G. R. Abecasis, and V. G. Cheung. In silico method for inferring genotypes in pedigrees. *Nature genetics*, 38(9):1002–1004, 2006. [cited at p. 48]

[43] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M.B. Srivastava. Participatory sensing. In *World Sensor Web Workshop*, 2006. [cited at p. 101]

[44] C. A. Cassa, B. Schmidt, I. S. Kohane, and K. D. Mandl. My sister's keeper?: genomic research and the identifiability of siblings. *BMC Medical Genomics*, 1(1):32, 2008. [cited at p. 47]

[45] A. Chaabane, G. Acs, and M.A. Kaafar. You are what you like! Information leakage through users' interests. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium*, 2012. [cited at p. 10, 14, 15, 22]

[46] J. Chen, A. Dholakia, E.Eleflhetiou, M. Fossotier, and X.-Y. Hu. Near optimum reduced-complexity decoding algorithm for LDPC codes. *IEEE ISIT*, 2002. [cited at p. 38]

[47] Y. Chen, B. Peng, X. Wang, and H. Tang. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. *NDSS*, 2012. [cited at p. 81]

[48] Y. Chen, R. Sion, and B. Carbunar. XPay: Practical anonymous payments for Tor routing and other networked services. In *Proceedings of the 8th ACM workshop on Privacy in the electronic society*, 2009. [cited at p. 86, 100]

[49] L. A. Cutillo, R. Molva, and T. Strufe. Safebook: A privacy-preserving online social network leveraging on real-life trust. *Communications Magazine, IEEE*, 47(12):94–101, 2009. [cited at p. 22]

[50] G. Danezis and E. De Cristofaro. Simpler protocols for privacy-preserving disease susceptibility testing. In *GenoPri*, 2014. [cited at p. 69]

[51] G. Danezis, R. Dingledine, and N. Mathewson. Mixminion: Design of a type III anonymous remailer protocol. In *Security and Privacy, 2003. Proceedings. 2003 Symposium on*, 2003. [cited at p. 86, 99]

[52] E. De Cristofaro, S. Faber, P. Gasti, and G. Tsudik. Genodroid: Are privacy-preserving genomic tests ready for prime time? *Proceedings of the ACM workshop on Privacy in the electronic society - WPES*, pages 97–108, 2012. [cited at p. 50, 81]

[53] E. De Cristofaro, S. Faber, and G. Tsudik. Secure genomic testing with size-and position-hiding private substring matching. In *Proceedings of the 12th ACM Workshop on privacy in the electronic society*, pages 107–118. ACM, 2013. [cited at p. 81]

[54] R. Dey, C. Tang, K. Ross, and N. Saxena. Estimating age privacy leakage in online social networks. In *Proceedings of the 33rd Annual IEEE International Conference on Computer Communications*, pages 2836–2840, 2012. [cited at p. 10, 12, 23, 66]

[55] C. Diaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In *Privacy Enhancing Technologies*, pages 54–68. Springer, 2003. [cited at p. 39]

[56] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *Proceedings of 13th USENIX Security Symposium*, 2004. [cited at p. 4, 86, 99, 100]

[57] R. Dingledine, N. Mathewson, and P. Syverson. Challenges in deploying low-latency anonymity (draft). Technical report, 2005. [cited at p. 100]

[58] P.S. Dodds, R. Muhamad, and D.J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 2003. [cited at p. 14, 15, 23]

[59] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, et al. Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. *Science*, 327(5961):78–81, 2010. [cited at p. 40, 57, 77]

[60] F. du Pin Calmon and N. Fawaz. Privacy against statistical inference. In *Allerton*. IEEE, 2012. [cited at p. 81]

[61] D. Eppstein, M.T. Goodrich, M. Löffler, D. Strash, and L. Trott. Category-based routing in social networks: Membership dimension and the small-world phenomenon. *Arxiv preprint arXiv:1110.4499*, 2011. [cited at p. 24]

[62] Y. Erlich and A. Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014. [cited at p. 29]

[63] D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics (4th Edition)*. Addison Wesley Longman, Harlow, Essex, UK, 1996. [cited at p. 33]

[64] S. E. Fienberg, A. Slavkovic, and C. Uhler. Privacy preserving GWAS data sharing. *Proceedings of the IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, Dec. 2011. [cited at p. 69, 81]

[65] D. Figueiredo, J. Shapiro, and D. Towsley. Incentives to promote availability in peer-to-peer anonymity systems. In *13th IEEE International Conference on Network Protocols*, 2005. [cited at p. 86, 100]

[66] A. Finder. For some, online persona undermines a résumé. *The New York Times*, 2006. [cited at p. 9]

[67] M. Fishelson and D. Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18(suppl 1):S189–S198, 2002. [cited at p. 47]

[68] M.J. Freedman and R. Morris. Tarzan: A peer-to-peer anonymizing network layer. In *CCS*, 2002. [cited at p. 86, 99, 100]

[69] J. Freudiger, M. H. Manshaei, J.-P. Hubaux, and D. C. Parkes. On non-cooperative location privacy: A game-theoretic analysis. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 324–337. ACM, 2009. [cited at p. 67, 99]

[70] A. Fréville. The multidimensional 0–1 knapsack problem: An overview. *European Journal of Operational Research*, 155(1):1–21, 2004. [cited at p. 72, 77, 81]

[71] E. J. Friedman, J. Y. Halpern, and I. A. Kash. Efficiency and nash equilibria in a scrip system for P2P networks. In *Proceedings of the 7th ACM conference on Electronic commerce*, 2006. [cited at p. 85]

[72] J. Gitschier. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *Am. J. Hum. Genet.*, 84:251–258, 2009. [cited at p. 47]

[73] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29:1872–1892, 2011. [cited at p. 14]

[74] N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song. Evolution of social-attribute networks: measurements, modeling, and implications using Google+. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 131–144. [cited at p. 20]

[75] R. Griffiths. 5 important Facebook Timeline privacy settings. http://www.freshtechtips.com/2012/01/facebook-timeline-privacy-settings.html, January 2012. [cited at p. 9]

[76] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2005. [cited at p. 22]

[77] J. Grossklags, B. Johnson, and N. Christin. The price of uncertainty in security games. In *Economics of Information Security and Privacy*, pages 9–32. Springer, 2010. [cited at p. 51]

[78] P. Gundecha, G. Barbier, and H. Liu. Exploiting vulnerability to secure user privacy on a social networking site. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 511–519, 2011. [cited at p. 15, 22]

[79] P. Gundecha, G. Barbier, and H. Liu. Exploiting vulnerability to secure user privacy on a social networking site. In *KDD*. ACM, 2011. [cited at p. 44]

[80] S. Gurses. *Multilateral privacy requirements analysis in online social network services*. PhD thesis, KU Leuven, 2010. [cited at p. 21]

[81] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science: 339 (6117)*, January 2013. [cited at p. 29, 47, 49, 69]

[82] M. Hachman. Facebook now totals 901 million users, profits slip. http://www.pcmag.com/article2/0,2817,2403410,00.asp, April 2012. [cited at p. 17]

[83] P.E. Hart, N.J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4:100–107, 1968. [cited at p. 10, 12]

[84] J. He, W. Chu, and Z. Liu. Inferring privacy information from social networks. *Intelligence and Security Informatics*, pages 154–165, 2006. [cited at p. 22]

[85] B. Henne, C. Szongott, and M. Smith. SnapMe if you can: Privacy threats of other peoples' geo-tagged media and what we can do about it. In *Proceedings of the sixth ACM conference on Security and privacy in wireless and mobile networks*, pages 95–106. ACM, 2013. [cited at p. 66]

[86] N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4, Aug. 2008. [cited at p. 47, 69, 81]

[87] R. Honicky, E.A. Brewer, E. Paulos, and R. White. N-smarts: networked suite of mobile atmospheric real-time sensors. In *Proceedings of the second ACM SIGCOMM workshop on Networked systems for developing regions*, 2008. [cited at p. 102]

[88] D. Hughes, G. Coulson, and J. Walkerdine. Free riding on Gnutella revisited: The bell tolls? *IEEE Distributed Systems Online*, 6(6), 2005. [cited at p. 85]

[89] M. Humbert, E. Ayday, J. P. Hubaux, and A. Telenti. Addressing the concerns of the Lacks Family: Quantification of kin genomic privacy. *CCS '13: Proceedings of 20th ACM Conference on Computer and Communications Security*, 2013. [cited at p. 5]

[90] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Reconciling privacy with utility in genomics. In *Proceedings of the ACM Workshop on Privacy in Electronic Society*, 2014. [cited at p. 5]

[91] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. On non-cooperative genomic privacy. In *Financial Cryptography and Data Security*. Springer, 2015. [cited at p. 5]

[92] M. Humbert, M. H. Manshaei, J. Freudiger, and J.-P. Hubaux. Tracking games in mobile networks. In *Decision and Game Theory for Security*, pages 38–57. Springer, 2010. [cited at p. 67]

[93] M. Humbert, M. H. Manshaei, and J.-P. Hubaux. One-to-n scrip systems for cooperative privacy-enhancing technologies. In *49th Annual Allerton Conference on Communication, Control, and Computing*, pages 682–692. IEEE, 2011. [cited at p. 5]

[94] M. Humbert, T. Studer, M. Grossglauser, and J.-P. Hubaux. Nowhere to hide: Navigating around privacy in online social networks. In *Computer Security–ESORICS 2013*, pages 682–699. Springer, 2013. [cited at p. 5]

[95] J. Ioannidis, S. Ioannidis, A. Keromytis, and V. Prevelakis. Fileteller: Paying and getting paid for file storage. In *Financial Cryptography*, 2003. [cited at p. 85]

[96] S. Jahid, S. Nilizadeh, P. Mittal, N. Borisov, and A. Kapadia. Decent: A decentralized architecture for enforcing privacy in online social networks. *Arxiv preprint arXiv:1111.5377*, 2011. [cited at p. 22]

[97] P. Jain and P. Kumaraguru. Finding nemo: Searching and resolving identities of users across online social networks. *arXiv preprint arXiv:1212.6147*, 2012. [cited at p. 23]

[98] R. Jansen, N. Hopper, and Y. Kim. Recruiting new Tor relays with BRAIDS. In *CCS*, 2010. [cited at p. 100]

[99] E. T. Jaynes. Where do we stand on maximum entropy. *The Maximum Entropy Formalism*, pages 15–118, 1978. [cited at p. 90]

[100] C. S. Jensen, A. Kong, and U. Kjærulff. Blocking Gibbs sampling in very large probabilistic expert systems. *International Journal of Human Computer Studies*, 42(6):647–666, 1995. [cited at p. 48]

[101] S. Jha, L. Kruger, and V. Shmatikov. Towards practical privacy for genomic computation. *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 216–230, 2008. [cited at p. 81]

[102] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1087. ACM, 2013. [cited at p. 69, 81]

[103] A. D. Johnson and C. J. O'Donnell. An open access database of genome-wide association results. *BMC Medical Genetics 10:6*, 2009. [cited at p. 31]

[104] C.Y. Johnson. Project Gaydar: An MIT experiment raises new questions about online privacy. *Boston Globe*, 2009. [cited at p. 22]

[105] M. I. Jordan. Graphical models. *Statistical Science*, pages 140–155, 2004. [cited at p. 47]

[106] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin. A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on Information Technology in Biomedicine*, 12(5):606–617, 2008. [cited at p. 40, 69, 81]

[107] I. A. Kash, E. J. Friedman, and J. Y. Halpern. Optimizing scrip systems: Efficiency, crashes, hoarders, and altruists. In *Proceedings of the 8th ACM conference on Electronic commerce*, 2007. [cited at p. 85, 96]

[108] I. A. Kash, E. J. Friedman, and J. Y. Halpern. Manipulating scrip systems: Sybils and collusion. In *Auctions, Market Mechanisms and Their Applications*, 2009. [cited at p. 85]

[109] F. Kerschbaum, M. Beck, and D. Schönfeld. Inference control for privacy-preserving genome matching. In *GenoPri*, 2014. [cited at p. 81]

[110] P.D. Killworth and H.R. Bernard. The reversal small-world experiment. *Social networks*, 1:159–192, 1979. [cited at p. 14]

[111] B. Kirkpatrick, E. Halperin, and R. M. Karp. Haplotype inference in complex pedigrees. *Journal of Computational Biology*, 17(3):269–280, 2010. [cited at p. 48]

[112] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 2000. [cited at p. 24]

[113] J. Kleinberg. Small-world phenomena and the dynamics of information. *15th Advances in Neural Information Processing Systems (NIPS)*, pages 431–438, 2001. [cited at p. 24]

[114] D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003. [cited at p. 63, 64, 65]

[115] G. Kótyuk and L. Buttyán. A machine learning based approach for predicting undisclosed attributes in social networks. In *4th IEEE International Workshop on Security and Social Networking*, pages 361–366, 2012. [cited at p. 10, 22]

[116] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *STACS 99*, pages 404–413. Springer, 1999. [cited at p. 60]

[117] B. Krishnamurthy and C.E. Wills. Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, pages 37–42, 2008. [cited at p. 22]

[118] B. Krishnamurthy and C.E. Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 7–12, 2009. [cited at p. 22]

[119] F. Kschischang, B. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47, 2001. [cited at p. 30, 33, 36]

[120] R. Kumar, D. Liben-Nowell, and A. Tomkins. Navigating low-dimensional and hierarchical population networks. *Algorithms–ESA 2006*, pages 480–491, 2006. [cited at p. 24]

[121] H. Kunreuther and G. Heal. Interdependent security. *Journal of Risk and Uncertainty*, 26(2-3):231–249, 2003. [cited at p. 52]

[122] A. Laszka, M. Felegyhazi, and L. Buttyán. A survey of interdependent security games. *CrySyS*, 2, 2012. [cited at p. 49, 50, 51, 67]

[123] S. Lattanzi, A. Panconesi, and D. Sivakumar. Milgram-routing in social networks. In *Proceedings of the 20th international conference on World wide web*, 2011. [cited at p. 24]

[124] S. L. Lauritzen and N. A. Sheehan. Graphical models for genetic analyses. *Statistical Science*, pages 489–514, 2003. [cited at p. 47, 54]

[125] Y. Li, C. Willer, S. Sanna, and G. Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387, 2009. [cited at p. 48]

[126] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 2005. [cited at p. 23, 24]

[127] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th international conference on World wide web*, pages 1145–1146, 2009. [cited at p. 22]

[128] K. Loesing. Measuring the Tor network: Evaluations of client requests to directories. Technical report, Tor Project, 2009. [cited at p. 86]

[129] G. Magno, G. Comarela, D. Saez-Trumper, M. Cha, and V. Almeida. New kid on the block: Exploring the Google+ social graph. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 159–170, 2012. [cited at p. 15, 17]

[130] H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12, 2011. [cited at p. 10]

[131] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001. [cited at p. 10]

[132] D. Meier, Y. A. Oswald, S. Schmid, and R. Wattenhofer. On the windfall of friendship: Inoculation strategies on social networks. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 294–301. ACM, 2008. [cited at p. 52, 60]

[133] S. Milgram. The small world problem. *Psychology today*, 2:60–67, 1967. [cited at p. 12, 23]

[134] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007. [cited at p. 24]

[135] A. Mislove, B. Viswanath, K.P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260, 2010. [cited at p. 10, 12, 66]

[136] T.-W. Ngan, R. Dingledine, and D. S. Wallach. Building incentives into Tor. In *Financial Cryptography*, 2010. [cited at p. 100]

[137] D. R. Nyholt, C. E. Yu, and P. M. Visscher. On Jim Watson's APOE status: Genetic information is hard to hide. *European Journal of Human Genetics*, 17:147–149, 2009. [cited at p. 30, 71]

[138] A. M. Olteanu, K. Huguenin, R. Shokri, and J.-P. Hubaux. Quantifying the effect of co-location information on location privacy. In *Privacy Enhancing Technologies Symposium*, 2014. [cited at p. 67]

[139] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988. [cited at p. 30, 33]

[140] H. Pishro-Nik and F. Fekri. On decoding of low-density parity-check codes on the binary erasure channel. *IEEE Transactions on Information Theory*, 50:439–454, March 2004. [cited at p. 33]

[141] J. Preston. Seeking to disrupt protesters, Syria cracks down on social media. http://www.nytimes.com/2011/05/23/world/middleeast/23facebook.html?_r=1, May 2011. [cited at p. 2, 9, 25]

[142] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*, volume 414. John Wiley & Sons, 2009. [cited at p. 95]

[143] M. Rennhard and B. Plattner. Practical anonymity for the masses with morphmix. In *Financial Cryptography*, 2004. [cited at p. 86, 99, 100]

[144] S. I. Resnick. *Adventures in Stochastic Processes*. Birkhauser Boston, 1992. [cited at p. 89]

[145] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009. [cited at p. 81]

[146] F. Santos, M. Humbert, R. Shokri, and J.-P. Hubaux. Collaborative location privacy with rational users. *Proceedings of the Second international conference on Decision and Game Theory for Security*, pages 163–181, 2011. [cited at p. 99]

[147] A. Serjantov and G. Danezis. Towards an information theoretic metric for anonymity. In *Privacy Enhancing Technologies*, pages 41–53. Springer, 2003. [cited at p. 39]

[148] N. A. Sheehan. On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review*, 68(1):83–110, 2000. [cited at p. 48]

[149] J. Shi, R. Zhang, Y. Liu, and Y. Zhang. PriSense: Privacy-preserving data aggregation in people-centric urban sensing systems. In *IEEE INFOCOM 2010. 29th IEEE International Conference on Computer Communications*, 2010. [cited at p. 99, 102]

[150] W. Shih. A branch and bound method for the multiconstraint zero-one knapsack problem. *Journal of the Operational Research Society*, pages 369–378, 1979. [cited at p. 72]

[151] R. Shokri, P. Papadimitratos, G. Theodorakopoulos, and J.-P. Hubaux. Collaborative location privacy. *Proceedings of IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS)*, pages 500–509, 2011. [cited at p. 99]

[152] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. Quantifying location privacy. In *IEEE Symposium on Security and Privacy*, 2011. [cited at p. 39]

[153] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec. Protecting location privacy: optimal strategy against localization attacks. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 617–627. ACM, 2012. [cited at p. 67]

[154] F. Stajano, L. Bianchi, P. Liò, and D. Korff. Forensic genomics: kin privacy, driftnets and other open questions. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, 2008. [cited at p. 29, 47]

[155] G. Suarez-Tangil, J. Tapiador, P. Peris-Lopez, and A. Ribagorda. Evolution, detection and analysis of malware for smart devices. *IEEE Communications Surveys and Tutorials*, 2013. [cited at p. 50]

[156] B. Sullivan. Govt. agencies, colleges demand applicants' facebook passwords. http://redtape.msnbc.msn.com/_news/2012/03/06/10585353-govt-agencies-colleges-demand-applicants-facebook-passwords?chromedomain=usnews, 2012. [cited at p. 2, 9, 25]

[157] J. Sweeney and R. J. Sweeney. Monetary theory and the great capitol hill babysitting co-op crisis: Comment. *Journal of Money, Credit and Banking*, 9(1):86–89, 1977. [cited at p. 85]

[158] L. Sweeney, A. Abu, and J. Winn. Identifying participants in the personal genome project by name. *Available at SSRN 2257732*, 2013. [cited at p. 29, 47, 49, 69]

[159] A. Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific journal of Mathematics*, 5:285–309, 1955. [cited at p. 96]

[160] A. Thomas, A. Gutin, V. Abkevich, and A. Bansal. Multilocus linkage analysis by blocked Gibbs sampling. *Statistics and Computing*, 10(3):259–269, 2000. [cited at p. 48]

[161] K. Thomas, C. Grier, and D. Nicol. unFriendly: multi-party privacy risks in social networks. In *Privacy Enhancing Technologies*, pages 236–252, 2010. [cited at p. 21, 23]

[162] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969. [cited at p. 15, 20, 23]

[163] J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik. Privacy preserving error resilient DNA searching through oblivious automata. *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, 2007. [cited at p. 81]

[164] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the Facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011. [cited at p. 9, 17, 24]

[165] V. Vishnumurthy, S. Chandrakumar, and E.G. Sirer. Karma: A secure economic framework for peer-to-peer resource sharing. In *Workshop on Economics of Peer-to-Peer Systems*, 2003. [cited at p. 85]

[166] N. Vratonjic, K. Huguenin, V. Bindschaedler, and J.-P. Hubaux. How others compromise your location privacy: The case of shared public ips at hotspots. In *Privacy Enhancing Technologies*, pages 123–142. Springer, 2013. [cited at p. 66]

[167] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: Information leaks in genome wide association study. *Proceedings of the 16th ACM CCS*, pages 534–544, 2009. [cited at p. 47]

[168] R. Wang, X. Wang, Z. Li, H. Tang, M. K. Reiter, and Z. Dong. Privacy-preserving genomic computation through program specialization. *Proceedings of the 16th ACM CCS*, pages 338–347, 2009. [cited at p. 81]

[169] T. Watkins. Suddenly, Google Plus is outpacing Twitter to become the world's second largest social network. *Business Insider*, 2013. http://www.businessinsider.com/google-plus-is-outpacing-twitter-2013-5. [cited at p. 13]

[170] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002. [cited at p. 15, 24]

[171] A. F. Westin. *Privacy and Freedom*. Atheneum, 1967. [cited at p. 2]

[172] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B.Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218, 2009. [cited at p. 24]

[173] A. Yamada, T. H. J. Kim, and A. Perrig. Exploiting privacy policy conflicts in online social networks. Technical report, 2012. [cited at p. 12, 23]

[174] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540, 2009. [cited at p. 22]

[175] X. Zhou, B. Peng, Y. F. Li, Y. Chen, H. Tang, and X. Wang. To release or not to release: Evaluating information leaks in aggregate human-genome data. *ESORICS*, 2011. [cited at p. 47]

Index

MATHIAS HUMBERT

EPFL IC ISC LCA1
BC 210 (BC Building)
Station 14
CH-1015 Switzerland

mathias.humbert@epfl.ch
mathias.humbert@gmail.com
http://people.epfl.ch/mathias.humbert
+41 21 693 66 16

EDUCATION

***Sept. 2009 - present* — Ph.D. in Computer, Communication and Information Sciences**

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Ph.D. thesis: *When Others Impinge upon Your Privacy: Interdependent Risks and Protection in a Connected World*

Advisor : Prof. Jean-Pierre Hubaux, Laboratory for Communications and Applications

***Sept. 2007 - July 2009* — M.Sc. in Communication Systems**

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

M.Sc. thesis: *Location Privacy amidst Local Eavesdroppers*

Key courses: cryptography and security, algorithms in public-key cryptology, algebra, TCP/IP networking, signal processing for audio and acoustics, economics

***Sept. 2007 - May 2008* —Exchange Year in the EECS department at U.C. Berkeley**

University of California, Berkeley, USA

Key courses: digital signal processing, information theory and coding, fundamentals of wireless communication, digital image processing, convex optimization, information and services economy.

***Sept. 2004 - July 2007* — B.Sc. in Communication Systems**

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

B.Sc. thesis: *Polynomial Time Primality Testing*

Key courses : calculus, linear algebra, probability and statistics, physics, electromagnetics, programming, computer networks, algorithms, digital communication, stochastic processes, signal and speech processing, optimization, information systems, project management

PROFESSIONAL EXPERIENCE

***Sept. 2009 - present* — Research and Teaching Assistant**

Laboratory for Communications and Applications, EPFL, Switzerland

Teaching assistant for: *Computer Networks* (Prof. Hubaux; Fall'10'11, and Prof. Argyraki; Fall'12), *Logic Systems* (Dr. Kluter; Spring'11'12), *Pattern Classification and Machine Learning* (Prof. Seeger; Fall'13), and *Mobile Networks* (Prof. Hubaux; Spring' 14)

More than twenty student semester projects supervised.

***Sept. 2006 – July 2009* — Student Assistant**

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Student assistant for: *Algorithms* (Prof. Shokrollahi), *Circuits and Systems* (Prof. Hasler), and *Signal Processing for Communications* (Prof. Urbanke)

***June 2008 – Aug. 2008* — Intern in the Information Systems Department**

International Federation of Red Cross and Red Crescent Societies, Geneva, Switzerland
Introduction to datacenter (racks, servers, switches, routers …), Information Technology Infrastructure Library (ITIL), development of Definitive Software Library, documentation of test server room (policy, procedures, racks diagram), introduction to monitoring

***Sept. 2001* – *June 2007* — Tutor**

Centre Vaudois d'Aide à la Jeunesse, Lausanne, Switzerland
Private lessons provided to secondary school students (mathematics, physics, German, ...)

PROFESSIONAL SERVICES

Reviewer for conferences and journal

ACM CCS'11'12, FC'11, PETS'11'12'13, USENIX Security'14, Middleware'14, IEEE Journal on Selected Areas in Communications

COMPUTING

OS: Linux, Windows, Mac OS; **Sci.**: Python, Java, C++, MatLab, VHDL; **web/db**: HTML, MySQL

HONORS AND AWARDS

***Sept. 2007 – July 2009* — Excellence Scholarship (for Master degree)**

Scholarship as an outstanding student of EPFL

PUBLICATIONS

M. Humbert, E. Ayday, J.-P. Hubaux and A. Telenti, **On Non-cooperative Genomic Privacy,** 19th International Conference on Financial Cryptography and Data Security (FC), January 2015

M. Humbert, E. Ayday, J.-P. Hubaux and A. Telenti, **Reconciling Utility with Privacy in Genomics**, ACM Workshop on Privacy in the Electronic Society (WPES), November 2014

M. Humbert, E. Ayday, J.-P. Hubaux and A. Telenti, **Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy**, 20th ACM Conference on Comptuer and Communications Security (CCS), November 2013

M. Humbert, T. Studer, M. Grossglauser and J.-P. Hubaux, **Nowhere to Hide: Navigating around Privacy in Online Social Networks**, 18th European Symposium on Research in Computer Security (ESORICS), September 2013

E. Ayday, J.-L. Raisaro, M. Humbert and J.-P. Hubaux, **Towards Quantifying and Preventing the Leakage of Genomic Data Using Privacy-Enhancing Technologies**, 22nd USENIX Security Symposium, August 2013 (poster)

F. Santos, M. Humbert, R. Shokri and J.-P. Hubaux, **Collaborative Location Privacy with Rational Users**, Second Conference on Decision and Game Theory for Security (GameSec), November 2011

M. Humbert, M. Manshaei and J.-P. Hubaux, **One-to-n Scrip Systems for Cooperative Privacy-Enhancing Technologies**, 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), September 2011

M. Humbert, M. Manshaei, J. Freudiger and J.-P. Hubaux, **Tracking Games in Mobile Networks**, First Conference on Decision and Game Theory for Security (GameSec), November 2010

PATENT

E. Ayday, J.-P. Hubaux, J. L. Raisaro, A. Telenti, J. Fellay, P. J. McLaren, J. Rougemont, M. Humbert, **Genomic Privacy Protection,** European Patent Nr.: 12184372.6, September 2012

VOLUNTEER EXPERIENCE

***2010 – present* — Member of the City Council of Yverdon-les-Bains, Switzerland**
Chairman of my political group since March 2013

***2010 – present* — Treasurer of Amnesty International, groups of Lausanne**

***2004 – 2011* — Member of Amnesty International campus group**
Co-president, treasurer, and webmaster, at different time periods

***2007 – 2008* — Member of the IC School Council, EPFL**

***2006 – 2007* — Class representative, EPFL**

***2007* — Member of the consular polling station for French Presidential Election, Yverdon-les-Bains, Switzerland**

MISC.

Languages
French: mother tongue
English: full professional proficiency
German: good knowledge (studied for eight years at school)

Interests
Reading, music (piano), politics, economics
Sports: Soccer (outdoor and indoor; champion of the Berkeley IM Cup), tennis, squash, cycling, swimming, running

International House Alumnus, Berkeley, USA.