Interdependent Privacy Games: The Case of Genomics

Mathias Humbert[1], Erman Ayday[2⋆], Jean-Pierre Hubaux[1], and Amalio Telenti[3]

[1] Laboratory for Communications and Applications, EPFL, Lausanne, Switzerland
firstname.lastname@epfl.ch
[2] Bilkent University, Department of Computer Science, erman@cs.bilkent.edu.tr
[3] The J. Craig Venter Institute, La Jolla, USA, atelenti@jcvi.org

Abstract. Over the last few years, the vast progress in genome sequencing has highly increased the availability of genomic data. Today, individuals can obtain their digital genomic sequences at reasonable prices from many online service providers. Individuals can store their data on personal devices, reveal it on public online databases, or share it with third parties. Yet, it has been shown that genomic data is very privacy-sensitive and highly correlated between relatives. Therefore, individuals' decisions about how to manage and secure their genomic data are crucial. People of the same family might have very different opinions about (i) how to protect and (ii) whether or not to reveal their genome. We study this tension by using a game-theoretic approach. First, we model the interplay between two purely-selfish family members. We also analyze how the game evolves when relatives behave altruistically. We define closed-form Nash equilibria in different settings. We then extend the game to N players by means of multi-agent influence diagrams that enable us to efficiently compute Nash equilibria. Our results notably demonstrate that altruism does not always lead to a more efficient outcome in genomic-privacy games. They also show that, if the discrepancy between the genome-sharing benefits that players perceive is too high, they will follow opposite sharing strategies, which has a negative impact on the familial utility.

Keywords: Genomic Privacy, Interdependent Privacy, Game Theory, Altruism.

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. Any individual can obtain the sequencing of a significant part of his genome for less than $100. This availability raises many questions regarding the management (storage, sharing, etc.) and, ultimately, the privacy of genomic data. The genome contains very sensitive information about its owner such as his ethnicity, kinship, and predisposition

⋆ This work was carried out while the author was at EPFL.

to diseases. If this data is leaked, there could be serious consequences such as genetic discrimination, divorce [1] and blackmail (considering e.g., fatherhood issues) [10]. As genomic data is personal data, we could let individuals manage it independently of each other. However, as shown in [15], the genomic data of close relatives is highly correlated, thus leading to interdependent privacy risks. Hence, all genome-related decisions should be made by considering that genomic data is not only personal, but also *familial* data.

Nevertheless, thousands of individuals already spontaneously share their genomic data online, either anonymously[1] or with their real identity (e.g., on OpenSNP.org). Even for individuals who do not share their genomic data online, important decisions regarding the storage security of their genomes have to be made. Some will 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. Therefore, in this work, we consider that an individual whose DNA has been sequenced must make decisions on (i) whether to share his genomic data, and (ii) how much to invest in securing the storage of this data.

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 [4]. In particular, interdependent security (IDS) games have been proposed [20] for scenarios where agents make decisions that affect not only their own security risks but also those of others. Following the IDS works, we define two interdependent privacy (IDP) games between family members with different perceived benefits, costs and privacy levels. First, we study the interplay between two family members. With 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. Furthermore, we consider some altruistic[2] 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. Moreover, we quantify genomic-privacy loss with real genomic data, which provides very tangible results.

Our results show that, if the discrepancy is too high between the players' perceptions of the genome-sharing benefits, they will follow opposite strategies, creating externalities. These misaligned incentives lead to inefficient equilibria that result in a familial utility lower than when incentives are aligned. Our

[1] Anonymization has been proven to not be an effective technique for protecting identities of the data owners in the genomic context [13, 27].

[2] Each player takes into account the other players' utility when making a decision.

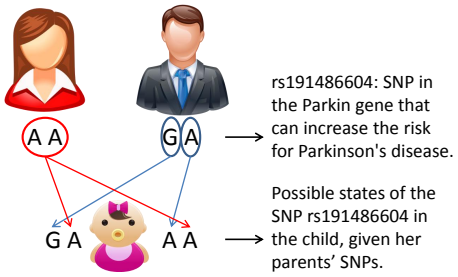


Fig. 1: Inheritance of a SNP (rs191486604) that can increase the risk for Parkinson's disease from the parents to the child. Homozygous-minor form (AA) of this SNP on the Parkin gene is associated with very severe, very young disease. For example, the father has the heterozygous form (GA) of the SNP (which does not increase the risk for the disease). However, the father might still want to protect his genome's privacy considering the fact that his child might have the homozygous-minor form of the SNP (depending on the genome of the mother).

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.

2 Genomic Background

Human DNA is composed of sequence of letters A, T, G, and C and each individual has a unique sequence. The most common DNA variation in the human population is known as single nucleotide polymorphism (SNP) and it occurs when a nucleotide at a specific position on the DNA varies between individuals of a given population. The predispositions of an individual to certain diseases can be determined based on his SNPs (along with his clinical factors), hence SNPs carry privacy-sensitive information about individuals (and their family members). A SNP includes two nucleotides (or alleles), and each of these nucleotides is either the major allele, which the most frequently observed nucleotide or the minor allele, which is the rare nucleotide. For each SNP position, a child inherits one allele from his mother and one from his father. Therefore, the content of a SNP position can be in one of the following states: (i) homozygous-major genotype, if an individual receives the same major allele from both parents (which we encode with 0); (ii) heterozygous genotype, if he receives a different allele from each parent (one major and one minor, which we encode with 1); or (iii) homozygous-minor genotype, if he inherits the minor allele from both parents (which we encode with 2). Note that each allele of a parent is inherited with equal probability of 0.5.

We illustrate the inheritance of a SNP from parents to offspring in Fig. 1. In this example, we illustrate a SNP (rs191486604, with major allele G and minor allele A) in the Parkin gene resulting in the amino acid Gly430 to Asp mutation

that increases the risk for Parkinson's disease [2]. This is a condition characterized by progressive problems with movement and balance.[3] Mutations in Parkin are associated with the juvenile form of Parkinson disease, which appears before age 20, and some cases of the more common, late-onset form that begins after age 50. In this example, the father has a heterozygous SNP (rs191486604) which does not increase the risk for the Parkinson's disease, therefore the father might not worry about his own privacy. However, (from the father's point of view), there is a possibility that his children might have a homozygous-minor SNP at rs191486604 (assuming that the mother might have a heterozygous or homozygous minor SNP) which would increase the risk of the Parkinson's disease for his children. Therefore, the father might still want to protect his genome's privacy considering this situation (altruistic behaviors regarding the close family members). Similarly, as the mother has a homozygous-minor SNP at rs191486604, she might have motivation to protect both her and her children's genomic privacy.

3 Model

Users: We consider a set of N users from a family whose genotypes are sequenced. We focus on the most common DNA variant, the single nucleotide polymorphism (SNP), described in the previous section. We assume that all users have the same number and set Ω of SNPs sequenced. Users have to make choices regarding the investment in 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 [7], there are various advantages to keeping a person's genome on a smartphone. It is portable, highly personal, and has very good computational and storage capabilities. Unfortunately, malware in smartphones has exploded over the last few years [26], and keeping a mobile device secure causes non-negligible costs. Alternatively, a user could decide to outsource the storage of his genomic data to a third party. A user might also want to publicly share his SNPs, essentially because his perceived benefits outweigh the perceived cost (loss) for his genomic privacy.[4] 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 reasons for gathering individuals' genotypes can be multiple. For instance, he could sell the collected genomic data to life or health insurance companies that would then use it to genetically discriminate against potential insurees. As usually assumed in IDS games, the adversary is considered to be an exogenous, persistent

[3] There are many genes associated with various forms of Parkinson, but homozygous-minor form (AA) of rs191486604 on the Parkin gene is associated with very severe, very young disease.

[4] See, e.g., http://opensnp.wordpress.com/2011/11/17/first-results-of-the-survey-on-sharing-genetic-information/ to understand users' motivations for and fears about genome sharing.

threat [20]. Thus, we do not model him as a strategic agent, but rather as probability $h(\cdot)$ of a successful breach in the targeted system. If a user decides to publicly disclose his SNPs online, the probability of a breach is equal to 1.

4 Genomic Privacy Games

The genomes of close family members are highly correlated. Thus, individuals' behaviors regarding their genomes will not only affect their personal genomic privacy, but also those of their relatives. Game theory enables us to model the interplay between users with dependent payoffs and potentially conflicting interests, and 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 game G_s, for each player P_i, the strategy $x_i \in \mathcal{S}$ 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 [12, 20], we consider here the discrete model, i.e., $x_i \in \{0, 1\}$. $x_i = 1$ means "to invest in securing his own device", and $x_i = 0$ means "to not invest", by putting his data on his device or outsourced to an untrusted third party (that could be itself attacked). The strategy profile is then defined as $\boldsymbol{x} = [x_1, \cdots, x_N]^T$. In 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 a finer granularity of disclosure is studied in detail in a cooperative context in [16]. 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 $\boldsymbol{d} = [d_1, \cdots, d_N]^T$.

• **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.[5] For instance, 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 between the SNP values inferred

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

by the adversary $\hat{y}_i^k$'s and the actual values y_i^k's, $\forall g_k \in \Omega$ [15].[6] Defining Y_i^k as the random variable representing SNP g_k of player P_i, the genomic privacy of P_i is

$$E_i = \frac{1}{|\Omega|} \sum_{k:g_k \in \Omega} \sum_{\hat{y}_i^k \in \{0,1,2\}} P(Y_i^k = \hat{y}_i^k | Y_O = \boldsymbol{y_O}) \left\| y_i^k - \hat{y}_i^k \right\|_1, \quad (1)$$

where Y_O represents the SNPs observed by the adversary. This set depends on the strategies of the players in G_s and G_d. We will denote $E_{i,0}$ to be the genomic privacy when no SNP is observed, i.e., when $P(Y_i^k = \hat{y}_i^k | Y_O = \boldsymbol{y_O}) = P(Y_i^k = \hat{y}_i^k)$. This initial privacy level is computed by using the minor allele frequencies (MAFs) given by population statistics [15]. In general, as the observation depends on the strategy profile $\boldsymbol{x}$ (respectively $\boldsymbol{d}$), E_i will be a function of $\boldsymbol{x}$ (respectively $\boldsymbol{d}$) in game G_s (respectively G_d). As assumed in several IDS games (e.g., [19]), the probability of successful breach is set to zero when a player invests in security, i.e., $h(x_i = 1) = 0$. Otherwise, $h(x_i = 0) = p_a$ with $0 < p_a \leq 1$. For game G_d, $h(d_i = 1) = 1$ as discussed in Section 3, and $h(d_i = 0) = 0$.[7] 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 $\boldsymbol{x} = (x_i, \boldsymbol{x}_{-i})$ or $\boldsymbol{d} = (d_i, \boldsymbol{d}_{-i})$. 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 access the player's genome and then there is only an indirect privacy loss. Defining $h(\boldsymbol{x}_{-i})$ 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

$$u_i(x_i, \boldsymbol{x}_{-i}) = b_i^g - (x_i c_i + h(x_i) E_{i,0} + (1 - h(x_i))\, h(\boldsymbol{x}_{-i})\, (E_{i,0} - E_i(\boldsymbol{x}_{-i}))), \quad (2)$$

and his payoff in game G_d is

$$u_i(d_i, \boldsymbol{d}_{-i}) = b_i^g + d_i b_i^d - ((1 - d_i) c_i + d_i E_{i,0} + (1 - d_i)\, (E_{i,0} - E_i(\boldsymbol{d}_{-i}))).^{8} \quad (3)$$

• **Social Welfare:** We define the *social welfare* function as the sum of the payoffs of all players: $U(\boldsymbol{x}) = \sum_{i:P_i \in \mathcal{P}} u_i(\boldsymbol{x})$ for G_s, and $U(\boldsymbol{d}) = \sum_{i:P_i \in \mathcal{P}} u_i(\boldsymbol{d})$ for G_d.
• **Altruism:** Finally, we consider that family members are usually not purely selfish regarding their relatives, hence some altruistic factors play a role in their decisions. Following an idea introduced in [22] 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 $k(i, j) \in \mathbb{N}^*$ that represents

[6] Note that a SNP value is encoded by the set $\{0, 1, 2\}$ whose elements represent the number of minor alleles in the SNP.

[7] In G_d, we assume that a player who does not share his SNPs will always invest in security. Note also that G_d is a special case deriving from G_s.

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

Table 1: Normal form of the two-player game G_s.

$P_1 \backslash P_2$	$x_2 = 1$	$x_2 = 0$
$x_1 = 1$	$(b_1^g - c_1, b_2^g - c_2)$	$(b_1^g - c_1 - p_a(E_{1,0} - E_{1\|2}), b_2^g - p_a E_{2,0})$
$x_1 = 0$	$(b_1^g - p_a E_{1,0}, b_2^g - c_2 - p_a(E_{2,0} - E_{2\|1})$	$(b_1^g - p_a E_{1,0} - (1-p_a)p_a(E_{1,0} - E_{1\|2}),$ $b_2^g - p_a E_{2,0} - (1-p_a)p_a(E_{2,0} - E_{2\|1}))$

the degree of kinship between relatives i and j.[9] $\alpha = 0$ 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 (2)):

$$u_i^a(x_i, \boldsymbol{x}_{-i}) = u_i(x_i, \boldsymbol{x}_{-i}) + \sum_{j:P_j \in \mathcal{P}, j \neq i} \alpha^{k(i,j)} u_j(x_i, \boldsymbol{x}_{-i}). \tag{4}$$

5 Two-Player Games

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

5.1 Selfish Players

We start our analysis with game G_s whose strategic representation is shown in Table 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 1. *For any value $c \in [0, \infty)$, there exists at least one pure Nash equilibrium (NE) in G_s. The NE are defined by the best responses (x_1^*, x_2^*):*

$$(x_1^*, x_2^*) = \begin{cases} (1,1) & \text{if } c < \min(t_1, t_2) \\ (1,1), mNE & \text{if } \min(t_1, t_2) < c < \max(t_1, t_2) \\ (1,1), (0,0) & \text{if } \max(t_1, t_2) < c < p_a \min(t_1^0, t_2^0) \\ (0,0), mNE & \text{if } p_a \min(t_1^0, t_2^0) < c < p_a \max(t_1^0, t_2^0) \\ (0,0) & \text{if } c > p_a \max(t_1^0, t_2^0) \end{cases} \tag{5}$$

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 (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.

[9] $k = 1$ 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.

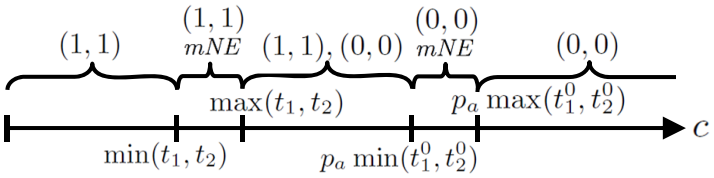


Fig. 2: Dependence of the NE of game G_s with respect to the investment cost c.

Proof. The NE is derived from the best responses of both players. For player P_i, the best response to the strategy x_j of player P_j, $j \neq i$ is defined as

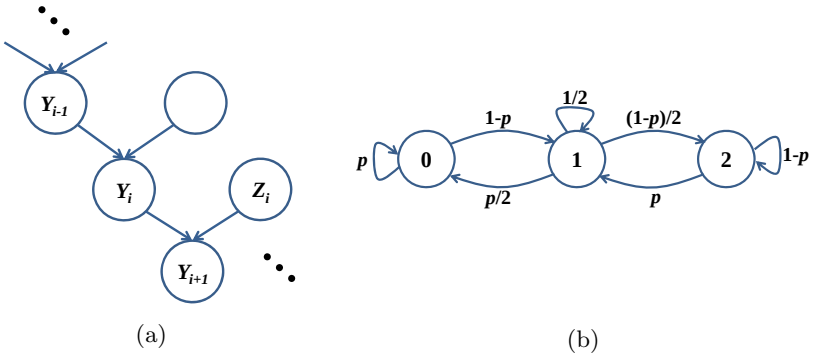
$$x_i^* = \arg\max_{x_i} u_i(x_i, x_j).$$

These best responses can be easily derived by inspecting the payoffs in Table 1 above. Then, there exists a pure NE if and only if two strategies x_i^* and x_j^* are the mutual best responses of each other.

Fig. 2 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 [15], the authors show that, in the general case, belief propagation on factor graphs can be used to compute the posterior marginal probability $P(Y_i^k|Y_O)$ given some observed genomic data, and thus to quantify genomic privacy. We now 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 $P(Y_i^k|Y_O)$, thus for $E_{i,0}$ and $E_{i|j}$. As we assume that all players have the same set of SNPs Ω sequenced and potentially exposed, and that the adversary can access either the whole sequence of SNPs or nothing (as he either successfully breaches the system or not), linkage disequilibrium (correlations) between the SNPs would not help the adversary very much, thus it is not used in the computation of genomic privacy here. Hence, 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, so drop the superscript k. Assuming Y_i is the random variable representing a SNP of an individual at generation i in a familial branch (see Fig. 3a), and p is the major allele frequency of the SNP, we have the following lemma.

Lemma 2. *The sequence $\{Y_n\}$ is a discrete stochastic process. Moreover, it is a first-order homogeneous Markov chain, i.e., the conditional probability of Y_{i+1} given (direct) ancestors in one of the parents' family branches is formally defined as $P(Y_{i+1} = y_{i+1}|Y_i = y_i, Y_{i-1} = y_{i-1}, \dots) = P(Y_{i+1} = y_{i+1}|Y_i = y_i)$. Its transition matrix P is defined as follows:*

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



(a) (b)

Fig. 3: (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.

where $p_{mn} = P(Y_{i+1} = n | Y_i = m)$, 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 Fig. 3a, where each node in the BN represents the SNP of a relative [21]. 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 Y_{i+1} is conditionally independent of any ancestor in the subnetwork (whose leaf node is Y_i) given Y_i. Thus, $P(Y_{i+1} = y_{i+1} | Y_i = y_i, Y_{i-1} = y_{i-1}, \dots) = P(Y_{i+1} = y_{i+1} | Y_i = y_i)$. Finally, the transition probability $P(Y_{i+1} = y_{i+1} | Y_i = y_i)$ is equal to

$$\sum_{z_i \in \{0,1,2\}} P(Y_{i+1} = y_{i+1} | Y_i = y_i, Z_i = z_i) P(Z_i = z_i), \tag{6}$$

where $P(Y_{i+1} = y_{i+1} | Y_i = y_i, Z_i = z_i)$ is given by the Mendelian inheritance probabilities, and $P(Z_i = z_i)$ by the major allele frequency p (Z_i is not observed). Equation (6) directly leads to the transition matrix P.

We have noticed that the reverse process, which is the conditional probability of Y_{i-1} given direct descendants 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 Fig. 3b.

Lemma 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 $P(Y_{i+k} | Y_i)$ of a relative k-degrees apart from another

individual i whose SNP is observed and equal to m is, by definition of the Markov chain, given by $\boldsymbol{\pi}_{i+k} = \boldsymbol{\pi}_i P^k$, where $\boldsymbol{\pi}_i$ is a row vector that is equal to 1 in the m^{th} coordinate and 0 elsewhere. Note also that the stationary distribution, defined as the vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi} = \boldsymbol{\pi} P$, is equal to the vector of prior probabilities $(P(Y_i))$, given by the major allele probability p:

$$\boldsymbol{\pi} = \left(p^2 \; 2p(1-p) \; (1-p)^2\right). \tag{7}$$

This follows the intuition, as $\boldsymbol{\pi}$ is defined to be equal to any of the columns of P^k when k tends to infinity. When the observed relative j is far enough from the targeted individual i in the family tree, the genome of j has no influence on i's genome. 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 $P(Y_{i_1} = b | Y_{i_2} = a)$ of i_1 given his sibling i_2. Defining Y^m_{i-1} and Y^f_{i-1} to be the mother and father SNP variable respectively, entry m_{ab} of matrix M is derived as follows:

$$P(Y_{i_1} = b | Y_{i_2} = a) = \sum_{\substack{Y^m_{i-1} \in \{0,1,2\} \\ Y^f_{i-1} \in \{0,1,2\}}} P(Y_{i_1} = b, Y^m_{i-1}, Y^f_{i-1} | Y_{i_2} = a) \tag{8}$$

$$= \sum_{\substack{Y^m_{i-1} \in \{0,1,2\} \\ Y^f_{i-1} \in \{0,1,2\}}} P(Y_{i_1} = b | Y^m_{i-1}, Y^f_{i-1}) P(Y^m_{i-1} | Y^f_{i-1}, Y_{i_2} = a) P(Y^f_{i-1} | Y_{i_2} = a), \tag{9}$$

where we used the chain rule to go from (8) to (9), and the fact that $P(Y_{i_1} | Y^m_{i-1}, Y^f_{i-1}, Y_{i_2}) = P(Y_{i_1} | Y^m_{i-1}, Y^f_{i-1})$, since two siblings are conditionally independent given both their parents. $P(Y_{i_1} = b | Y^m_{i-1}, Y^f_{i-1})$ is given by the Mendelian inheritance probabilities, $P(Y^f_{i-1} | Y_{i_2} = a)$ is given by matrix P, and

$$P(Y^m_{i-1} | Y^f_{i-1}, Y_{i_2} = a) = \frac{P(Y_{i_2} = a | Y^f_{i-1}, Y^m_{i-1}) P(Y^m_{i-1})}{P(Y_{i_2} = a | Y^f_{i-1})}, \tag{10}$$

using the Bayes rule and the fact that $P(Y^m_{i-1} | Y^f_{i-1}) = P(Y^m_{i-1})$, as two parents are independent if no child is observed. Again, one can compute every factor of (10) by using the inheritance probabilities, matrix P, and the major allele frequency p. Matrix M is equal to

$$\begin{pmatrix} p^2 + pq + \frac{q^2}{4} & pq + \frac{q^2}{2} & \frac{q^2}{4} \\ \frac{p^2}{2} + \frac{pq}{4} & \frac{p^2}{2} + \frac{3}{2}pq + \frac{q^2}{2} & \frac{pq}{4} + \frac{q^2}{2} \\ \frac{p^2}{4} & \frac{p^2}{2} + pq & \frac{p^2}{4} + pq + q^2 \end{pmatrix}, \tag{11}$$

where $q = 1 - p$.

Defining the 3×3 distance matrix D with elements $d_{ij} = |i - j|$ and the (column) vector $\boldsymbol{y}_i$ whose m^{th} coordinate is equal to 1 and others 0 (where m is the SNP value), we have the following lemma.

Lemma 3. *The genomic privacy E_i of individual i at any SNP is:*

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

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

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

(i) If no observations are made, then $P(Y_i = \hat{y}_i|Y_O) = P(Y_i = \hat{y}_i)$, the prior probability, which is given by the major allele frequency p. This is equal to $\boldsymbol{\pi}$ given in (7). The second element $||y_i - \hat{y}_i||_1$ is simply expressed in matrix format by $D\boldsymbol{y}_i$. Hence, $E_{i,0} = \boldsymbol{\pi} D\boldsymbol{y}_i$.

(ii) If i is a k^{th}-degree relative of j, then the conditional probability distribution $P(Y_i|Y_j)$ is given by $\boldsymbol{\pi}_{j+k} = \boldsymbol{\pi}_j P^k$ from Lemma 2, leading to $E_{i|j} = \boldsymbol{\pi}_j P^k D\boldsymbol{y}_i$.

(iii) In case of indirect relationship, 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 j who is a sibling of the direct ancestor a_i of i. If a_j and j are u^{th}-degree relatives, $\boldsymbol{\pi}_{a_j} = \boldsymbol{\pi}_j P^u$. Then, as a_j and a_i are siblings, we make use of matrix M defined in (11) to compute the conditional probability of a_i's SNP given a_j's SNP value. Thus, $\boldsymbol{\pi}_{a_i} = \boldsymbol{\pi}_{a_j} M = \boldsymbol{\pi}_j P^u M$. Finally, if a_i and i are v^{th}-degree relatives, we have $\boldsymbol{\pi}_i = \boldsymbol{\pi}_{a_i} P^v = \boldsymbol{\pi}_j P^u M P^v$. Hence, we get $E_{i|j} = \boldsymbol{\pi}_j P^u M P^v D\boldsymbol{y}_i$.

To illustrate the third case of Lemma 3, let us take for example two close relatives, uncle and nephew. If j is the uncle of i, then the genomic privacy of i given j at a certain SNP is $E_{i|j} = \boldsymbol{\pi}_j P^1 M P^0 D\boldsymbol{y}_i = \boldsymbol{\pi}_j PMD\boldsymbol{y}_i$ whereas, if j is the nephew of i, the genomic privacy of i is $E_{i|j} = \boldsymbol{\pi}_j MPD\boldsymbol{y}_i$.

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

Theorem 1. *For any value $c \in [0, \infty)$, the pure Nash equilibrium is:*

$$(x_1^*, x_2^*) = \begin{cases} (1,1) & \text{if } c < \max(t_1, t_2) \\ (1,1),(0,0) & \text{if } \max(t_1,t_2) < c < p_a \min(t_1^0, t_2^0) \\ (0,0) & \text{if } c > p_a \min(t_1^0, t_2^0) \end{cases} \tag{12}$$

if $\max(t_1,t_2) < p_a \min(t_1^0, t_2^0)$, where $t_i^0 = \frac{1}{|\Omega|} \sum_{l:g_l \in \Omega} \boldsymbol{\pi}^l D\boldsymbol{y}_i^l$, $t_i = \frac{p_a}{|\Omega|} \big(\sum_{l:g_l \in \Omega} ((1-p_a)\boldsymbol{\pi}^l + p_a \boldsymbol{\pi}_j^l P_l^k) D\boldsymbol{y}_i^l \big)$ if i and j are direct k^{th}-degree relatives, and $t_i = \frac{p_a}{|\Omega|} \big(\sum_{l:g_l \in \Omega} ((1-p_a)\boldsymbol{\pi}^l + p_a \boldsymbol{\pi}_j^l P_l^u M P_l^v) D\boldsymbol{y}_i^l \big)$ if i and j are not in direct line, u and v as defined in Lemma 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 Ω the genomic privacy expressions computed in Lemma 3 and embedding them into the NE computed in Lemma 1 (keeping only pure NE), after some reordering, we get the NE in (12), as well as the expressions t_i's and t_i^0.

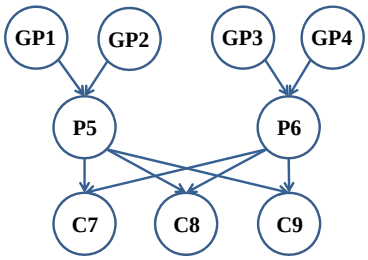
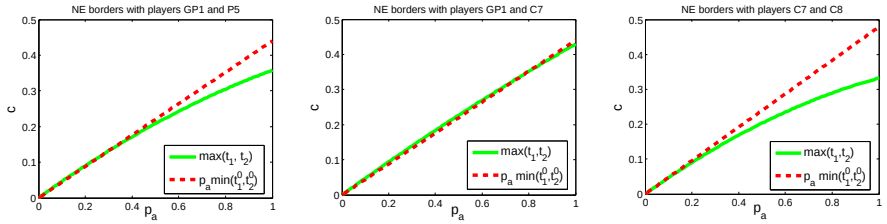


Fig. 4: Bayesian network representation of nine relatives of the CEPH/Utah pedigree 1463.

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 [9]. We filter 8 of the 11 children out, thus keeping 9 relatives in total. Fig. 4 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. Fig. 5 shows the thresholds separating the three different cases of NE in Theorem 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,[10] 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 Fig. 5c, as the difference between the initial privacy levels $E_{i,0}$'s and posterior levels $E_{i|j}$ is the most significant (see Table 3). 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

[10] This happens for $p_a < 0.29$ in Fig. 5a and $p_a < 0.78$ in Fig. 5b.



(a) (b) (c)

Fig. 5: Thresholds of Theorem 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.

players, leading (for a tiny subset of values of p_a of c) to inefficient NE, such as $(0, 1)$, 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 invest in security, or do not.

We now move to the disclosure game G_d. Table 2 shows the resulting payoffs for two players P_1 and P_2. The following theorem determines its NE.

Theorem 2. *For any value $b_1^d \in [0, \infty)$, and $b_2^d \in [0, \infty)$, the pure Nash equilibrium is:*

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

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

Table 2: Normal form of the two-player game G_d.

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

Table 3: Genomic privacy levels of grandparent GP1, parent P5, children C7 and C8, from the Utah family shown in Fig. 4.

(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

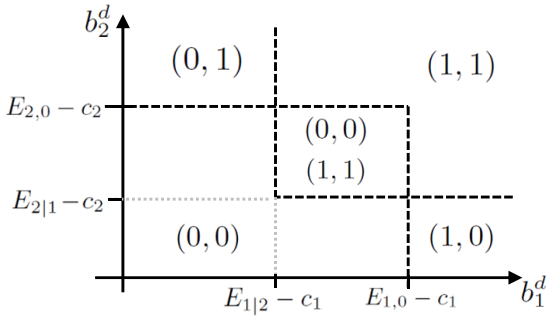


Fig. 6: Dependence of the NE w.r.t. the genome-sharing benefits b_1^d and b_2^d.

Proof. The NE is derived from the best responses of both players. As for the proof of Lemma 1, the best responses can be easily derived by inspecting the payoffs in Table 2 above. Then, there exists a pure NE if and only if two strategies x_i^* and x_j^* are the mutual best responses of each other.

Fig. 6 illustrates the NE computed in Theorem 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 that 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 3 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 perceived by the players is high enough, these players follow opposite strategies, one putting the other's privacy at risk by sharing his genome.

5.2 Altruistic Players

In this subsection, we analyze how the equilibria evolve when the players are not purely selfish, but also consider 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 happens.

To evaluate how the NE is affected by altruistic behavior, we focus on 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, given the other player's strategy, no player can reduce his altruistic payoff u^a by unilaterally changing his strategy. Defining $b_i = b_i^d + c_i$ for the ease of presentation, we have the following theorem.

Theorem 3. *For any value $b_1 \in [0, \infty)$, and $b_2 \in [0, \infty)$, the pure FNE is:*

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

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

Proof. The NE is derived similarly to Theorem 2 but with different payoff functions. In an altruistic context, player P_i maximizes u_i^a, which takes into account the other player's payoff, instead of u_i. Hence, we just need to replace the payoffs in Table 2 by the new altruistic payoffs u_i^a, and compute the subsequent best responses in order to get the pure FNE.

These different NE are depicted in Fig. 7 by circled numbers separated by (thick) dotted lines. Note the shift upwards and to the right 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 value of b_i higher 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.

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^*.

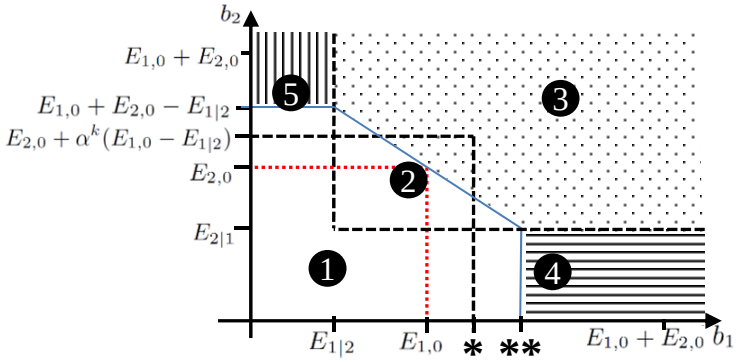


Fig. 7: Familial NE and social optima with respect to b_1 and b_2. Circled numbers represent the five different cases of Theorem 3, in order, separated by (thick) dotted lines in the figure. The red (small) dotted lines represent the borders of Fig. 6. The four different texture patterns represent the strategies of the social optimum, depicted in Theorem 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 $E_{1,0}+E_{2,0}-E_{2|1}$.

Theorem 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:*

$$(d_1^*, d_2^*) = \begin{cases} (0,0) & \text{if } (b_1+b_2 < E_{1,0}+E_{2,0}) \wedge (b_1 < E_{1,0}+E_{2,0}-E_{2|1}) \wedge \\ & \quad (b_2 < E_{1,0}+E_{2,0}-E_{1|2}) \\ (1,0) & \text{if } (b_1 > E_{1,0}+E_{2,0}-E_{2|1}) \wedge (b_2 < E_{2|1}) \\ (0,1) & \text{if } (b_2 > E_{1,0}+E_{2,0}-E_{1|2}) \wedge (b_1 < E_{1|2}) \\ (1,1) & \text{if } (b_1+b_2 > E_{1,0}+E_{2,0}) \wedge (b_2 > E_{2|1}) \wedge (b_1 > E_{1|2}) \end{cases} \tag{13}$$

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

Proof. This theorem is derived by simply summing the payoffs of both players in Table 2 for all four strategy combinations and selecting the combination of strategies that leads to the maximum sum for any value of b_1 and b_2.

The socially optimal strategies are represented schematically with respect to b_1 and b_2 by the texture of Fig. 7. Given this social optimum $U^*(\boldsymbol{s})$, the price of anarchy (PoA), which measures how the game efficiency decreases due to selfishness, is defined as $U^*(\boldsymbol{s})/\min_{NE} U(\boldsymbol{s})$ [18]. The price of stability (PoS) also measures this inefficiency but, assuming that players coordinate amongst themselves, considers the best NE instead of the worst one, i.e., is defined as $U^*(\boldsymbol{s})/\max_{NE} U(\boldsymbol{s})$ [5].

Following the notion of windfall of friendship (WoF) proposed in [22], we define the windfall of kinship (WoK) as the ratio between the social welfare of the worst FNE and the social welfare of the worst NE:

$$\kappa(\alpha, k) = \frac{\min_{FNE} U(\boldsymbol{s})}{\min_{NE} U(\boldsymbol{s})} \tag{14}$$

Given this definition, we can state the following theorem.

Theorem 5. *If b_1, b_2 are such that*

$$\begin{cases} b_1 + b_2 > E_{1,0} + E_{2,0} \\ b_1 < E_{1,0} + \alpha^k (E_{2,0} - E_{2|1}) \\ b_2 < E_{2,0} + \alpha^k (E_{1,0} - E_{1|2}), \end{cases} \tag{15}$$

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 is 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 Fig. 7). We know, from Theorem 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 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 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 (15), the two others being given by the area we are focusing on. If $b_1 + b_2 < E_{1,0} + E_{2,0}$, 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 Fig. 7.

This theorem tells us that, contrary 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 (15). 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 an outcome worse than in the selfish NE. However, note that the WoK is always less than or equal to the PoA. Indeed, as for any $\alpha \in [0, 1]$, $k \geq 1$, $\min_{FNE} U(\boldsymbol{s}) \leq U^*(\boldsymbol{s})$, it directly follows from (14) that $\kappa(\alpha, k) \leq \text{PoA}$.

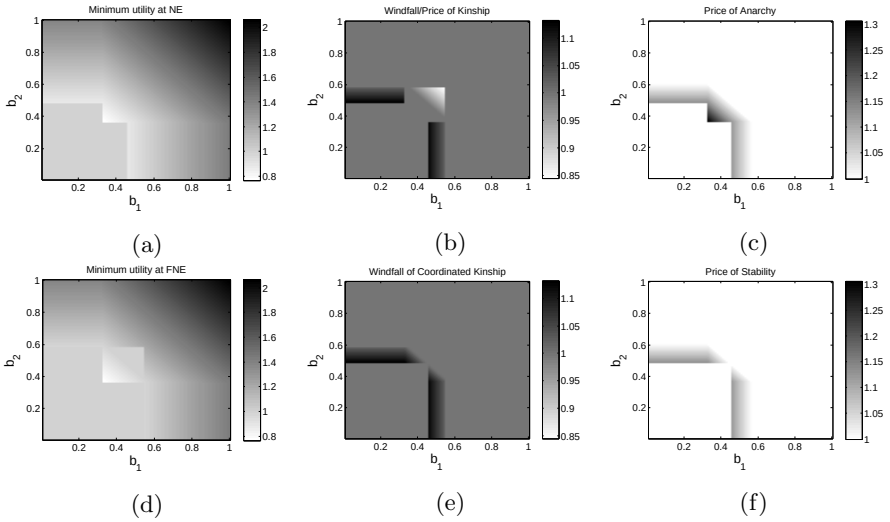


Fig. 8: Evaluation of the (in)efficiency of the NE and FNE with respect to b_1 and b_2. (a) Minimum social welfare at NE, (b) windfall/price of kinship, (c) price of anarchy, (d) minimum social welfare at FNE, (e) windfall of coordinated kinship, and (f) price of stability in G_d with GP1 and P5, $\alpha = 0.8$, and $b_1^g = b_2^g = 0.5$.

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:

$$\gamma(\alpha, k) = \frac{\max_{FNE} U(\boldsymbol{s})}{\max_{NE} U(\boldsymbol{s})} \tag{16}$$

This new definition enables us to state the following theorem.

Theorem 6. *For any $b_1 \in [0, \infty)$, $b_2 \in [0, \infty)$, $k \geq 1$, and $\alpha \in [0, 1]$, it holds that:*

$$1 \leq \gamma(\alpha, k) \leq PoS \leq PoA. \tag{17}$$

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

$$\frac{U^*(\boldsymbol{s})}{\max_{NE} U(\boldsymbol{s})} \leq \frac{U^*(\boldsymbol{s})}{\min_{NE} U(\boldsymbol{s})}, \tag{18}$$

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

$$\frac{\max_{FNE} U(\boldsymbol{s})}{\max_{NE} U(\boldsymbol{s})} \leq \frac{U^*(\boldsymbol{s})}{\max_{NE} U(\boldsymbol{s})}, \tag{19}$$

thus $\gamma(\alpha, k) \leq$ PoS. We know from Theorem 5 that $\kappa(\alpha, k) < 1$ in the triangle defined by (15). 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 (15), 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, $\max_{FNE} U(\boldsymbol{s}) / \max_{NE} U(\boldsymbol{s}) = 1$ 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, 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 5.1. 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. Fig. 8 shows the resulting graphs. First, we notice the shift upwards and to the right of $(0,0)$ between NE and FNE; it follows the borders shown in Fig. 7. We also see that minimum social welfare is minimal in the squares standing in the middle of both Figs. 8a and 8d. Looking at Fig. 8b, we clearly notice that the WoK is smaller than 1 for the values of b_1 and b_2 close to 0.5, thus confirming Theorem 5. However, as soon as both players coordinate amongst themselves, the ratio between the social welfare of FNE and the social welfare of NE (WoCK) becomes always greater than or equal to 1, as illustrated in Fig. 8e. Finally, we note that PoA and PoS are always greater than or equal to 1, that PoS $\leq$ PoA, and that PoS $\geq$ WoCK, thus confirming Theorem 6.
Discussion: In conclusion, if players cannot coordinate amongst themselves, their altruistic prudence about the disclosure of their genomes can lead to a worse social outcome than in the purely selfish setting, as shown in Theorem 5 and in Fig. 8b.

6 N-Player Game

In this section, we extend the genomic privacy game to consider $N > 2$ relatives. Contrary 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 form if more than one other family member discloses 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 [17]. A MAID is an extension of the Bayesian network framework that embeds, in addition to random variables, decision and utility variables, and enables us to consider multiple strategic agents, thus 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 Fig. 9. The chance[11] variable Y_i

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

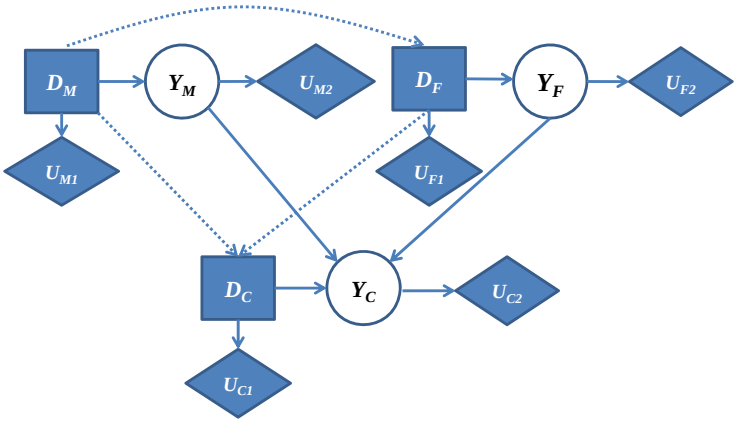


Fig. 9: Multi-agent influence diagram representing a trio (mother, father, child) with one decision variable (square), one chance variable (circle) representing the SNPs of the individual, and two utility variables (diamonds) per person. Full lines represent probabilistic or deterministic dependencies, whereas dotted lines 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.

is defined as $P(Y_i = y_i) = 1$ (other values having probability 0) if $d_i = 1$, and $P(Y_i = \hat{y}_i|Y_O)$ if $d_i = 0$. Note that, we represent the chance variable Y_i for a single SNP, but in fact there are $|\Omega|$ chance variables that directly depend on d_i, and are independent of each other. A child's SNP is probabilistically determined by his parents' genomes, as explained in Section 2. We also define two utility variables: $u_{i1} = b_i^g + d_i b_i^d - E_{i,0}$, which directly depends on d_i, and $u_{i2} = E_i$, which directly depends on the chance variable 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 with 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 Fig. 9, the following decision ordering is shown: mother, father and then child. Under these assumptions, we can state the following lemma.

Lemma 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 proof directly follows from the concept of *s-reachability*, defined in Definition 5.3 of [17]. 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. If a decision node D_i is observed by D_j (dotted edge in Fig. 9), 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

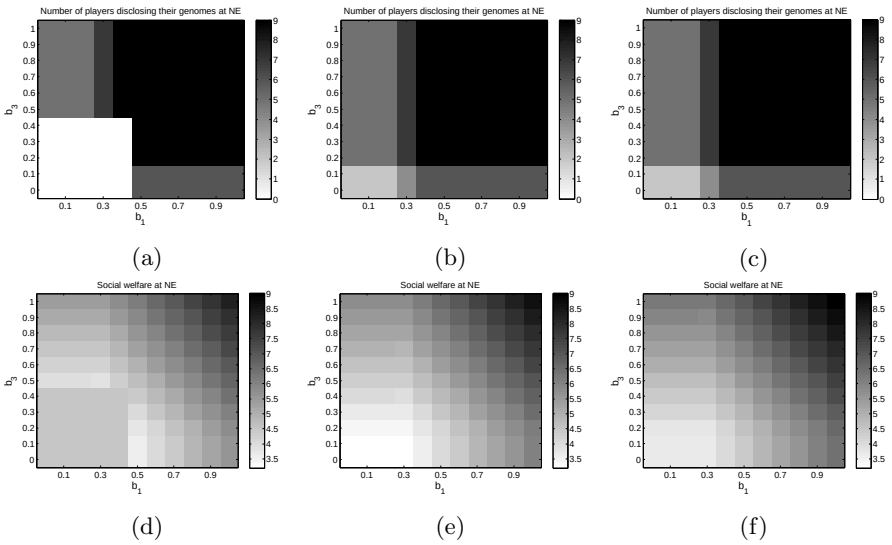


Fig. 10: Outcome of the N-player game. Number of players disclosing their genomes (first row) and social welfare (second row) at NE in the N-player game G_d. We set $b_2 = 0.4$ in (a) and (d), $b_2 = 0.6$ in (b) and (e), and $b_2 = 0.8$ in (c) and (f).

as $\delta(d_j|d_i), \forall d_i \in \{0,1\}$, thus this decision will not be affected by a change in 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, by using Lemma 4, for any sequence of strategic decision among players, an acyclic relevance graph[12]. 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 Fig. 9, 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 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 $\boldsymbol{d}^*$ that is a Nash equilibrium of $\mathcal{M}_d$.*

This theorem essentially follows from Algorithm 6.1 and Theorem 6.1 of [17]. Note that, in our scenario, under the 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 7 is similar to the one defined by backward induction in extensive-form games. However, the MAID approach enables us to run inference on $\mathcal{M}_d$ in order to compute the expected utilities

[12] See the definition of a relevance graph in Definition 5.4 of [17].

given the decision rules of every player, and to eventually find a NE in $\mathcal{O}(|\Omega|2^N)$ instead of $\mathcal{O}(|\Omega|3^{2N})$ in the extensive-form game.

We numerically compute the NE of the N-player game G_d by using the Utah family dataset. We assume the sequence of decisions to be the following: GP1, GP2, GP3, GP4, P5, P6, C7, C8, and C9. We skip the details of the algorithm and inference, and we provide the main numerical results. We focus on 1,000 randomly chosen SNPs of chromosome 1,[13], and we 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 Fig. 10), 0.6 (second column of Fig. 10) and 0.8 (third column of Fig. 10). In the first row of Fig. 10, we see the number of players who disclose their genomes at NE. In Fig. 10a, because b_2 is quite small (0.4), if b_1 and b_2 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, this will automatically lead the parents to do the same because their genomic privacy will be reduced by their relatives' decision. We see this in the left strip where $b_3 \geq 0.5$ and $b_1 \leq 0.2$: Five relatives disclose their SNPs, the three children and the two parents. By increasing b_1 to 0.3, then two of the four grandparents have the incentive to share their SNPs, considering their privacy levels. We notice that when b_2 increases to 0.6 (Fig. 10b) and 0.8 (Fig. 10c), 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 two 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 the 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 Fig. 10d and Fig. 10e 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 created by the parents disclosing their SNPs, whereas grandparents and children have no incentives to do the same. Hence, misaligned incentives have a negative impact on the social welfare of a family. In future work, we intend to extend this model to altruistic players and see if this improves the global outcome. 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. In future work, we plan to extend our current model in order to take into account the inherent dynamic nature of life.

[13] As in Section 5, LD is not used as we assume the same set Ω of SNPs potentially shared by the players and targeted by the adversary.

7 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 [23]. Henne et al. study how OSN pictures uploaded by friends can reveal information about one's own location [14]. Dey et al. analyze the risk of age inference in OSNs, notably by relying on information posted by users' friends and friends-of-friends [8]. 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 the location privacy of others [28]. Olteanu et al. study how users reporting co-locations with other users (e.g., on online social networks) can decrease others' location privacy [24]. In order to precisely quantify the effect of co-location information, they propose an optimal inference algorithm and two polynomial-time approximate inference algorithms. Humbert et al. propose a framework to quantify the damage to genomic privacy caused by relatives [15]. We extend this framework to study the interplay between rational agents with different motivations and utilities related to their genomic privacy, considering selfish and altruistic behaviors.

Acquisti et al. were among the first to propose an economic model for formalizing incentives and interactions between rational agents in the context of privacy [3]. 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, by using game theory, the behavior of selfish mobile nodes that want to protect their location privacy at a minimum cost [11]. 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 [6]. They show how negative externalities can lead to inefficient equilibria in scenarios where two users decide about the adoption of an app. Pu and Grossklags go one step further by studying large groups of users who take others' preferences into account when making their own decisions [25]. These works build upon the literature on IDS games, surveyed in [20]. We follow a similar approach for genomic privacy. In addition, precisely quantify by using real data the possible direct and indirect privacy losses with a probabilistic framework. The non-linear dependencies between players in genomic privacy are also novel compared to previous work.

8 Conclusion and Future Work

In this work, focusing on the privacy of genomic data, we have studied the strategic decisions of family members about whether to disclose their genomes and how to secure their storage on personal devices. By using a game-theoretic approach, we have modeled the interplay between family members with different incentives and have predicted their behaviors at equilibrium. First, we extensively studied

a two-player game between two either selfish or altruistic family members. Then, using multi-agent influence diagrams we have extended this to an N-player game. We believe that the proposed models can help the family members choose how to protect the privacy of their genomic data while still helping medical research and benefiting from the merits of genomics. In future work, we will study games with altruistic behaviors in the N-player game.

References

1. http://www.vox.com/2014/9/9/5975653/with-genetic-testing-i-gave-my-parents-the-gift-of-divorce-23andme.
2. http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=191486604.
3. A. Acquisti, R. Dingledine, and P. Syverson. On the economics of anonymity. In *Financial Cryptography*, pages 84–102. Springer, 2003.
4. R. Anderson and T. Moore. The economics of information security. *Science*, 314(5799):610–613, 2006.
5. 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.
6. G. Biczók and P. H. Chia. Interdependent privacy: Let me share your data. In *Financial Cryptography and Data Security*, pages 338–353. Springer, 2013.
7. E. De Cristofaro, S. Faber, P. Gasti, and G. Tsudik. Genodroid: Are privacy-preserving genomic tests ready for prime time? In *ACM WPES*, 2012.
8. R. Dey, C. Tang, K. Ross, and N. Saxena. Estimating age privacy leakage in online social networks. In *IEEE INFOCOM*, 2012.
9. 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.
10. Y. Erlich and A. Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014.
11. J. Freudiger, M. H. Manshaei, J.-P. Hubaux, and D. C. Parkes. On non-cooperative location privacy: A game-theoretic analysis. In *ACM CCS*, 2009.
12. J. Grossklags, B. Johnson, and N. Christin. The price of uncertainty in security games. In *Economics of Information Security and Privacy*. Springer, 2010.
13. M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science: 339 (6117)*, Jan. 2013.
14. 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 *ACM WiSec*, 2013.
15. M. Humbert, E. Ayday, J. P. Hubaux, and A. Telenti. Addressing the concerns of the Lacks Family: Quantification of kin genomic privacy. In *ACM CCS*, 2013.
16. M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Reconciling utility with privacy in genomics. In *ACM WPES*, 2014.
17. D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003.
18. E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *STACS*, 1999.
19. H. Kunreuther and G. Heal. Interdependent security. *Journal of Risk and Uncertainty*, 26(2-3):231–249, 2003.

20. A. Laszka, M. Felegyhazi, and L. Buttyán. A survey of interdependent security games. *CrySyS Lab Technical Report No. CRYSYS-TR-2012-11-15*, 2012.
21. S. L. Lauritzen and N. A. Sheehan. Graphical models for genetic analyses. *Statistical Science*, pages 489–514, 2003.
22. D. Meier, Y. A. Oswald, S. Schmid, and R. Wattenhofer. On the windfall of friendship: Inoculation strategies on social networks. In *ACM EC*, 2008.
23. A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in online social networks. In *ACM WSDM*, 2010.
24. A. M. Olteanu, K. Huguenin, R. Shokri, and J.-P. Hubaux. Quantifying the effect of co-location information on location privacy. In *PETS*, 2014.
25. Y. Pu and J. Grossklags. An economic model and simulation results of app adoption decisions on networks with interdependent privacy consequences. In *Decision and Game Theory for Security*, pages 246–265. Springer, 2014.
26. 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.
27. L. Sweeney, A. Abu, and J. Winn. Identifying participants in the personal genome project by name. *Available at SSRN 2257732*, 2013.
28. 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.