Simulating the Large-Scale Erosion of Genomic Privacy Over Time

Michael Backes, Pascal Berrang, Mathias Humbert, Xiaoyu Shen, and Verena Wolf

Abstract—The dramatically decreasing costs of DNA sequencing have triggered more than a million humans to have their genotypes sequenced. Moreover, these individuals increasingly make their genomic data publicly available, thereby creating privacy threats for themselves and their relatives because of their DNA similarities. More generally, an entity that gains access to a significant fraction of sequenced genotypes might be able to infer even the genomes of unsequenced individuals. In this paper, we propose a simulation-based model for quantifying the impact of continuously sequencing and publicizing personal genomic data on a population's genomic privacy. Our simulation probabilistically models data sharing and takes into account events such as migration and interracial mating. We exemplarily instantiate our simulation with a sample population of 1,000 individuals and evaluate the privacy under multiple settings over 6,000 genomic variants and a subset of phenotype-related variants. Our findings demonstrate that an increasing sharing rate in the future entails a substantial negative effect on the privacy of all older generations. Moreover, we find that mixed populations face a less severe erosion of privacy over time than more homogeneous populations. Finally, we demonstrate that genomic-data sharing can be much more detrimental for the privacy of the phenotype-related variants.

Index Terms—Genomic privacy, simulations, inference, graphical models

1 Introduction

Since the first sequencing of the human genome in 2001, at least a million humans have had their DNA genotypes sequenced [1]. The rapidly decreasing costs of DNA sequencing will ensure that this number keeps rising, presumably at a much higher pace than ever before. Moreover, individuals increasingly share their genomic data publicly, e.g., to help medical research. For example, there are already thousands of genotypes available on the OpenSNP platform [2]. In addition to such open platforms, popular genotyping service providers such as 23andMe already possess millions of individuals' genotypic data and are sharing them with third parties such as pharmaceutical companies [3], [4]. Furtermore, portable sequencing sensors such as minION promise to pioneer fast and pervasive DNA sequencing [5], [6]. Finally, the whole genomes of significant subsets of individuals from specific populations are now available [7].

This increasingly comprehensive, widely available genomic information bears great promise for medical research and for becoming the key enabler for highly personalized medical treatments. But it also comes with unprecedented privacy risks not only for the individuals that sequenced their DNA [8], [9], [10], but also for their relatives because of their DNA similarities [11]. Hence, we, in particular, encounter the problem that even the privacy of those individuals who decide not to sequence their DNA is affected by other sequencings, and that an entity that gains access to a significant fraction of genomes from a given population might be able to probabilistically infer the unsequenced genomes from publicly available data.

The goal of this paper is to simulate the erosion of genomic privacy over time. More precisely, we aim at quantifying the effect of continuous large-scale sequencing of genotypes on the privacy of a population under various realistic scenarios. First of all, we evaluate the impact of individuals sharing their genomes on the privacy of others based on a probabilistic population model. Second, we assess the influence on the genomic privacy of geopolitical events, such as migration, and of sociological parameters such as the fraction of interracial mating. To the best of our knowledge, this is the first work to assess the large-scale erosion of genomic privacy over time.

We run our simulations on a sample population of 1,000 individuals distributed over five generations. First, we evaluate the evolution of genomic privacy on 6,000 genomic variants located on chromosome 19. We note that the global population's genomic privacy erodes superlinearly in the sharing rate, i.e., the sharing behavior of others has a detrimental effect on the privacy of everyone. We also observe that an increasing sharing rate of genomic data in the future can also have a substantial negative effect on the privacy of all previous generations. Moreover, we find that mixed populations, due to their large genomic diversity, face a less severe erosion of genomic privacy over time than more homogeneous populations. However, the average genomic privacy level is already quite low without any observed

- *M. Backes is with CISPA, Helmholtz Center i.G., Saarbrücken 66123, Germany, and Saarland University, Saarbrücken 66123, Germany. E-mail: backes@cispa.saarland.*
- *P. Berrang, X. Shen, and V. Wolf are with CISPA, Saarland University, Saarbrücken 66123, Germany. E-mail: cispa@paberr.net, s8xishen@stud.uni-saarland.de, verena.wolf@uni-saarland.de.*
- *M. Humbert is with the Swiss Data Science Center, ETH Zurich and EPFL, Switzerland. E-mail: mathias.humbert@epfl.ch.*

Manuscript received 18 June 2018; accepted 21 June 2018. Date of publication 24 July 2018; date of current version 5 Oct. 2018.
(Corresponding author: Pascal Berrang.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2859380

data (baseline). This can be explained by the fact that most of the population carries the same variants in general. Finally, focusing on a subset of sensitive variants (e.g., correlated with a disease), we observe that, for most of these variants, the baseline genomic privacy is much higher–success rate is lower–than the one with all variants of chromosome 19. As a consequence, the decrease in privacy induced by genomic-data sharing is much more significant in the case of these very sensitive variants, calling for a more cautious behavior regarding individual data sharing.

The paper is organized as follows. In Section 2, we introduce the various population parameters used throughout the paper, as well as the adversarial model. In Section 3, we describe how we simulate a large population, and how we efficiently infer hidden genomic data in this population. In Section 4, we present different realistic instantiations of our population model and their corresponding results. We review the related literature in Section 5 before concluding in Section 6.

2 Population and Threat Models

We consider a probabilistic population model over a variable number of k generations. Starting with generation 0, which consists of n_f individuals–so called founders–, the individuals then mate, have children and share their genome with a certain probability. We introduce, hereafter, all parameters used in our simulations.

Birthrate. The number of children of a couple in our population is randomly determined based on a Poisson distribution with mean λ (known as the birthrate). The Poisson distribution is used rather than the Gaussian distribution because it generates discrete and positive values only.

Sharing Genomic Data. $\gamma(i)$ represents the proportion of individuals in the ith generation of the population who have their DNA sequenced and who share it online or with strong attackers such as prominent direct-to-consumer companies having access to millions of genotypes in their database (e.g., Ancestry). Since it is most likely that this proportion will increase in future generations, we allow instantiations of this parameter to depend on the actual generation. The proportion may range from $\gamma(i) = 0$ if no individual has his/her DNA sequenced and shared to $\gamma(i) = 1$ if everyone in this generation shares his/her genomic data.

Mating Behavior. Our model forbids mating between individuals up to kinship degree 2, including sisters, brothers, and cousins. To account for interracial mating, α represents the probability of an individual mating an individual from a different ethnical group. So, $\alpha = 0$ means there is no interracial mating, whereas $\alpha = 0.5$ means that chances are equally high that the partner is either randomly chosen from the individuals of the same ethnicity or from the individuals of any other ethnicity. As we focus in this work on autosomal chromosomes (non-sexual chromosomes), we do not distinguish males and females when selecting partners, for simplicity. This does not have any impact on the privacy of all non-sexual chromosomes.

Immigration. The last but not least relevant population parameter we explore is the degree of population diversity stemming from immigration. δ represents the immigration rate, that we define as the proportion of immigrants per generation (relative to the current generation's population).

Adversarial Model. We assume the adversary can gain access to a significant fraction of sequenced genomes, be it because they are publicly available or because of access to the databases of direct-to-consumer testing companies such as 23andMe or Ancestry (which own millions of genotypes already). Moreover, we assume the adversary can gather background knowledge on the family relationships, e.g., from genealogical databases or online social networks.

3 Computational Model

In order to assess the genomic privacy erosion at large, we rely on simulations, since this is the standard approach for complex models of population dynamics. It ensures scalability and easy adaption and extension. The simulations are split into two separate steps, namely (1) generating the population and (2) calculating the extent to which the unobserved genomes of the population can be inferred from the observed genomic data.

3.1 Generating Populations

The first step aims at generating a realistic population and its genomes, based on the mating, birthrate, and immigration parameters presented in the previous section. To this end, we construct a large pedigree of k generations based on n_f founders in the 0th generation by successively generating future generations. For each generation except the first (i.e., $i > 0$), we add immigrants to our population according to the immigration rate δ.

Then, we choose partners for as many individuals as possible. The partner is randomly chosen from the set of individuals from the same origin (excluding those up to relationship degree 2) with probability $1 - \alpha$ and from the set of individuals of other ethnicities with probability α. Each pair of individuals has c children, where c is randomly sampled from the Poisson distribution with mean λ. Following the Mendelian inheritance laws, we sample the genome of each child independently from the genomes of the parents. More precisely, we first randomly pick one of the two alleles at a given position of the mother, then we repeat this uniform sampling for the father, and finally merge the two resulting alleles together to derive the child's base pair at this position. We repeat this process over the whole varying (i.e., polymorphic) positions in the genome, independently from any other position. We do not take into account the linkage disequilibrium structure when creating the next generation's offsprings. This modeling assumption allows us to generate populations of thousands of individuals very efficiently.

Fig. 1 shows a sample pedigree of $k = 5$ generations based on $n_f = 200$ founders. If we assume that the founders are of different ethnicities than the immigrants, it can be easily recognized that $\alpha > 0$, because there are immigrants mating with descendants of the founders. Note that, since the illustration only shows a subset of all individuals, some arrows have been omitted.

3.2 Inferring Hidden Genomes

In the second step, we assume that a certain percentage of people in the whole population get their genomes sequenced, and that they release them according to the parameter $\gamma(i)$. We thus randomly select a fraction of $\gamma(i) \cdot |Y_i|$ individuals

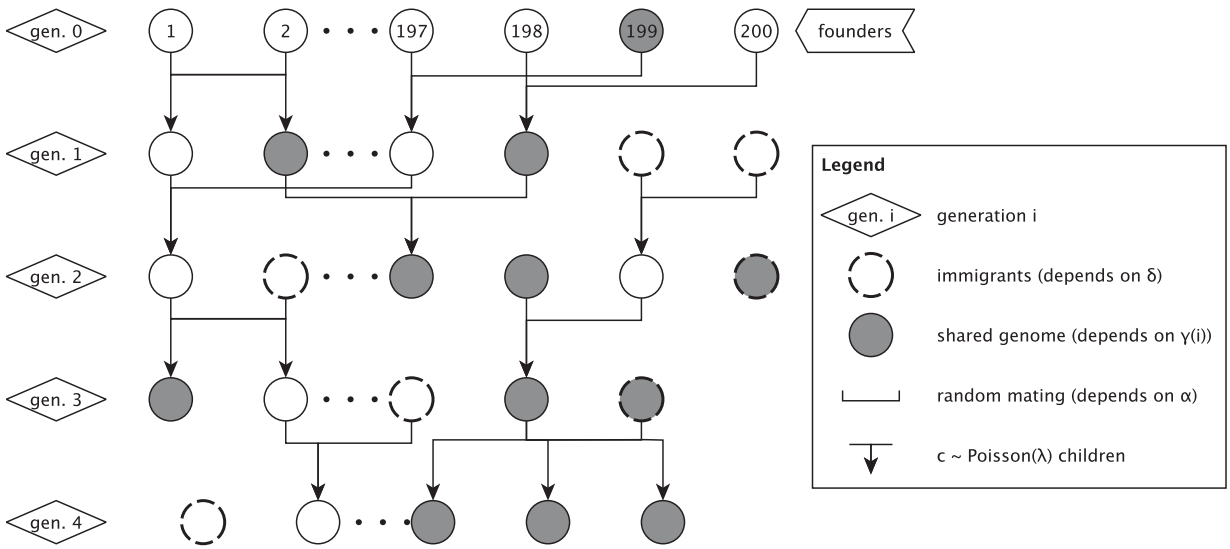


Fig. 1. A sample population annotated with the different parameters of our model.

from the population Y_i at the ith generation. These selected genomes are then assumed to be observed by the adversary. We represent these observed genomic data with the random variable $\mathbf{X}_{\text{obs}}$.

In order to efficiently infer the rest of the population's genomes (assumed to not be disclosed) based on the observed genomes, we rely upon the belief propagation algorithm (also called message-passing or sum-product algorithm). This algorithm propagates evidence (i.e., observed genomes) to other variables (i.e., unobserved genomes) in a Bayesian network encompassing the dependencies between the individuals' genomes [12], [13], [14]. Bayesian networks are probabilistic graphical models that allow to represent conditional (in)dependencies between random variables. These models are especially well suited for pedigrees as inheritance laws induce conditional independencies between someone's genome and all his ancestors given his parents' genome [15], [16].

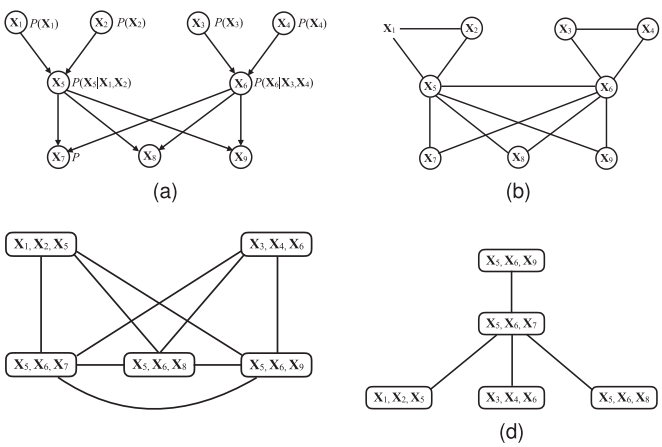


Fig. 2. Transformation of a Bayesian network into a junction tree: Example with a three-generation pedigree containing four grandparents, two parents, and three children. (a) Original Bayesian network and its nodes' probabilities (whose product is equal to the global distribution shown in Formula 1). (b) Moral graph obtained by transforming all directed edges into undirected ones, and by connecting parents together. (c) Clique graph obtained by clustering nodes belonging to the same cycle together. (d) Junction tree constructed by forming a maximum spanning tree of cliques.

Assuming $P(\mathbf{X})$ represents the joint probability distribution of all the m genomic variants of n individuals (where n is the size of the population) in our simulation, inference is in general exponential in $n \times m$, which is computationally intractable when n and m are large. However, due to the Mendelian inheritance laws, and under the assumption that the variants are independent of each other, we can split this global joint distribution into smaller local probability functions

$$P(\mathbf{X}) = \prod_{g_i \in \mathcal{G}} \prod_{r_j \in founders} P(\mathbf{X}_j^i) \prod_{r_k \in \mathcal{R} \setminus founders} P(\mathbf{X}_k^i \,|\, \mathbf{X}_{m(k)}^i, \mathbf{X}_{f(k)}^i), \quad (1)$$

where $\mathcal{G}$ is the set of genomic variants, $\mathcal{R}$ is the set of individuals in the population, and $m(k)$ and $f(k)$ are the mother and the father of individual r_k.

This factorization allows us to deal with much smaller probability distributions, represented by n nodes in m independent Bayesian networks. As shown in Fig. 2a, in every Bayesian network, the n (equal to 9 in the figure) nodes are connected to each other by directed edges representing the conditional probability $P(\mathbf{X}_k^i \,|\, \mathbf{X}_{m(k)}^i, \mathbf{X}_{f(k)}^i)$ given by Mendelian laws. The concrete values of this conditional probability are depicted in Table 1. Each child node in the graph has two parent nodes, exactly like in real biological life. Only the founders have no parent in the Bayesian network. For those, the probability of each variant is given by the prior probability $P(\mathbf{X}_j^i)$. If the value $\mathbf{X}_j^i$ is not observed (i.e., shared) this probability is typically given by the minor allele frequencies. Concretely, assuming the minor allele frequency of SNP g_i in a given population is equal to f_i, the prior probability is defined as: $P(\mathbf{X}_j^i = 0) = (1 - f_i)^2$, $P(\mathbf{X}_j^i = 1) = 2f_i(1 - f_i)$, and $P(\mathbf{X}_j^i = 2) = f_i^2$.

Because of the sibling relationships, the Bayesian network representing the general population contains undirected cycles. In order to remove these loops, we transform the original Bayesian network into a junction tree, or clique tree, as shown in Fig. 2. It is worth noting that, although the step from the moral graph to the clique graph is in general computationally hard, in our case, the cliques are straightforwardly created by merging each child node with its two

TABLE 1
Conditional Probability Table of $P(\mathbf{X}_k^i \mid \mathbf{X}_{m(k)}^i, \mathbf{X}_{f(k)}^i)$ Appearing in the Right-Hand Product of Formula (1)

		Father		
		$\mathbf{X}_{f(k)}^i = 0$	$\mathbf{X}_{f(k)}^i = 1$	$\mathbf{X}_{f(k)}^i = 2$
Mother	$\mathbf{X}_{m(k)}^i$=0	(1,0,0)	(0.5,0.5,0)	(0,1,0)
	$\mathbf{X}_{m(k)}^i$=1	(0.5,0.5,0)	(0.25,0.5,0.25)	(0,0.5,0.5)
	$\mathbf{X}_{m(k)}^i$=2	(0,1,0)	(0,0.5,0.5)	(0,0,1)

The vth element in each of the table's triplet represents $P(\mathbf{X}_k^i = v | \mathbf{X}_{m(k)}^i, \mathbf{X}_{f(k)}^i)$.

parent nodes. Hence, every child and its parents form a clique of size 3, as shown in Fig. 2c.

On the resulting junction tree, the belief propagation algorithm converges in only two iterations: (i) passing messages upwards, from the leaves of the tree to the root, and (ii) passing messages downward, from the root clique to the leaves. Messages are computed according to the Pearl's belief propagation rules, similarly to Formulas (31)-(33) depicted in [13]. The computational complexity of the belief propagation algorithm is linear in the number of nodes n, in the number of variants m, and exponential in the maximal clique size (also called treewidth). This size is equal to 3 in our case, which is negligible compared to n and m. Therefore, running belief propagation on the junction tree enables us to infer the whole population's genomes with complexity linear in $n \times m$.

Note that the assumption of variants being independent of each other can be justified as we assume here that individuals either release all their genomic data or none. Thus, considering linkage disequilibrium—i.e., dependencies between genomic variants—would not bring much more inference power to the attacker. As the evaluation of [17] shows, the LD correlations have lower impact on privacy when a larger subset of the targeted SNPs are observed in relatives' genomes. We can thus assume that, by observing the full set of considered SNPs of anyone sharing his genome, the LD correlations do not help the adversary improve his inference. This assumption was also made in previous works [18], [19], and it allows us to significantly reduce the computational complexity of our algorithm and make it tractable for thousands of variants and one thousand individuals in the considered population.

The belief propagation algorithm eventually outputs the marginal posterior probabilities of all individuals at every genomic position given the observed genomes, i.e., $P(\mathbf{X}_j^i \mid \mathbf{X}_{\text{obs}})$ for all $g_j \in \mathcal{G}$ and $r_i \in \mathcal{R}$. As suggested by Wagner [20], we rely upon the success of the inference attack, $P(\mathbf{X}_j^i = x_j^i \mid \mathbf{X}_{\text{obs}})$, where x_j^i is the actual value of the variant, as a metric to measure privacy. More precisely, the success rate quantifies the loss of genomic privacy. When we consider multiple variants, we average the success rate over all considered variants. For instance, to measure the success rate of an adversary inferring all variants of an individual r_j, we rely upon the following formula:

$$\frac{1}{|\mathcal{G}|} \sum_{g_i \in \mathcal{G}} P(\mathbf{X}_j^i = x_j^i \mid \mathbf{X}_{\text{obs}}). \tag{2}$$

Fig. 1 exemplarily shows a sample population with individuals who have shared their genomes. In this illustration, we depict a sharing rate that increases from generation to generation. In our simulations, we use the genomic data of the grey nodes to infer the genomes for the rest of the population (white nodes) using our belief propagation algorithm.

4 Simulation Results

In this section, we first introduce concrete instantiations of our parameters, and then present the most interesting findings of our experiments.

4.1 Model Instantiations

For all our simulations, we set the birthrate equal to the official U.S. rate (2012), i.e., $\lambda = 1.88$. As for the sharing rate, we consider two different settings. The first instantiation assumes a uniform sharing rate $\gamma(i) = \gamma_{\text{global}}$ for all generations. We also study the case where younger generations share more data than older ones. In order to simulate this behavior, we assume a linearly increasing sharing rate $\gamma(i) = \frac{i \cdot |Y|}{10 \cdot |Y_i|} \gamma_{\text{global}}$, where $|Y|$ is the size of the whole population. This is equal to $\frac{i \cdot k}{10} \gamma_{\text{global}}$ if the size of the population remains stable over generations. Of course, γ_{global} has to be set according to k such that $\gamma(i)$ never exceeds 1.

Now, we present the various combinations of the other parameters and the underlying populations we consider. We label the combinations using the following scheme:

$$\langle base\ population \rangle - \langle \delta \rangle I - \langle \alpha \rangle M - \langle sharing\ rate \rangle U,$$

$\langle sharing\ rate \rangle$ is set to either uni(form) or lin(ear), as defined above. The base population can either be CEU, which are Americans with European ancestors, or MIX, which are Americans with mixed ancestors (70 percent European, 13 percent Mexican, 12 percent African, 3 percent Chinese, and 2 percent Bangladeshi ancestors). We construct our different populations from founders (generation 0) with real genomic data gathered from the 1,000 Genomes Project [21].

CEU-0I-0.5M-uni:
Homogeneous CEU population, no immigration, uniform sharing rate.

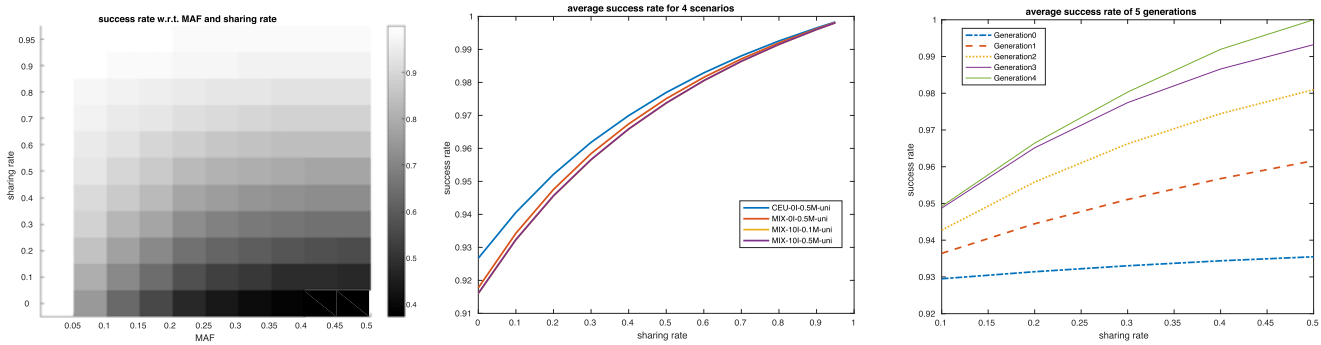
CEU-0I-0.5M-lin:
Homogeneous CEU population, no immigration, linear sharing rate.

MIX-0I-0.5M-uni:
Mixed American population, no immigration, uniform sharing rate.

MIX-10I-0.5M-uni:
Mixed American population, 10 percent immigration rate per generation, random mating, uniform sharing rate. The immigrants are randomly chosen from a population that consists of 10 percent CEU, 10 percent ACB (African Caribbeans in Barbados), 40 percent JPT (Japanese in Tokyo), 40 percent CLM (Colombians from Medellin).

MIX-10I-0.1M-uni:
Mixed American population, 10 percent immigration rate per generation, low interracial mating, uniform sharing rate. The immigrants are randomly selected as above.

CEU-xI-0.5M-lin:
Homogeneous CEU population, x immigration rate per

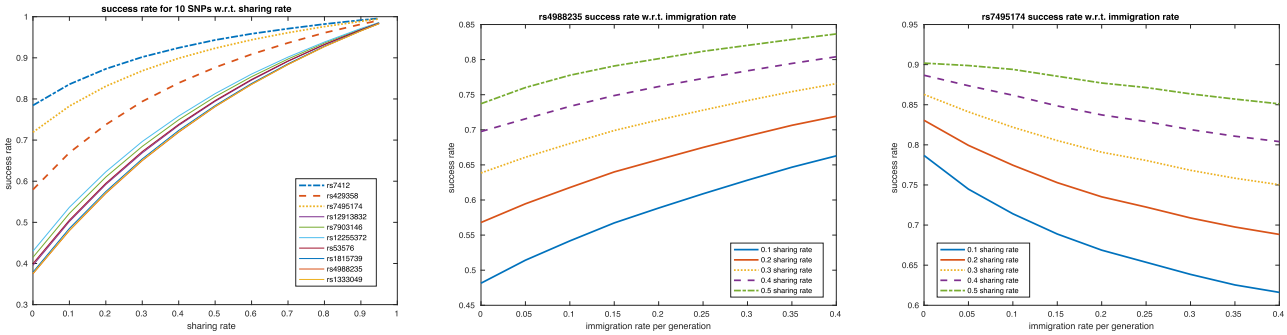


(a) Success rate with varying sharing rates and minor allele frequencies (**CEU-0I-0.5M-uni**).

(b) Success rate with varying sharing rate and four different scenarios, namely **CEU-0I-0.5M-uni**, **MIX-0I-0.5M-uni**, **MIX-10I-0.1M-uni** and **MIX-10I-0.5M-uni.**

(c) Success rate with varying linearly-increasing sharing rate, for all generations (**CEU-0I-0.5M-lin**). Note that generation 0 never shares any data.

Fig. 3. Evolution of privacy with simulations using 6,000 SNPs on chromosome 19.



(a) Success rate of 10 selected SNPs with respect to the sharing rate (**CEU-0I-0.5M-uni**).

(b) Success rate of inferring the SNP rs4988235 (associated with lactose intolerance) depending on the immigration and sharing rates (**CEU-xI-0.5M-lin**).

(c) Success rate of inferring the SNP rs7495174 (associated with eye color) depending on the immigration and sharing rates (**CEU-xI-0.5M-lin**).

Fig. 4. Evolution of privacy with simulations using individual SNPs associated with phenotypes.

generation (varying in the experiment), random mating, linear sharing rate. The immigrants are picked from an EAS (East Asian) population, since this population differs most from the CEU population for the SNPs highlighted in our paper.

Note that we make use of Python to sample the different populations, and of the Bayes Net Toolbox (implemented in Matlab) for the belief propagation algorithm [22].

4.2 Results

We provide here the most interesting findings of our simulations, using first 6,000 SNPs on chromosome 19 (Fig. 3) and then a set of 10 SNPs that are highly variable among populations and are linked to certain phenotypes (Fig. 4). We select $n_f = 200$ founders from the 1,000 Genomes Project and generate 4 additional generations from these individuals, generating a population of around 1,000 individuals. We sample 10 different populations for every settings, and we generate 10 different subset of individuals sharing their genome for a given sharing rate, and average the results.

Fig. 3a depicts the success rate with respect to the minor allele frequencies (MAF) and the sharing rate. The minor allele frequency is defined as the frequency at which the least common allele occurs in a given population. As expected, the success rate monotonically increases with the sharing rate. Moreover, we see that the absolute success increase is higher for SNPs with high MAFs. This holds true since inferring the SNP with high chance is easier if the major allele occurs more frequently within a population by just relying on (public) MAF statistics. It is worth noting here that most of the 6,000 SNPs we use have a low to very low minor allele frequency: Out of 6,000 SNPs, 5,165 have their MAFs between 0 and 0.05, and 228 between 0.05 and 0.1. The other bins (from 0.1 to 0.5) in Fig. 3a only contain between 61 and 102 SNPs. This implies that most of the SNPs on chromosome can be inferred by relying only on the MAFs with high success rate.

Fig. 3b shows the evolution of the success rate, averaged over all 6,000 SNPs, for increasing sharing rate. We observe that the baseline sucess rate is already very high (from around 0.92 to 0.93) for all scenarios, due to the large number of SNPs with low MAFs, confirming our previous finding. Moreover, the homogeneous CEU population gives slightly worse privacy provision than the more mixed populations. However, the interracial mating rate does not have any significant impact on the average privacy. Note that the *MIX-10I-0.1M-uni* curve is not visible, as it is similar to the *MIX-10I-0.5M-uni* curve.

Fig. 3 shows the impact of an increasing sharing rate of younger generations. The x-axis sharing rate is γ_{global} and the founding generation never shares anything. This generation's privacy is nevertheless slightly affected by descendants' sharing behavior. We clearly observe the privacy erosion for

younger generations when sharing increasingly more genomic data.

Next, we focus on a small subset of 10 sensitive SNPs that are linked to various phenotypes, such as diseases, and are listed as "popular" on SNPedia [23]. SNPedia is a website that aggregates current scientific knowledge on the relationship between SNPs and phenotypes. These SNPs consist of 2 SNPs associated with the Alzheimer's disease, 2 associated with eye color, 2 associated with type-2 diabetes, 1 associated with empathy, 1 associated with muscle strength, 1 associated with lactose intolerance and 1 associated with coronary heart disease. It is worth noting that, since some of these SNPs are not part of the 1000 Genomes dataset, we simulate the missing ones by sampling artificial SNPs with the allele frequencies provided on dbSNP [24].

First of all, we notice in Fig. 4a that, despite a very homogeneous population, the baseline success rate is much smaller (around 0.4 for 7 out of 10 SNPs) with these sensitive SNPs than the average over all SNPs on chromosome 19. We also notice in this figure a superlinear increase of the success rate with respect to the sharing rate, for all 10 SNPs. Specifically, we observe that the success rate jumps from 0.4 to 0.8 for 7 out of 10 SNPs if half of the population decides to share his genome.

One of the most interesting parameters for individual SNPs is the immigration rate. Since there are sometimes large differences in the allele frequencies of individual SNPs between populations, genomic privacy can be highly affected by immigration. In general, there is no clear trend on how immigration influences the inference success of individual SNPs: immigration can both increase or decrease the success rate depending on the genetic diversity it brings. Fig. 4b shows the influence of immigration from Eastern Asia (EAS) onto a SNP associated with lactose intolerance. If the amount of immigrating individuals increases, also the inference success of this SNP increases, since its minor allele frequency in the immigrating population is much smaller than the initial (CEU) population. As this population increases the genetic homoeneity, it also increases the success rate, and thus decreases the overall privacy.

On the other hand, Fig. 4c displays the influence of the same immigrating population onto a SNP associated with eye color. Here, the new population brings more genetic diversity into the population, which leads to an enhancement of privacy. Out of the 10 SNPs, 4 fall into the latter category of SNPs where more immigration yields a better global privacy level in the end.

5 Related Work

Inference algorithms based on graphical models have been previously proposed in the context of pedigree analysis. Directed acyclic graphs such as Bayesian networks or hidden Markov models have been proposed to represent multilocus pedigrees [16]. Fishelson and Geiger propose a Bayesian network framework to represent linkage analysis problems, and they compute exact multipoint likelihood by relying on variable elimination [25]. Lauritzen and Sheehan present Bayesian network models, and the detailed junction tree algorithm in the context of genetic analyses [15].

Approximate inference techniques, based on Markov chain Monte Carlo methods, have also been proposed for genetic analyses in the case of complex pedigrees [26], [27], [28]. All the aforementioned models were developed before whole-genome sequencing became affordable, and they were essentially used to infer hidden genotypes given observed phenotypes (such as diseases) in a family. More recently, researchers have proposed methods to infer hidden genotypic data given other observed genomic regions. Genotype imputation relies on intra-genome correlations to complete missing SNPs based on observed genotyped data [29]. Another approach relies on low-resolution genotypes and identity-by-descent genomic regions to infer high-density genotypes in pedigrees [30]. Finally, Kirkpatrick et al. propose to rely on Gibbs sampling for efficiently inferring haplotypes from genotypes in complex pedigrees [31]. The latter approach is approximate but enable to take into account up to 59 individuals in the inference process.

None of the aforementioned related papers addresses privacy issues. However, there has been some previous work on interdependent risks in genomic privacy [11], [17], [19]. Humbert et al. have first proposed to rely on graphical models and belief propagation to quantify kin genomic privacy [11]. In particular, they take into account relatives' genotypes, familial and intra-genome correlations, to infer hidden genomic data and quantify privacy of relatives. Then, they make some independence assumptions between the SNPs within the same genotype, and study the interaction between family members with different preferences regarding the sharing of their genome, and they derive the resulting impact on everyone's privacy [19]. Finally, the authors extend their graphical model to Bayesian networks that account for phenotypic information in addition to genotypes [17]. They also show experimentally that the discrepancy in inference power with and without intra-genome correlations becomes negligible when the adversary gets access to the whole range of targeted SNPs of family members sharing their genomic data.

All aforementioned work on kin genomic privacy focused on a single family, over three generations, considering up to 11 individuals in total. In this work, by using similar models and methods, such as belief propagation, we have shown that the inference algorithm can scale to one thousand individuals and thousands of SNPs. Moreover, by making use of several population parameters, we simulate various populations and thoroughly evaluate how immigration, mating and data-sharing behaviors affect genomic privacy.

6 Conclusion

To the best of our knowledge, this work is the first to propose a framework for predicting the risk of privacy erosion for large populations at a relatively long term. Based on a probabilistic population model, we simulate and quantify the effect of large-scale availability of personal genomic data on the privacy of a large population.

Our findings show that indeed, an increasing proportion of individuals uploading their genomic data threatens not only the privacy of these persons, but also the privacy of the general population. Moreover, we observe that an increasing sharing rate of genomic data in the future can also have a substantial negative effect on the privacy of all older generations. We find that mixed populations can slow down the erosion of genomic privacy over time compared to more homogeneous populations. This effect can be mostly explained by the larger genomic diversity in mixed populations.

Considering a scenario in which nobody shares its genomic data (baseline), the average genomic privacy level is already quite low, since most of the population carries the same variants in general. Thus, individuals sharing their genome especially affects the genomic privacy of variants that are varying a lot within the population. Such variants are often connected to sensitive information such as diseases. Focusing on a subset of such sensitive variants, we observe that, for most of them, the baseline genomic privacy is much higher than the one with all variants of a specific chromosome. Moreover, the effect of sharing the genome is much higher on the genomic privacy of those variants than on the global privacy. We thus conclude that such variants should be cautiously handled, and, if possible, not shared at all.

Our work demonstrates that more research about the implications of large-scale availability of personal genomic data is necessary. Future directions could, for example, include an equational, probabilistic form of a simpler population model. Another promising direction is to incorporate multiple populations and regions with different sharing rates and parameters (e.g., different continents), and more sophisticated immigration models. Finally, other types of biomedical data (such as microRNA or gene expressions) are becoming increasingly available, it would be crucial to evaluate the privacy erosion stemming from sharing those data as well [32], [33], [34].

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0656) and by the German Research Foundation (DFG) via the Collaborative Research Center "Methods and Tools for Understanding and Controlling Privacy" (SFB 1223), project A5.

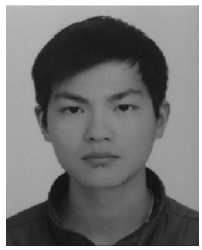
References

[1] Power of one million, (2018, Aug.). [Online]. Available: http://blog.23andme.com/news/one-in-a-million/, Accessed on: Jul. 14, 2016.
[2] OpenSNP, (2018, Aug.). [Online]. Available: https://opensnp.org/genotypes, Accessed on: Jul. 14, 2016.
[3] A. Regalado, (2016, Jun.). [Online]. Available: https://www.technologyreview.com/s/601506/23andme-sells-data-for-drug-search/, Accessed on: Aug. 01, 2018.
[4] How 23andme is monetizing your DNA, (2018, Aug.). [Online]. Available: http://www.fastcompany.com/3040356/what-23andme-is-doing-with-all-that-dna, Accessed on: Jul. 14, 2016.
[5] S. Zaaijer, A. Gordon, R. Piccone, D. Speyer, and Y. Erlich, "Democratizing dna fingerprinting," *bioRxiv*, 2016. [Online]. Available: http://biorxiv.org/content/early/2016/06/30/061556
[6] MinION, (2018, Aug.). [Online]. Available: https://www.nanoporetech.com
[7] D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, et al., "Large-scale whole-genome sequencing of the icelandic population," *Nature Genetics*, vol. 47, no. 5, pp. 435–444, 2015.
[8] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Rev. Genetics*, vol. 15, pp. 409–421, 2014.
[9] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang, "Privacy in the genomic era," *ACM Comput. Surv.*, vol. 48, 2015, Art. no. 6.
[10] E. Ayday, E. De Cristofaro, J.-P. Hubaux, and G. Tsudik, "Whole genome sequencing: Revolutionary medicine or privacy nightmare?" *Comput.*, vol. 48, pp. 58–66, 2015.
[11] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the Lacks family: Quantification of kin genomic privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2013, pp. 1141–1152.
[12] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.
[13] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[14] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Burlington, MA, USA: Morgan Kaufmann Publishers, 1988.
[15] S. L. Lauritzen and N. A. Sheehan, "Graphical models for genetic analyses," *Statistical Sci.*, vol. 18, pp. 489–514, 2003.
[16] M. I. Jordan, "Graphical models," *Statistical Sci.*, vol. 19, pp. 140–155, 2004.
[17] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Quantifying interdependent risks in genomic privacy," *ACM Trans. Privacy Security*, vol. 20, 2017, Art. no. 3.
[18] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool," *Nature Genetics*, vol. 41, no. 9, pp. 965–967, 2009.
[19] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "On non-cooperative genomic privacy," in *Proc. Int. Conf. Financial Cryptography Data Security*, 2015, pp. 407–426.
[20] I. Wagner, "Genomic privacy metrics: A systematic comparison," in *Proc. Int. Workshop Genome Privacy Security*, 2015, pp. 50–59.
[21] 1000 genomes project, (2018, Aug.). [Online]. Available: http://www.1000genomes.org, Accessed on: Sep. 04, 2016.
[22] K. Murphy et al.,, "The bayes net toolbox for Matlab," *Comput. Sci. Statist.*, vol. 33, no. 2, pp. 1024–1034, 2001.
[23] Snpedia, (2018, Aug.). [Online]. Available: http://www.snpedia.com, Accessed on: Sep. 04, 2016.
[24] dbsnp, (2018, Aug.). [Online]. Available: https://www.ncbi.nlm.nih.gov/projects/SNP/, Accessed on: Jan. 12, 2017.
[25] M. Fishelson and D. Geiger, "Exact genetic linkage computations for general pedigrees," *Bioinf.*, vol. 18, no. suppl 1, pp. S189–S198, 2002.
[26] C. S. Jensen, U. Kjærulff, and A. Kong, "Blocking gibbs sampling in very large probabilistic expert systems," *Int. J. Human-Comput. Stud.*, vol. 42, no. 6, pp. 647–666, 1995.
[27] N. Sheehan, "On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees," *Int. Statistical Rev.*, vol. 68, no. 1, pp. 83–110, 2000.
[28] A. Thomas, A. Gutin, V. Abkevich, and A. Bansal, "Multilocus linkage analysis by blocked gibbs sampling," *Statist. Comput.*, vol. 10, no. 3, pp. 259–269, 2000.
[29] Y. Li, C. Willer, S. Sanna, and G. Abecasis, "Genotype imputation," *Annu. Rev. Genomics Human Genetics*, vol. 10, 2009, Art. no. 387.
[30] J. T. Burdick, W.-M. Chen, G. R. Abecasis, and V. G. Cheung, "In silico method for inferring genotypes in pedigrees," *Nature Genetics*, vol. 38, no. 9, pp. 1002–1004, 2006.
[31] B. Kirkpatrick, E. Halperin, and R. M. Karp, "Haplotype inference in complex pedigrees," *J. Comput. Biol.*, vol. 17, no. 3, pp. 269–280, 2010.
[32] E. E. Schadt, S. Woo, and K. Hao, "Bayesian method to predict individual SNP genotypes from gene expression data," *Nature Genetics*, vol. 44, pp. 603–608, 2012.
[33] M. Backes, P. Berrang, A. Hecksteden, M. Humbert, A. Keller, and T. Meyer, "Privacy in epigenetics: Temporal linkability of microRNA expression profiles," in *Proc. 25th USENIX Security Symp.*, 2016, pp. 1223–1240.
[34] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, "Membership privacy in microRNA-based studies," in *Proc. 23rd ACM Conf. Comput. Commun. Security*, 2016, pp. 319–330.

Michael Backes is a full professor at Saarland University. He is the chairman and founding director of the CISPA Helmholtz Center i.G. and head of the Information Security and Cryptography Group. His research covers various aspects of IT security and privacy and ranges from the design, analysis, and verification of protocols and systems, mechanisms for protecting end-user privacy, research on new attack vectors, to universal solutions in software and network security. He is a senior member of the IEEE.

Pascal Berrang received the BS degree in computer science and then started his PhD work under the supervision of Michael Backes in 2014. He is working toward the PhD degree in the Center for IT-Security, Privacy, and Accountability (CISPA), Saarland University, Germany. His research interests broadly cover the areas of IT security, privacy and cryptography, focused on the intersection with bioinformatics and online social networks. He is a student member of the IEEE.

Mathias Humbert completed his PhD thesis on interdependent privacy in the School of Computer and Communication Sciences at EPFL. He is a senior data scientist with the Swiss Data Science Center (ETH Zurich, EPFL). Prior to this, he was a post-doctoral researcher at CISPA, Saarland University (Germany). His research interests lie at the intersection of security and machine learning, with a special focus on bioinformatics. He is a member of the IEEE.



Xiaoyu Shen received the BS degree from Nanjing University, China, in 2015, and studied for one semester (2013-2014) at Utrecht University. He is working toward the master's degree in the Graduate School of Computer Science, Saarland University. His current research interests are machine learning and computational linguistics.

Verena Wolf received the diploma degree in computer science from University Bonn, in 2003 and the PhD degree from University Mannheim, in 2008. She has been a full professor with Saarland University since 2012 and leads the Modelling and Simulation Group with the Department of Computer Science. She is currently working on discrete stochastic modelling as well as efficient simulation methods and has been on the program committees of more than 50 international conferences.