Simulating the Large-Scale Erosion of Genomic Privacy Over Time

Michael Backes[1 2], Pascal Berrang[1], Mathias Humbert[1], Xiaoyu Shen[1], Verena Wolf[1]
[1] CISPA, Saarland University, [2] MPI-SWS

Abstract

The dramatically decreasing costs of DNA sequencing have triggered more than a million humans to date to have their genotypes sequenced. Moreover, these individuals increasingly make their genomic data publicly available, and thereby create unique privacy threats not only for themselves, but also for their relatives because of their DNA similarities. More generally, an entity that gains access to a significant fraction of sequenced genotypes from a given population might be able to infer even the genomes of unsequenced individuals by relying on available data.

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 by individuals and additionally takes into account the influence on genomic privacy of geopolitical events such as migration, and sociological trends such as interracial marriage. We exemplarily instantiate our simulation with a sample population of 1,000 individuals, and evaluate the evolution of privacy under different settings over either thousands of genomic variants or a subset of variants influencing the phenotype. Our findings notably demonstrate that an increasing sharing rate of genomic data in the future entails a substantial negative effect on the privacy of all older 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, even when no data is shared, the genomic privacy averaged over a large number of variants is already very low since mere population allele frequencies already reveal a lot of information about the values of the genomic variants. By focusing on a subset of sensitive variants, we observe a higher genetic diversity in the population. Thus, genomic-data sharing can be much more detrimental for the privacy of the most sensitive variants.

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], and 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 marriages. Note that this is a pioneering work in the sense that it is the first to assess the large-scale erosion of genomic privacy over time. As a consequence, there is no closely related work.

We run our simulations on a sample population of 1,000 individuals distributed over 5 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 older 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 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 than the one with all variants of chromosome 19.

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 Gaussian distribution with mean β (known as the birthrate) and standard deviation 1. We round each number generated with the Gaussian distribution to the closest non-negative integer. Note that the Poisson distribution could also be used as it generates discrete and positive values only.

Sharing Genomic Data. $\gamma(i)$ represents the proportion of individuals in the i-th generation of the population who have their DNA sequenced and who share it online or with strong attackers such as big direct-to-consumer companies having access to millions of genotypes in their database. 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 excludes relationships between individuals up to degree 2, including sisters, brothers, and also 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 all other ethnicities. As we focus in this work on autosomal chromosomes (non-sexual chromosomes), we do not distinguish males and females when selecting partners, for simplicity.

Immigration. The last but not least interesting 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 global players such as 23andMe or Illumina. 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, which 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 genomes of the population can be inferred from the observed data.

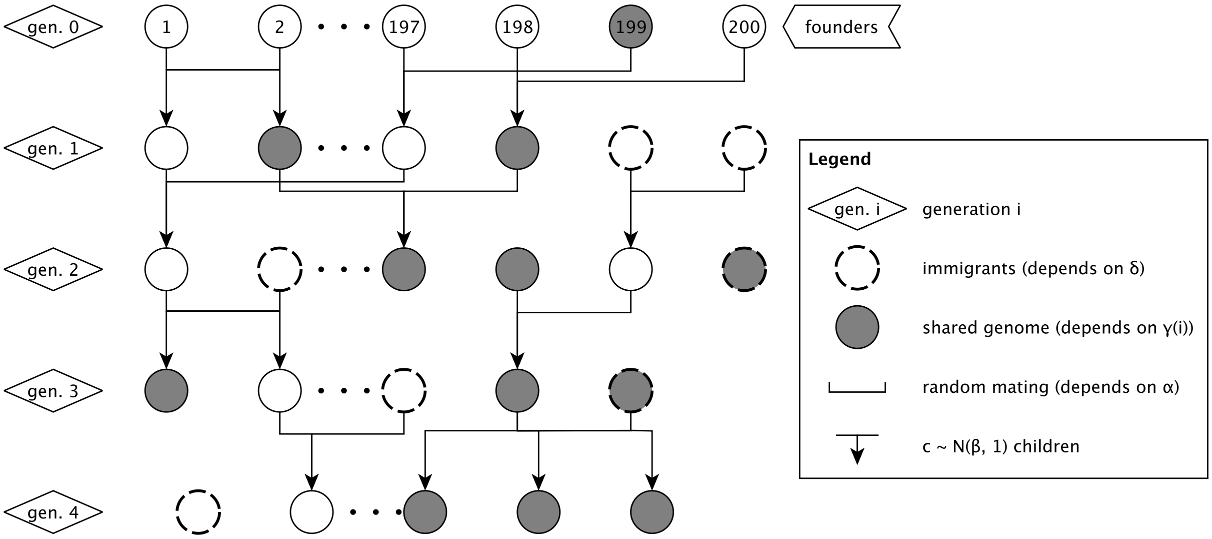


Figure 1: A sample population annotated with the different parameters of our model.

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 ($i > 0$), 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 $\max(0, \lfloor c \rceil)$ children (closest, non-negative integer to c), where c is randomly sampled from the Gaussian distribution $\mathcal{N}(\beta, 1)$. We sample the genome of each child from the genomes of the parents, following Mendelian inheritance laws. 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 local linkage disequilibrium structure when creating the next generation offsprings. This modeling assumption allows us to generate populations of thousands of individuals very efficiently.

Figure 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.

Inferring Hidden Genomes

In the second step, we assume that a certain percentage of people in the whole population get their genomes sequenced and released according to the parameter $\gamma(i)$. We thus randomly select a fraction of $\gamma(i) \cdot |Y_i|$ individuals from the population Y_i at the i-th generation. These selected genomes are then assumed to be observed by the adversary. We represent these observed genomic data as $\mathbf{X}_{\text{obs}}$.

In order to infer the rest of the population's genomes based on the observed genomes, we rely upon the belief propagation algorithm (also called message-passing or sum-product). 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].

If $P(\mathbf{X})$ represents the joint probability distribution of m genomic variants of n individuals (where n is the size of the population), inference is in general exponential in nm, 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:

$$P(\mathbf{X}) = \prod_{g_i \in \mathcal{G}} \prod_{r_j \in \textit{founders}} P(\mathbf{X}^i_j) \prod_{r_k \in \mathcal{R} \setminus \textit{founders}} P(\mathbf{X}^i_k \mid \mathbf{X}^i_{m(k)}, \mathbf{X}^i_{f(k)}), \tag{1}$$

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 k.

This factorization allows us to deal with much smaller probability distributions represented by n nodes in m independent Bayesian networks. In every Bayesian network, the n nodes are connected to each other by directed edges representing the conditional probabilities given by Mendelian laws, each child node in the graph having two parent nodes, exactly like in real biological life. Only the founders have no parent in the Bayesian network.

By transforming the original Bayesian networks into a junction trees, or clique trees (which remove the loops in the original Bayesian networks that appear when there are siblings in the population), the belief propagation algorithm converges in only two iterations (one forward and one backward). It is worth noting that, although this step is in general computationally hard, in our case, the cliques are straightforwardly created by merging each child node with its two parent nodes. So, every child and its parents form a clique of size 3. The computational complexity of the belief propagation algorithm is linear in the number of node n, in the number of variants m, but 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 rinfer the whole population's genomes with complexity linear to nm.

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. This assumption was also made in previous works[15,16], 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., $P(\mathbf{X}^i_j \mid \mathbf{X}_{\text{obs}})$ for all $g_j \in \mathcal{G}$ and $r_i \in \mathcal{R}$. As suggested by Wagner[17], we rely upon the success of the inference attack, $P(\mathbf{X}^i_j = x^i_j \mid \mathbf{X}_{\text{obs}})$, where x^i_j 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:

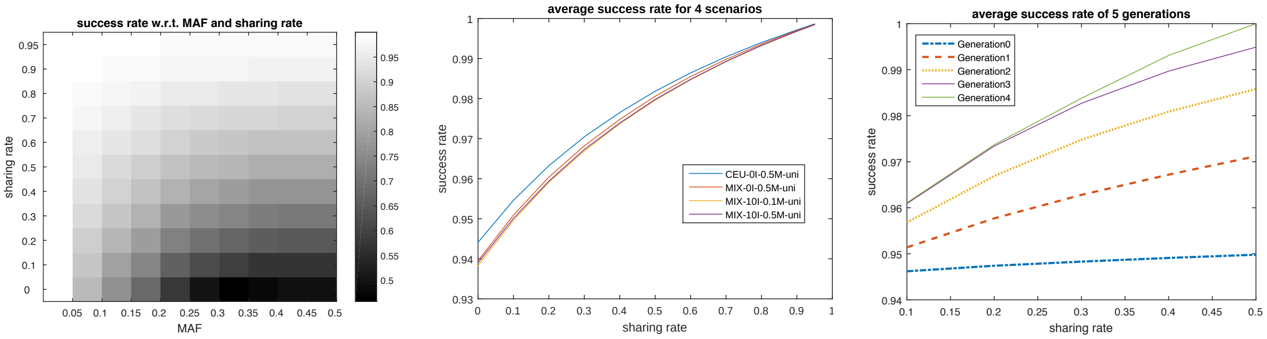
$$\frac{1}{|\mathcal{G}|} \sum_{g_i \in \mathcal{G}} P(\mathbf{X}^i_j = x^i_j \mid \mathbf{X}_{\text{obs}}) \tag{2}$$

Figure 1 also exemplarily shows individuals who have shared their genome. In this illustration, we rely on a sharing rate, which increases from generation to generation. In our simulations, we would then use the genomes of the grey nodes to infer the genomes for the rest of the population 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.

Model Instantiations



(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.

Figure 2: Simulations using 6,000 SNPs on chromosome 19.

For all our simulations, we set the birthrate equal to the official U.S. rate (2012), i.e., $\beta = 1.88$. As for the sharing rate, we consider two different settings. The first instantiation assumes a uniform sharing rate $\gamma(i) = \gamma_{\mathsf{global}}$ 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 $\gamma(i) = \frac{i \cdot |Y|}{10 \cdot |Y_i|}\gamma_{\mathsf{global}}$, where $|Y|$ is the size of the whole population. This is equal to $\frac{i \cdot k}{10}\gamma_{\mathsf{global}}$ 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:

$$\langle \textit{base population} \rangle - \langle \delta \rangle I - \langle \alpha \rangle M - \langle \textit{sharing rate} \rangle U$$

$\langle$*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% European, 13% Mexican, 12% African, 3% Chinese, and 2% Bangladeshi ancestors). We construct our different populations from founders (generation 0) with real genomic data gathered from the 1000 Genomes Project[18].

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.

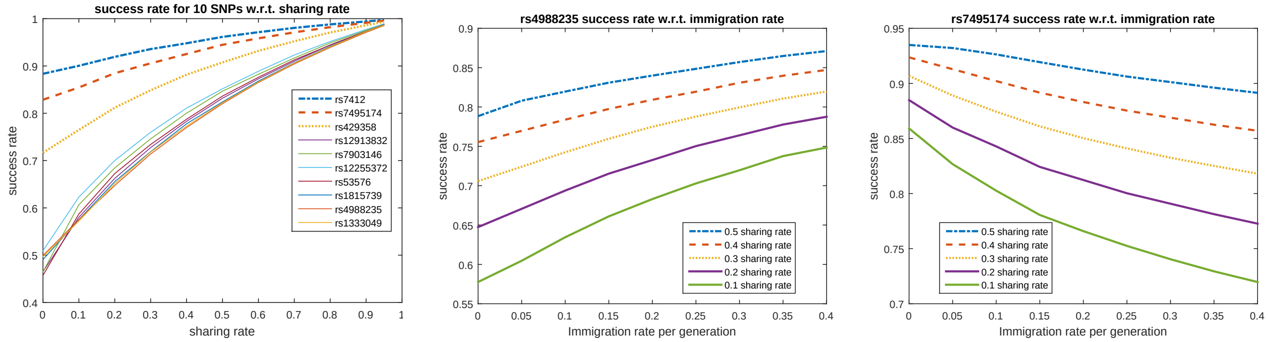
MIX-0I-0.5M-uni Mixed American population, no immigration, uniform sharing rate.

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

MIX-10I-0.1M-uni Mixed American population, 10% 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 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[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**).

Figure 3: Simulations using individual SNPs associated with phenotypes.

Results

We highlight here the most interesting findings of our simulation using first 6,000 SNPs on chromosome 19 (Figure 2) and then a set of 10 SNPs that are highly variable among populations and are linked to certain phenotypes (Figure 3). We select $n_f = 200$ founders from the 1000 Genomes Project and generate 4 additional generations from these individuals leading us to 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.

Figure 2a 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 Figure 2a only contain between 61 and 102 SNPs.

Figure 2b 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 (around 0.94) for all scenarios, due to the large number of SNPs with low MAFs. 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.

Figure 2c 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[20] that aggregates current 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 these missing ones by sampling artificial SNPs with the allele frequencies provided on dbSNP.

First of all, we notice in Figure 3a that, despite a very homogeneous population, the baseline success rate is much smaller (around 0.5 for 7 out of 10 SNPs) with these sensitive SNPs than the average over all SNPs on chromosome 19. 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. Figure 3b shows the influence of immigration from EAS (Eastern Asia) 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. Figure 3c, on the other hand, displays the influence of the same immigration onto a SNP associated with eye color. Here, the new diversity brought into the population 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 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.

Hence, 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 one chromosome. Moreover, the effect of sharing the genome is much higher on the genomic privacy of those variants than on the global privacy.

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 model. 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[21,22,23].

Acknowledgements

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. `http://blog.23andme.com/news/one-in-a-million/`, . Accessed: 2016-07-14.

2. OpenSNP. `https://opensnp.org/genotypes`. Accessed: 2016-07-14.

3. Regalado Antonio. 23andme sells data for drug search. *MIT Technology Review*, 2016.

4. How 23andme is monetizing your DNA. `http://www.fastcompany.com/3040356/what-23andme-is-doing-with-all-that-dna`, . Accessed: 2016-07-14.

5. Zaaijer Sophie, Gordon Assaf, Piccone Robert, Speyer Daniel and Erlich Yaniv. Democratizing dna fingerprinting. *bioRxiv*, 2016. doi: 10.1101/061556. URL `http://biorxiv.org/content/early/2016/06/30/061556`.

6. MinION. `https://www.nanoporetech.com`.

7. Gudbjartsson Daniel F, Helgason Hannes, Gudjonsson Sigurjon A, Zink Florian, Oddson Asmundur, Gylfason Arnaldur et al. Large-scale whole-genome sequencing of the icelandic population. *Nature genetics*, 47(5):435–444, 2015.

8. Erlich Yaniv and Narayanan Arvind. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15:409–421, 2014.

9. Naveed Muhammad, Ayday Erman, Clayton Ellen W, Fellay Jacques, Gunter Carl A, Hubaux Jean-Pierre et al. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48:6, 2015.

10. Ayday Erman, De Cristofaro Emiliano, Hubaux Jean-Pierre and Tsudik Gene. Whole genome sequencing: Revolutionary medicine or privacy nightmare? *Computer*, pages 58–66, 2015.

11. Humbert Mathias, Ayday Erman, Hubaux Jean-Pierre and Telenti Amalio. Addressing the concerns of the Lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1141–1152. ACM, 2013.

12. Koller Daphne and Friedman Nir. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

13. Kschischang Frank, Frey Brendan and Loeliger Hans-Andrea. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47, 2001.

14. Pearl Judea. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988.

15. Sankararaman Sriram, Obozinski Guillaume, Jordan Michael I and Halperin Eran. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

16. Humbert Mathias, Ayday Erman, Hubaux Jean-Pierre and Telenti Amalio. On non-cooperative genomic privacy. In *International Conference on Financial Cryptography and Data Security*, pages 407–426. Springer, 2015.

17. Wagner Isabel. Genomic privacy metrics: A systematic comparison. *International Workshop on Genome Privacy and Security*, 2015.

18. 1000 genomes project. `http://www.1000genomes.org`. Accessed: 2016-09-04.

19. Murphy Kevin and others . The bayes net toolbox for Matlab. *Computing science and statistics*, 33(2):1024–1034, 2001.

20. Snpedia. `http://www.snpedia.com`. Accessed: 2016-09-04.

21. Schadt Eric E, Woo Sangsoon and Hao Ke. Bayesian method to predict individual snp genotypes from gene expression data. *Nature genetics*, 44:603–608, 2012.

22. Backes Michael, Berrang Pascal, Hecksteden Anne, Humbert Mathias, Keller Andreas and Meyer Tim. Privacy in epigenetics: Temporal linkability of microrna expression profiles. In *25th USENIX Security Symposium*, 2016.

23. Backes Michael, Berrang Pascal, Humbert Mathias and Manoharan Praveen. Membership privacy in microRNA-based studies. In *23rd ACM Conference on Computer and Communications Security*, 2016.