Genomic Privacy Metrics: A Systematic Comparison

Isabel Wagner
Dept. of Computer Science, University of Hull, UK
i.wagner@hull.ac.uk

Abstract—The human genome uniquely identifies, and contains highly sensitive information about, individuals. This creates a high potential for misuse of genomic data (e.g., genetic discrimination). This paper investigates how genomic privacy can be measured in scenarios where an adversary aims to infer a person's genome by constructing probability distributions on the values of genetic variations. Specifically, we investigate 22 privacy metrics using adversaries of different strengths, and uncover problems with several metrics that have previously been used for genomic privacy. We then give suggestions on metric selection, and illustrate the process with a case study on Alzheimer’s disease.

I. INTRODUCTION

Full sequencing of the human genome was first completed by Celera, Inc in 2001, at a cost of about 300 million dollars. Today, next-generation DNA sequencers output genomic data at a cost under $10,000 per genome [1]. This advance has led to a rapid increase in the number of applications that use genomic data, for example in health care (e.g., personalized medicine and pharmacogenomics), research (e.g., genome-wide association studies that correlate the appearance of diseases with specific locations in the genome), and forensics (e.g., paternity tests).

Although most of these applications are highly beneficial, the wide availability of genomic data raises privacy concerns, because a genome sequence uniquely identifies an individual. Possible violations of genomic privacy range from the re-identification of anonymous participants in genome-wide association studies (possibly revealing a person’s disease status) to genetic discrimination (for example, denial of insurance because of an unfavorable genetic predisposition). Moreover, because related individuals have similar genomes, sensitive information can be inferred not only about an individual but also about her/his kin. Despite all these risks, technologies that preserve genomic privacy, and genomic privacy metrics, are still in their infancy.

Genomics. Although the human genome consists of about three billion DNA base pairs, variations between the genomes of individuals occur only in about 112 million bases1, or 0.5% of the genome. These variations are called single nucleotide polymorphisms (SNPs, pronounced "snips"). In the simplest case, a SNP consists of one base pair, and a person has two versions of it, called alleles: one inherited from the mother, the other from the father. The major allele is the variant that is most common among the population, while the minor allele is less common. The value of an SNP can be encoded as 0, 1, or 2, representing that at this SNP location a person has either two major alleles; one each of major and minor allele; or two minor alleles [2]. Population-wide frequencies of major and minor alleles can be computed from a sample of human genomes; this has been done by the 1000 Genomes project2. SNPs that cause a disease or increase the risk for it can be discovered by genome-wide association studies which compare the incidence of genetic variations between a group of individuals who have the disease and a healthy control group.

In this paper, we assume an adversary who has targeted an individual and aims to infer as many of the target’s SNPs as possible. We assume that the adversary applies some kind of inference attack which allows him to compute a probability distribution for each SNP. This is a reasonable assumption, because several inference attacks to achieve this have already been published, for example exploiting SNPs that are in linkage disequilibrium [3], exploiting SNPs from kin genomes [2], [4], exploiting systematic execution of genomic tests [5], and using statistical analyses to infer the presence of individuals in genome-wide association studies [6], [7].

Contributions. We systematically evaluate the privacy levels indicated by 22 different privacy metrics for adversaries of different strengths. We choose the adversary’s probabilities so that we can intuitively order the adversaries: an adversary is stronger if his estimates are closer to the true value. We evaluate whether the metrics can distinguish between different strengths at all, and whether their values are consistent with the intuitive ordering. Using the risk of late-onset Alzheimer’s disease as a case study, we illustrate the process of measuring privacy in a real-world scenario.

Combining insight from the systematic evaluation and the case study, we find several metrics that have consistent privacy levels and whose values have an intuitive interpretation: the adversary’s success rate, the amount of information leaked, health privacy (with information surprisal or relative entropy as base metric), information surprisal, relative entropy, and user-specified innocence. We recommend to always use multiple metrics to gain insight about privacy from as many different angles as possible.

Our systematic comparison of metrics for genomic privacy is an important step in genomic privacy research, because it enables researchers to make an informed choice about the selection of privacy metrics. It will thus help ensure that new privacy enhancing technologies (PETs) are evaluated in a consistent and comparable manner.

II. RELATED WORK

Evaluation of metrics. Among the literature in genomic privacy and the wider privacy literature, there are few studies that systematically evaluate privacy metrics. The closest to

---


2 http://www.1000genomes.org/
our work is [8], which investigates the behavior of anonymity metrics, among them entropy and some of its variations.

Genomic privacy metrics. In the context of genomic privacy, there are papers that apply existing metrics to genomic privacy scenarios [2], [3], [9], and papers that propose new metrics specific to genomic privacy [2], [3], [10]. These papers generally propose or describe one or more metrics, and then use these metrics to evaluate a PET in a given scenario. However, they do not evaluate how the metrics behave, or how they differ from other metrics. This paper aims to address this gap.

Requirements for privacy metrics. Several papers have proposed quality criteria for privacy metrics. They should be understandable by laypeople and indicate how effectively the adversary can succeed [11]; show both the privacy level and the portion of data not hidden [12]; consider accuracy, uncertainty, and correctness as three aspects of the adversary’s success [13]; and indicate not only the difficulty for the adversary, but also the amount of resources he needs to succeed [14].

In this paper, we focus on metrics that quantify the privacy level from a user’s perspective, assuming that the adversary already has a probability distribution on SNP values. The requirements relevant for this focus include the privacy level, the portion of data not hidden, and understandability.

The other requirements are relevant when the focus is on quantifying privacy from the adversary’s point of view. We do not investigate these in detail. In short, the adversary’s uncertainty and accuracy can be estimated with entropy and by constructing a confidence interval for the results of an attack, respectively [13]. The difficulty for the adversary and the resources he needs to conduct his attack are important aspects to judge the viability of an attack. However, resource-based metrics have only recently been introduced in the context of anonymity systems [15]; it is unclear how they may be applied to genomic privacy.

III. PRIVACY METRICS

We evaluated a set of 22 metrics that have been proposed as privacy metrics in the literature. The formal definitions are summarized in Table I, and Table II provides a reference for notation used. Eight metrics have already been used to study genomic privacy; the remaining metrics have been drawn from the wider privacy literature (see the Genomics Precedent column in Table I). The metrics can be grouped into per-SNP metrics that compute values for each SNP separately, and per-individual metrics that compute an aggregate value for all of an individual’s SNPs (see the per SNP column).

A. Excluded Metrics

We excluded a range of privacy metrics that may be useful in some scenarios, but do not fit our assumptions.

Differential privacy [29] offers privacy guarantees for database queries. However, our scenario assumes that the adversary is already one step further in that he has already acquired a probability distribution on the target’s SNP values. While differential privacy will not help evaluate privacy in our scenario, it could be used to prevent the adversary from acquiring a probability distribution in the first place. $k$-anonymity [30] states that an individual cannot be distinguished among at least $k-1$ other individuals. Since we assume that the adversary already knows the target individual, we know that $k = 1$, and so this metric does not help us analyze privacy further.

The Gini coefficient is a measure of inequality [31]. It has originally been used to measure the equality of income distribution among the population. In the context of genomics, it could measure the equality of SNP value distribution in the genome. It is difficult to see how this connects to privacy, and in fact our experiments (see Figure 2o) show that the value of the Gini coefficient does not respond to a change in adversary strength as we would expect.

B. Included Metrics

The amount of leaked information [7], [9] counts the number of leaked SNPs. A SNP is considered leaked when the adversary’s estimate for the true outcome is above a threshold.

An opposite to the amount of leaked information is user-specified innocence which counts the number of SNPs that remain private [27]. A SNP is considered private if the adversary’s estimate for the true outcome is below a threshold.

A threshold of 1 resp. 0 means that the adversary’s estimate has to be absolutely certain. The threshold can be adapted to a specific scenario, for example, the value of a SNP may be considered leaked even if the adversary’s probability is lower than 100%, or private if the adversary’s probability is higher than 0%.

The expected estimation error quantifies the adversary’s correctness by computing the expected distance between the adversary’s estimate and the true value for every SNP [2]. In the context of genomics, this distance is computed on the encoded SNP values. We therefore have to ensure that the genomics interpretation for the distance between two encoded values is meaningful. For example, the encoding proposed in [2] is meaningful, because the encoded value 1 (one each of major and minor allele) clearly lies between 0 (two major alleles) and 2 (two minor alleles). It is important to note that this metric may behave differently if a different encoding is used.

The same statements apply to the mean squared error which is computed as the squared difference between the true value and the adversary’s estimate, averaged over all SNPs [24].

Entropy quantifies the amount of information contained in a random variable. Used as a privacy metric, it indicates the adversary’s uncertainty [20]. Entropy can be normalized to a range of $[0, 1]$ by dividing it by Hartley entropy, i.e. the logarithm of the number of outcomes [2].

Hartley entropy, or max-entropy, has also been used as a privacy metric [23]. It is an optimistic metric because it only accounts for the number of outcomes, but not for additional information the adversary may have. In the context of genomics, however, the number of outcomes per SNP is known to be 3, and therefore max-entropy is not useful.

In contrast, min-entropy is a pessimistic metric because it is based only on the probability of the most likely outcome,
**TABLE I. PRIVACY METRICS**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Definition</th>
<th>Genomics Precedent</th>
<th>Inputs</th>
<th>H/L</th>
<th>Privacy Level</th>
<th>Intuitiveness</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adversary’s success rate</strong></td>
<td>( \sum_{i \in SNPs} p(x_i = y_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Amount of information leaked</strong></td>
<td>(</td>
<td>y_i</td>
<td>) so that ( \forall x_i \in SNPs: p(y_i = y_i) &gt; \alpha )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Asymmetric entropy</strong></td>
<td>(</td>
<td>p(x_i = y_i)(1-p(x_i = y_i))</td>
<td>) where ( w_i = (1-r_i)^2 ) if ( y_i = 0 ), ( w_i = 2r_i(1-r_i) ) if ( y_i = 1 ), ( w_i = r_i^2 ) if ( y_i = 2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Coefficient of determination</strong></td>
<td>( r^2 = \frac{1 - \sum_{x \in Y} (p(x) - \hat{p}(x))^2}{\sum_{x \in Y} (p(x) - \overline{p}(x))^2} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conditional entropy</strong></td>
<td>( H(X_i</td>
<td>Y_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conditional privacy loss</strong></td>
<td>( 1 - 2H(X_i</td>
<td>Y_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cumulative entropy</strong></td>
<td>( \sum_{i \in SNPs} H(X_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Expected estimation error</strong></td>
<td>( \sum_{i \in (0, 1, 2)} p(x_i = k) ) (</td>
<td>k - y_i</td>
<td></td>
<td>)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Entropy</strong></td>
<td>( H(X_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Genomic privacy</strong></td>
<td>( - \sum_{i \in SNPs} \sum_{y_i \in S} \log_2(p(x_i = 1) + p(x_i = 2)) ) (</td>
<td>W_i</td>
<td>)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Health privacy</strong></td>
<td>( \sum_{i \in SNPs} \sum_{y_i \in S} c_i G_i ) where ( G_i ) is a per-SNP base metric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Information surprisal</strong></td>
<td>( - \log_2 p(x_i = y_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inherent privacy</strong></td>
<td>( 2^{H(X_i)} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Max-entropy</strong></td>
<td>( H_0(X_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean squared error</strong></td>
<td>( \sum_{i \in SNPs}</td>
<td></td>
<td>x_i - \hat{x}_i</td>
<td></td>
<td>^2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Min-entropy</strong></td>
<td>( - \log_2 \text{max}_i {p(x_i = k)} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mutual information</strong></td>
<td>( I(X_i; Y_i) = H(X_i) - H(X_i</td>
<td>Y_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Normal entropy</strong></td>
<td>( H_0(X_i</td>
<td>Y_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Normalized mutual inf.</strong></td>
<td>( \frac{H_0(X_i</td>
<td>Y_i)}{H(X_i)} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Relative entropy</strong></td>
<td>( \sum_{k \in (0, 1, 2)} p(y_i = k) \log_2 \frac{p(x_i = k)}{p(x_i = k)} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>User-specified innocence</strong></td>
<td>( \sum_{i \in SNPs} p(y_i = y_i) ) (</td>
<td>k - y_i</td>
<td></td>
<td>)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variation of information</strong></td>
<td>( H(X_i) + H(Y_i) - 2H(X_i; Y_i) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \alpha \) high (H) or low (L) values indicate high privacy

\( \alpha \) Provided a good (+++) base metric is used

**TABLE II. NOTATION**

- \( k \in \{0, 1, 2\} \) Possible SNP values
- \( x_i \) Estimated value of SNP \( i \)
- \( y_i \) True value of SNP \( i \)
- \( p(x_i = y_i) \) Probability to guess true value of SNP \( i \) correctly
- \( p(x_i = k) \) Adversary’s estimate for the case that SNP \( i \) has value \( k \)
- \( r_i \) Minor allele frequency of SNP \( i \)
- \( \alpha \) Threshold for adversary’s probabilities
- \( c_i \) per-SNP weights for Health privacy metric
- \( W_i \) per-SNP severities for Genomic privacy metric

Regardless of whether this is also the true entropy [23]. Min-entropy is a conservative measure of how certain the adversary’s estimate is of his estimate.

**Information surprisal**, or self-information, quantifies how much information is contained in a specific outcome of a random variable [21]. In the context of genomics, the outcome is the true value of a SNP, and the information content is measured using the probability the adversary assigns to this outcome. Informally, information surprisal quantifies how surprised the adversary would be upon learning the true value of a SNP.

**Cumulative entropy** is based on the notion that the adversary’s uncertainty increases when privacy protection is applied at several independent points; with entropy as a measure for uncertainty, cumulative entropy is computed as the sum of individual entropies [19]. In the context of genomics, we can sum over all per-SNP entropies.

The relative entropy, or Kullback-Leibler divergence, between two random variables \( P \) and \( Q \) measures the information that is lost when \( Q \) is used to approximate \( P \) [26]. In the context of genomics, good choices for \( P \) and \( Q \) are the true value and the adversary’s estimate, respectively. This measures how many additional bits of information the adversary needs to reconstruct the true value.

**Conditional entropy**, or the entropy of \( X \) conditioned on \( Y \), measures the amount of information needed to fully describe \( X \), provided that \( Y \) is known [17]. As with relative entropy, \( X \) can be chosen as the true value and \( Y \) as the adversary’s estimate. This measures how much more information the adversary needs to find the true value.

**Mutual information** measures how much information is shared between two random variables [25] which can again be chosen as the true value and the adversary’s estimate. We can normalize mutual information using either Shannon entropy [32] or Hartley entropy [2].

**Variation of information** is derived from mutual information so that it fulfills the conditions for a distance metric in the mathematical sense, especially the triangle inequality [28]. It describes the distance between two random variables which can be chosen as above.

**Inherent privacy** [18], [22], **conditional privacy** [18], and **conditional privacy loss** [18] are derivations of base metrics (entropy, conditional entropy, and mutual information, respectively), each computed as \(^2\text{base metric}\). While the base metrics are interpreted as bits of information, these metrics can be interpreted as the number of binary questions an adversary has to ask to resolve his uncertainty.

**Asymmetric entropy** is another measure for the adversary’s uncertainty. It is tailored to genomics because it assumes that the adversary’s estimate is based on population-wide minor allele frequencies. The metric normalizes using the different maximum values for the per-SNP entropies that are caused by this [10].

The adversary’s success rate captures how likely it is for the adversary to succeed. In the context of genomics, we can
define success on a per-SNP basis as the probability of correctly inferring a SNP value, and aggregate to a per-individual metric by computing the average probability for all SNPs [3].

The coefficient of determination \( r^2 \) describes how well a statistical model approximates data. It is typically used for linear regression where a value of 1 indicates a perfect fit [16]. In the context of genomics, the adversary’s estimate can be used as statistical model, and the true SNP values represent the data.

Genomic privacy weights each SNP with a severity which is thought to be derived from scientific studies or tables provided by insurance companies. The adversary’s weighted estimates for those cases when the SNP is present as a variation, i.e. with its minor allele, are then summed up to a per-individual metric [3]. Its value depends strongly on the number of SNPs studied for each individual, and could thus be normalized to make it more comparable.

Health privacy focuses on those SNPs known to contribute to a specific disease; it can be computed separately for each disease of interest. Health privacy uses a base metric to compute per-SNP values, and then aggregates to a per-individual metric using a weighted and normalized sum [2]. Base metrics discussed in [2] are the expected estimation error, normalized entropy, and normalized mutual information. We extend this list and also investigate relative entropy, information surprisal, and min-entropy as base metrics.

IV. ADVERSARY ESTIMATES

We use two different types of adversary estimate to evaluate each metric. The first type is the reference estimate which assumes that the adversary chooses his estimate according to population-wide minor allele frequencies, for example taken from the dbSNP [33]. The two outcomes that contain the minor allele are assigned the minor allele frequency as the adversary’s probability, while the third outcome is assigned the remainder to ensure the probability sums up to 1.

The second type is a range of estimates that represent different levels of adversary strength, i.e. how close the adversary’s probabilities come to the target’s true SNP values. We compute estimates that are close to the true value as follows: A probability between a given threshold and 1 is sampled from the uniform distribution and assigned to the true outcome. The two smaller probabilities are also sampled from the uniform distribution so that the three probabilities add up to 1. We use thresholds of 0.2, 0.6, and 0.9 for this estimate. In a similar manner, we compute estimates that are far from the true value. We sample a probability between 0 and the threshold and assign it to the true outcome. The two remaining probabilities are sampled from the uniform distribution so that the three probabilities add up to 1. For this estimate, we use thresholds of 0.1, 0.4, and 0.8.

Figure 1 shows the probability that the adversary assigns to the true SNP value, depending on his estimate. In our evaluation, we use six levels of adversary strength, ranging from very weak to very strong, as well as the reference estimate.

Intuitively, we expect that privacy is higher if the adversary’s guesses are far from the true value, and lower if his guesses are close. For the reference estimate, we expect that the adversary’s guesses are close to the true value in many cases, because the estimates are chosen to match the majority of the population. It is important to note that an adversary using the reference estimate is very realistic since minor allele frequencies are easy to obtain. It is therefore important to find protection mechanisms that are effective against this kind of adversary.

V. EVALUATION

Our evaluation consists of three main parts. First, we evaluate how the 22 metrics behave using fixed parameter values for the parametrized metrics. Second, we evaluate the influence of different parameter settings on these metrics. Third, we evaluate privacy with respect to late-onset Alzheimer’s disease by restricting the set of SNPs to two SNPs that are known to significantly increase the risk.

Data. We used two publicly available data sources for our evaluation. First, we downloaded genomic data from 1193 individuals from openSNP\(^5\). This dataset consists of genomic data that users acquired from 23andme\(^6\) and FamilyTreeDNA\(^7\) and published on openSNP. On average, each user has data about 730k SNPs. We used this data as ground truth information for all metrics that rely on it (see Table I, column Inputs).

Second, we downloaded minor allele frequencies from the Database of Single Nucleotide Polymorphisms (dbSNP) [33]. The minor allele frequencies in this dataset are computed from a sample global population consisting of 1000 genomes. We used minor allele frequencies to construct the reference adversary estimate, and for the computation of asymmetric entropy.

We implemented our computations in Python, using SciPy\(^8\) for the entropy-based metrics, and scikit-learn\(^9\) for metrics based on mutual information.

Plots. The plots show one vertical violin plot [34] for each adversary type. The vertical bar shows the range of the data, with horizontal lines indicating the minimum, mean, and maximum. In addition, a kernel density plot on each side of the bar indicates the probability density. Violins marked 1 to 6 represent adversaries whose estimates range from very far to very close to the true value. The right-most violin represents the reference adversary who uses minor allele frequencies as reference information.

To illustrate the statistical distribution of the bulk of metric values, we added the median as well as the first and third

\(^5\) https://opensnp.org/

\(^6\) https://www.23andme.com/

\(^7\) https://www.familytreedna.com/

\(^8\) http://scipy.org/

\(^9\) http://scikit-learn.org
quartile to the plot. To emphasize how this distribution changes with adversary strength, we fitted cubic splines to the three series of medians and quartiles. We plot the median spline as a black line, and shade the area between the quartile splines. In addition, we print the value of the mean in boldface at the top of each violin.

A. Comparison

To get a high-level overview of how the different metrics behave, we computed their values using all genomes in our dataset, but only 500 SNPs each\(^\text{10}\). We computed 25 replications to make sure the results are not due to random variations in the adversary estimate. We also computed confidence intervals for the mean. The relative errors for these confidence intervals were below 0.5% in all cases, indicating that we performed enough replications to achieve highly precise results. We do not plot the confidence intervals since they are so narrow that the lower and upper bounds would collapse to a single line on top of the mean.

We used fixed parameter settings for the metrics (we evaluate the effect of changing parameter settings in Section V-B below). For genomic privacy, we used severities sampled from the uniform distribution between 0 and 1. For health privacy, we used 100 SNPs with equal weights, and the expected estimation error as base metric. We set the threshold for amount of information leaked to 0.7, and for user-specified innocence to 0.3.

Figure 2 shows the results. Each of the vertical violins aggregates the results for all SNPs, all individuals, and all replications we performed for each metric and each adversary strength level.

The most important requirement we look for in a privacy metric is how it represents the privacy level for an individual. As stated before, we expect privacy to be high for a weak adversary, and then to decrease with increasing adversary strength, i.e. from left to right in the plots.

Looking at the entropy-based metrics in the first two rows (Figures 2a–2f), we can see that these metrics do not reflect how close the adversary comes to the true value: they have similar values on both ends of the adversary-strength spectrum, and higher values in the middle. Considering that entropy measures the uncertainty in a random variable, is is clear why this is the case: uncertainty is highest for adversaries in the middle of our spectrum, because in that case probabilities tend towards the uniform distribution, which represents maximum uncertainty. While it is certainly good for privacy if the adversary is uncertain, uncertainty alone is not an accurate representation of a user’s privacy level.

The same is true for conditional entropy and variation of information, shown in Figures 2h and 2i. A similar behavior, albeit reversed and less pronounced, can also be observed for mutual information and conditional privacy loss (observe the horizontal bars for the mean in Figures 2g and 2j).

Relative entropy and information surprisal, shown in Figures 2k and 2l, are the only two information theoretic metrics that behave as we would expect. Their values decrease with increasing adversary strength.

The adversary’s success rate (Figure 2m) is clearly increasing with the adversary’s strength, allowing to distinguish five of the six adversaries. It is not surprising that the two strongest adversaries cannot be distinguished because we count a success if the highest-probability estimate corresponds to the true value, regardless of how high this probability is.

The values of the coefficient of determination are decreasing for most adversary strengths, as shown in Figure 2n. This is contrary to what we expect from the definition of the metric: the lowest privacy level – a perfect fit between the adversary’s estimate and the true outcome – is indicated by higher values of the coefficient of determination. The figure shows the reverse behavior. The metric does therefore not give a correct estimation of a user’s privacy level.

The amount of information leaked (Figure 2p) and user-specified innocence (Figure 2q) show the same situation from two different angles: information that is assumed to be leaked versus information that is assumed to remain private. With the parameter setting we used in this experiment, each metric can only distinguish between five of the six adversaries; the values for the weakest resp. strongest two adversaries are zero. In the other cases, the metrics behave as we expect, with increasing values for information leaked, and decreasing values for user-specified innocence. The value range for these two metrics depends on the number of SNPs in the study; since we investigated 500 SNPs, the maximum value is 500. It would thus be easy to normalize these metrics to a range of \([0, 1]\) by dividing by the number of SNPs. Depending on the study, however, the value as a count of leaked/private SNPs may also be of interest. The amount of information leaked is the only metric that explicitly counts the number of information items (SNPs) not hidden by a PET.

The expected estimation error (Figure 2r) does not show a big difference between adversary strengths, mainly because the range of values is relatively large compared to where the bulk of the values lie. However, it can be seen that the mean is decreasing with increasing adversary strength. This becomes much more evident when the expected estimation error (a per-SNP metric) is aggregated to a per-individual metric, for example when it is used as a base metric for health privacy. In this case, the decrease in value is much more pronounced, as can be seen in Figure 2u (we investigate different base metrics for health privacy in Section V-B).

The values for mean squared error (Figure 2s) are decreasing, but only for the middle four adversaries. The weakest and strongest two adversaries are hard to distinguish, which may be acceptable in some scenarios.

Genomic privacy (Figure 2t) has increasing values for increasing adversary strengths. The values depend on the number of SNPs that are studied, as well as on how the severity for each SNP is chosen. We will discuss options for normalization in the next section.

Regarding the performance of the reference adversary, we can see that in most cases it is situated in the middle of our adversary-strength spectrum. Among the metrics that give a useful indication of the privacy level, notable exceptions are the
Fig. 2. Privacy metrics evaluated according to adversary strength, ordered weakest to strongest from left to right
expected estimation error and health privacy, which place the reference estimate among the weakest adversaries, and genomic privacy, which places it on slightly stronger level.

B. Influence of Parameter Settings

1) Genomic Privacy: Genomic privacy is a per-individual metric that operates on a selection of SNPs that are weighted with their severity. The choices to select SNPs and severities are difficult, because they depend on how much we know about genomics today, which is bound to change and often buried in highly technical genomics research articles. It is therefore tempting to choose SNPs and severities arbitrarily when evaluating a new PET using genomic privacy. However, an arbitrary choice may lead to wrong conclusions about the viability of a PET. More studies may be needed to investigate how the genomic privacy metric behaves when the selected SNPs are risk factors for real diseases and the severities have been determined from the magnitude of influence found in scientific studies.

We investigated the influence of severities on genomic privacy (Figures 3a–3c). In Figure 3a, all the severities were chosen to be 1, whereas Figures 3b and 3c reproduce severity distributions from [3], i.e. truncated Gaussian random variables with \( (\mu_{\text{high}}, \sigma_{\text{high}}) = (0.25, 0.15) \) and \( (\mu_{\text{low}}, \sigma_{\text{low}}) = (0.75, 0.15) \). We can see that the metric’s absolute value changes, but not its behavior.

Since the absolute value of genomic privacy depends on the number of SNPs and the choice of severities, we would ideally want to normalize it. However, normalization is not straightforward, because as the adversary’s probabilities get arbitrarily small, the metric’s values can get arbitrarily large. The original paper [3] normalized with respect to the highest privacy value one of three individuals could achieve. This is not a satisfactory solution because it depends on how the scenario is defined and on the individuals in the study, and does not enable cross-scenario comparisons. We studied two normalization strategies: using the number of SNPs (see Figures 3d and 3e), and using the product of the number of SNPs and the sum of severities (Figure 3f). The first option removes the influence of the number of SNPs studied, but still shows the influence of the chosen severities. The second option removes the influence of the severities as well. This is very similar to the normalization strategy for health privacy.

2) Health Privacy: Health privacy presents a way how a per-SNP metric can be aggregated into a per-individual metric. It relies on three parameters: the selection of SNPs, the weights assigned to each, and the base metric that computes per-SNP values. Since the metric is normalized using the sum of SNP weights, the number of SNPs and the composition of the weights do not influence magnitude of the final value. The value of health privacy therefore depends mostly on the value of the base metric.

Figure 4 shows the value of health privacy for six different base metrics. It can be seen that health privacy behaves very similar to its base metric: entropy, min-entropy and normalized mutual information (Figures 4a – 4c) have their highest values in the middle of our adversary spectrum and thus do not give a good indication of the achieved privacy level. Expected estimation error, relative entropy, and information surprisal (Figures 4d – 4f) have strictly decreasing values for increasing adversary strengths and are thus useful to quantify the privacy level. This means that health privacy is a convenient way of aggregating per-SNP metrics into a single per-individual value, provided that an appropriate base metric is chosen.

3) Amount of Information Leaked and User-Specified Innocence: In our experiments, we found that setting the threshold for leaked information to 1, or close to 1, resulted in either zero or 100% leaked SNPs for most adversary strength levels. The reverse is true for user-specified innocence. This setting therefore doesn’t allow to discriminate between adversaries of different strengths. In our comparative evaluation, we chose thresholds of 0.7 for the amount of information leaked and 0.3 for user-specified innocence, because these values allow to discriminate between five of the six adversary strengths. A higher expressive power can be achieved by combining the two metrics, which will indicate how many SNPs have to be considered leaked; how many will certainly remain private; and as additional information, for how many the leakage status is uncertain.

VI. CASE STUDY: ALZHEIMER’S DISEASE

We applied the privacy metrics to study privacy with respect to late-onset Alzheimer’s disease. The purpose of this case study is twofold: first, we show the considerations necessary prior to applying privacy metrics to a real scenario, and second, we show how the results can be interpreted. This will allow us to draw further conclusions about the usefulness of metrics and metric combinations.

We identified four tasks that are necessary to measure privacy in a real scenario. The first task is the choice of SNPs. There are hundreds of studies correlating SNPs with Alzheimer’s disease risk in the genomics literature. We focused on two SNPs whose association with Alzheimer’s disease has been known for a long time, and which are therefore present in every genome we downloaded from openSNP: rs7412 and rs429358 [35]. New associations with Alzheimer’s are discovered frequently (see references in [35]). However, these have only been sequenced for a subset of the genomes in the openSNP dataset. To avoid distorting the metrics because of missing SNPs, we have conducted our experiments using only the two SNPs mentioned above.

After choosing SNPs, the second task is the selection of privacy metrics. For this case study, we used as many metrics as possible to illustrate their effects. However, we did omit some metrics. Specifically, we removed most entropy-based metrics because the comparison above showed that they all have the same basic shape.

After choosing SNPs and metrics to study them with, the third task is to fix the metric parameters. Two metrics, genomic privacy and health privacy, use weights for individual SNPs. We therefore need to determine how much each SNP contributes to the overall disease risk. This would usually be done using scientific studies or tables released by insurance companies [2], [3]. For the sake of our case study, we chose equal weights and severities for the two SNPs. The amount of information leaked and user-specified innocence use a parameter for the threshold probability. This choice is highly dependent on the privacy preferences of individual users. For our case study, we
assumed that SNPs where the adversary’s estimate of the true value is above 70% are considered leaked, while SNPs with an estimate below 30% are considered private.

After fixing SNPs, metrics, and parameters, the privacy measurement can be conducted. The fourth and final task is then to interpret the results. Figure 5 shows that the results of our case study are similar to the comparative evaluation above.

As before, asymmetric entropy and normalized entropy (Figures 5a and 5b) measure the adversary’s uncertainty, and therefore cannot reliably indicate the user’s privacy level.

Relative entropy (Figure 5c) indicates how much additional information, measured in bits, the adversary needs to reconstruct the true values. This amount of information is very similar to the amount of surprise the adversary will experience upon learning the true value, as indicated by the information surprisal metric (Figure 5d, also measured in bits).

The values of the coefficient of determination (Figure 5e) are markedly different from the comparative evaluation above. However, we still cannot distinguish adversaries of different strengths. This metric is therefore not useful for our case study.

The expected estimation error shown in Figure 5f behaves in a similar way as above. However, it is worth pointing out that the error is higher for reference adversaries than for any other adversary type. This would indicate that the reference adversary performs worse (has a higher error) even than the adversary whose estimate is furthest from the true value. This is hard to imagine, and not backed up by the other metrics.

The distribution of the adversary’s success rate (Figure 5g) has three peaks at 0, 0.5, and 1, and thus clearly shows that the adversary can infer either 0, 1, or 2 SNP values in this scenario. The interpretation is intuitively clear. Similar to the success rate, the distributions of the amount of information leaked and user-specified innocence have three peaks, showing how often 0, 1, or 2 SNP values have leaked or remained private, respectively (Figures 5h and 5i). The figures also show how these frequencies change as the adversary gets stronger.

The mean squared error (Figure 5j) shows how far, on average, the adversary’s guess is from the true value. However, since the error is squared and computed on encoded SNP values, the values need some explanation.

As in the comparative evaluation, the genomic privacy metric behaves well and allows to distinguish all six adversary strengths (Figure 5k). However, is is unclear what the absolute value of genomic privacy means and how it can be interpreted.

Contrary to the comparative evaluation, here we show health privacy using information surprisal as the base metric, which gives a consistent estimation of the privacy level (Figure 5l). Again we see that the values and interpretations for health privacy are the same as for its base metric.
Based on our findings, we rated each metric based on how easy it is to understand what its values mean, and how easily it can be interpreted, both in general (Section III) and in the context of our evaluations (Sections V and VI). We summarize the ratings in Table I (column Intuitiveness).

VII. CONCLUSION

We conducted a comparison of metrics for genomic privacy. Our model assumes that the adversary aims to infer SNP values for a target individual, and that he uses an inference attack to construct probability distributions on SNP values. Based on this model, we evaluated 22 published privacy metrics for different adversary types and strengths. We also evaluated the metrics in a case study of an individual’s risk for late-onset Alzheimer’s disease.

Based on our findings, we rated the metrics according to two criteria: how well they are able to estimate a user’s privacy level, and how easy they are to interpret. We found that only six out of 22 metrics scored highly on both criteria.

In addition, we identified several metrics that reveal additional information when used in combination. Thus, we suggest to combine metrics that measure very different quantities, for example the adversary’s success rate (probability), amount of leaked information and user-specified innocence (counts), and relative entropy (additional bits of information needed).

Further research into genomic privacy metrics is necessary in two areas. The first area concerns the adversary’s difficulty in performing an attack, and especially the resources needed for it (e.g., in terms of time, computational power, or money). There are not many metrics addressing this issue in general (e.g., [15]), and none of them have been applied to genomic privacy. The second area is the quantification of kin privacy. Information about one individual’s genome always reveals information about their kin. However, it is difficult to quantify the privacy loss of an individual when a relative’s genome is revealed, and despite initial attempts [2] there is no satisfactory solution yet.

Our systematic comparison of metrics for genomic privacy will enable researchers to make informed decisions about the selection of privacy metrics, and thus help ensure that new PETs are evaluated in a consistent and comparable manner.

REFERENCES


