

# Evaluating the Strength of Genomic Privacy Metrics

ISABEL WAGNER, De Montfort University

The genome is a unique identifier for human individuals. The genome also contains highly sensitive information, creating a high potential for misuse of genomic data (for example, genetic discrimination). In this paper, I investigated how genomic privacy can be measured in scenarios where an adversary aims to infer a person's genomic markers by constructing probability distributions on the values of genetic variations. I measured the strength of privacy metrics by requiring that metrics are monotonic with increasing adversary strength and uncovered serious problems with several existing metrics currently used to measure genomic privacy. I provide suggestions on metric selection, interpretation, and visualization, and illustrate the work flow using a case study on Alzheimer's disease.

## 1. INTRODUCTION

In 2001, Celera, Inc was the first to sequence a full human genome at a cost of about 300 million dollars. At the time of this writing, full genome sequences can be obtained at a cost of little more than \$1,000 per genome [Wetterstrand 2016]. This has enabled a dramatic increase in the use of genomic data 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). Unfortunately, the wide availability of genomic data also raises important 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 (revealing a person's disease status) to genetic discrimination (for example, denial of insurance because of 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 these privacy concerns, currently, there is a lack of methods to measure how private a particular genomic technology is (i.e. genomic privacy metrics). As a result, technologies that preserve genomic privacy are still in their infancy.

In this paper, we investigate the strength of genomic privacy metrics. We consider an adversary who targets an individual and aims to infer as much of the target's genome sequence as possible. We assume that the adversary uses an inference attack to compute a probability distribution for each variation in the target genome. This is a reasonable assumption, because several inference attacks have already been described, for example exploiting linkage disequilibrium [Ayday et al. 2013], exploiting information from kin genomes [Humbert et al. 2013; Humbert et al. 2014], exploiting systematic execution of genomic tests [Goodrich 2009], and using statistics to infer whether an individual participated in a genome-wide association study [Homer et al. 2008; Wang et al. 2009].

**Contributions.** We measured the strength of 23 genomic privacy metrics in three scenarios, for adversaries of different strengths. The key indicator of a metric's strength was defined as monotonicity, i.e. that metrics should show decreasing privacy for increasing adversary strength. Adversary strength, in turn, was measured by how close their inferences of genomic variants were to the true value. We tested each of the 23 metrics in three possible attack scenarios: (1) a comparative evaluation with a large number of individuals; (2) an evaluation of kin privacy considering only related individuals, and (3) an evaluation focusing on risk factors for Alzheimer's disease.

Of the metrics we tested, we found that only 7 out of 23 metrics were strong across adversary types and scenarios, 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, percentage incorrectly classified,

relative entropy, and user-specified innocence. Furthermore, we find that none of the metrics we tested are sufficiently reliable when used by themselves. Therefore, we recommend to combine multiple strong metrics to gain insight on as many different aspects of privacy as possible.

Our systematic comparison of genomic privacy metrics enables researchers, clinicians, and policy-makers to make an informed choice about the selection of privacy metrics and privacy-enhancing technologies. In addition, two new visualization methods that we introduced, namely heat maps and radar plots, will further help to ensure that new privacy enhancing technologies are evaluated in a consistent and comparable manner.

## 2. BACKGROUND

### 2.1. Genomics

Although the human genome consists of about three billion DNA base pairs, genomes from two human individuals differ only in about 0.2–0.4% of base pairs [Tishkoff and Kidd 2004]. Most commonly, this genetic variation comes from differences in single bases, called single nucleotide polymorphisms (SNPs, pronounced *snips*) [Sachidanandam et al. 2001]. In most cases, a SNP has only two variants (alleles) in the human population. Usually, of the two SNP alleles one is more common than the other (called the major allele,  $A$ , and the minor allele,  $a$ , respectively). Because the genome of a somatic human cell is diploid, that is, it is comprised of two sets of chromosomes – one set inherited from the father, and the other set inherited from the mother – each SNP is present in two copies. Therefore, a given SNP can be encoded as 0, 1, or 2 corresponding to the combinations  $AA$ ,  $Aa$ , and  $aa$  [Humbert et al. 2013]. Population-wide frequencies of alleles  $A$  and  $a$  can be estimated from a sample of human genomes; this has been done in the 1000 Genomes project<sup>1</sup>. Genome-wide association studies can identify SNPs associated with diseases by comparing the incidence of SNP variations between individuals who have and do not have a particular disease.

### 2.2. Privacy Metrics

Many privacy metrics have been proposed for different domains [Wagner and Eckhoff 2015]. However, many studies have shown their shortcomings, for example inconsistent metrics [Wagner 2015], metrics that are hard to understand [Diaz et al. 2007], and metrics that work only in narrow scenarios [Kalogridis et al. 2010]. This is problematic, because use of a weak privacy metric can lead to an overestimation of privacy and result in privacy violations, for example the re-identification of individuals in published health data, thus linking individuals to their medical conditions [Sweeney 2002]. Privacy metrics that suffer none of these shortcomings can still be weak if used on their own because some metrics are complementary – they measure different aspects of privacy and thus need to be used in combination to form a more complete measurement of privacy [Murdoch 2014; Liu and Mittal 2016]. These shortcomings show that existing privacy metrics exhibit a lack of consistency, reproducibility, and wider applicability. However, it is unknown which privacy metrics, and in which application domains, produce consistently good measurements of privacy. This can not only impede and slow down privacy research [He et al. 2015; Shokri et al. 2011; Murdoch 2014], but also lead to real-world privacy violations [Sweeney 2002], and has recently led to calls for research on privacy metrics [He et al. 2015; Shokri et al. 2011; Murdoch 2014].

### 2.3. Genomic privacy metrics

Broadly, privacy metrics measure characteristics of privacy enhancing technologies and quantify how much privacy a technology offers [Clauß and Schiffner 2006], for example, the adversary’s probability to break a user’s anonymity [Serjantov and Danezis 2002], or

<sup>1</sup><http://www.1000genomes.org/>

the maximum amount of bits of private information an adversary can infer [Diaz et al. 2003]. In the context of genomic privacy, most research applies existing privacy metrics to genomic privacy scenarios [Ayday et al. 2014; Humbert et al. 2013; Ayday et al. 2013; Samani et al. 2015]. Some researchers also propose new metrics specific to genomic privacy [Ayday et al. 2013; Ayday et al. 2013; Humbert et al. 2013]. These papers generally propose or describe one or more metrics, and then use these metrics to evaluate a privacy enhancing technology in a given scenario. However, they do not evaluate the strengths of the metrics, or how they differ from other metrics. This paper aims to address this gap. The closest to our work is [Murdoch 2014], which investigates the behavior of anonymity metrics, among them entropy and some of its variations. In previous work, we have published an initial evaluation of metrics for genomic privacy [Wagner 2015].

#### 2.4. Requirements for genomic privacy metrics

Traditionally, a strong privacy metric is one that can (1) indicate, in terms understandable to lay people, how effectively the adversary can succeed [Alexander and Smith 2003]; (2) show both the privacy level and the portion of data not protected [Bertino et al. 2008]; (3) consider accuracy, uncertainty, and correctness as three aspects of the adversary’s success [Shokri et al. 2011]; and (4) indicate not only the difficulty for the adversary, but also the amount of resources he needs to succeed [Syverson 2013]. Most of these criteria apply to specific privacy metrics, but cannot be used to compare the strengths of different metrics.

In this work, we introduce a new criterion for strong privacy metrics – monotonicity, which requires privacy metrics to show decreasing privacy for increasing adversary strength (Section 5). Because monotonicity can be quantified, we believe that it can be used to compare the strengths of privacy metrics. Furthermore, we rate understandability based on the results of our case study (Section 6).

### 3. PRIVACY METRICS

From our previous survey of privacy metrics [Wagner and Eckhoff 2015], we selected 23 metrics that were applicable to our genomic privacy scenario. The metrics are summarized in Table II, and Table I provides a reference for notation used. Nine metrics have previously been applied in genomic privacy; the remaining metrics have been drawn from the wider privacy literature (see the *Genomics Precedent* column in Table II). 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).

#### 3.1. Excluded Metrics

We excluded a range of privacy metrics that did not fit our assumptions.

Differential privacy [Dwork 2006] 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 [Malin 2005] 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 *genomic privacy* metric introduced by Ayday et al. [2013] assumes that the adversary only aims to infer whether an individual’s genome has a specific SNP or not. In contrast, in this paper we assume that the adversary aims to infer the values of SNPs, and we study only SNPs that individuals do have.

Table I. 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

### 3.2. Included Metrics

We group our description of included metrics by the output they measure, according to the taxonomy proposed in [Wagner and Eckhoff 2015].

Table II. Privacy Metrics

Metric	per SNP	Genomics Prece- dent	Inputs	H/L <sup>2</sup>	Priv. Level <sup>4</sup>	Intuitive- ness
Adversary's success rate	–	✓	estimate, truth	L	++	++
Amount of information leaked	–	✓	estimate, truth, $\alpha$	L	++	++
Asymmetric entropy	–	✓	estimate, truth, prior	H	–	–
Asymmetric entropy (per SNP)	–	✓	estimate, truth, prior	H	o	–
Coefficient of determination $r^2$	–	–	estimate, truth	L	–	o
Conditional entropy	✓	–	estimate, truth	H	o	–
Conditional privacy loss	✓	–	estimate, truth	L	o	–
Cumulative entropy	–	–	estimate	H	o	+
Entropy $H(X_i)$	✓	–	estimate	H	o	+
Expected estimation error	✓	✓	estimate, truth	H	+	o
Health privacy	–	✓	base metric, $c_i$	H/L	+ / ++ 3	+ <sup>3</sup>
Information surprisal	✓	–	estimate, truth	H	+	+
Inherent privacy	✓	–	estimate	H	o	–
Max-entropy $H_0(X_i)$	–	–	estimate	H	–	–
Mean error	–	✓	estimate, truth	H	++	o
Mean squared error	–	–	estimate, truth	H	++	o
Min-entropy $H_\infty(X_i)$	✓	–	estimate	H	–	o
Mutual information	✓	–	estimate, truth	L	o	o
Normalized entropy	✓	✓	estimate	H	o	+
Normalized mutual inf.	✓	✓	estimate, truth	H	o	o
Perc. incorrectly classified	–	–	estimate, truth	H	++	++
Relative entropy	✓	–	estimate, truth	H	+	+
User-specified innocence	–	–	estimate, truth, $\alpha$	H	++	++
Variation of information	✓	–	estimate, truth	L	–	o

<sup>2</sup>high (H) or low (L) values indicate high privacy

<sup>3</sup>Provided a good/very good (+/++) base metric is used

<sup>4</sup>Privacy level  $\leq 30$ : –;  $\in [30, 70[$ : o;  $\in [70, 90]$ : +;  $> 90$ : ++

**3.2.1. Metrics Measuring the Adversary's Error.** 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 [Humbert et al. 2013]. In the context of genomics, this distance is computed on the encoded SNP values. Therefore, we have to ensure that the SNP encoding has a meaningful genomics interpretation. For example, the encoding proposed by Humbert et al. [2013] is meaningful, because the encoded value 1 (one each of major and minor allele) lies between 0 (two major alleles) and 2 (two minor alleles). This

metric may behave differently with a different encoding.

$$priv_{EEE} = \sum_{k \in \{0,1,2\}} p(x_i = k) \|k - y_i\|$$

The *mean squared error* is computed as the squared difference between the true value and the adversary's estimate, averaged over all SNPs [Oya et al. 2014].

$$priv_{MSE} = \frac{1}{|\text{SNPs}|} \sum_{x_i \in \text{SNPs}} \{\|x_i - y_i\|^2\}$$

Other variations of the adversary's error are the *mean error* [Samani et al. 2015] and the *mean error with normalized distance* [Humbert et al. 2015].

*Percentage incorrectly classified* measures how often the highest probability in the adversary's estimate does not correspond to true SNP value [Narayanan and Shmatikov 2009].

$$priv_{PIC} = \frac{|\text{incorrect SNPs}|}{|\text{SNPs}|}$$

**3.2.2. Metrics Measuring the Adversary's Uncertainty.** *Entropy* quantifies the amount of information contained in a random variable. Used as a privacy metric, it indicates the adversary's uncertainty [Serjantov and Danezis 2002].

$$priv_{ENT} = H(X_i) = - \sum_{k \in \{0,1,2\}} p(x_i = k) \log_2 p(x_i = k)$$

Entropy can be normalized to a range of  $[0, 1]$  by dividing it by Hartley entropy, that is, the logarithm of the number of outcomes [Humbert et al. 2013].

$$priv_{NE} = \frac{H(X_i)}{H_0(X_i)}$$

*Hartley entropy*, or max-entropy, has also been used as a privacy metric [Clauß and Schiffner 2006]. 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 and has been excluded from the evaluation.

$$priv_{MXE} = H_0(X_i) = \log_2 |x_i| = \log_2 3$$

*Min-entropy* is a pessimistic metric because it is based only on the probability of the most likely outcome, regardless of whether this is also the true outcome [Clauß and Schiffner 2006]. Min-entropy is a conservative measure of how certain the adversary is of his estimate.

$$priv_{MNE} = H_\infty(X_i) = -\log_2 \max p(x_i)$$

*Cumulative entropy* is based on the notion that the adversary's uncertainty increases when privacy protection is applied at several independent points. Cumulative entropy is computed as the sum of individual entropies [Freudiger et al. 2007]. In the context of genomics, we sum over the entropies computed for each SNP.

$$priv_{CE} = \sum_{i=1}^{|\text{SNPs}|} H(X_i)$$

*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 [Diaz et al. 2007]. For

genomic privacy,  $X$  can be chosen as the true SNP value and  $Y$  as the adversary's estimate. This measures how much more information the adversary needs to find the true value.

$$priv_{COE} = H(X_i|Y_i) = H(X_i) - I(X_i; Y_i)$$

*Inherent privacy* [Agrawal and Aggarwal 2001; Andersson and Lundin 2008] and *conditional privacy* [Andersson and Lundin 2008] are derivations of base metrics (entropy and conditional entropy, 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.

$$priv_{IP} = 2^{H(X_i)}, \quad priv_{CP} = 2^{H(X_i|Y_i)}$$

*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, which results in a different maximum value for entropy for each SNP [Ayday et al. 2013].

$$priv_{AE} = \sum_{i=1}^{|\text{SNPs}|} \frac{p(x_i = y_i)(1 - p(x_i = y_i))}{(-2w_i + 1)p(x_i = y_i) + w_i^2}, \text{ where } w_i = \begin{cases} (1 - r_i)^2 & \text{if } y_i = 0 \\ 2r_i(1 - r_i) & \text{if } y_i = 1 \\ r_i^2 & \text{if } y_i = 2 \end{cases}$$

Asymmetric entropy can also be used as a per-SNP metric to measure privacy for individual SNPs.

**3.2.3. Metrics Measuring Information Gain/Loss.** The *amount of leaked information* [Wang et al. 2009; Ayday et al. 2014] counts the number of leaked SNPs. A SNP is considered leaked when the adversary's estimate for the true outcome is above the threshold  $\alpha$ . A threshold of 1 means that a SNP is considered leaked only if the adversary is absolutely certain. Many scenarios will adopt a more conservative threshold to cover situations when the adversary is reasonably, but not absolutely, certain.

$$priv_{ALI} = |u| \text{ so that } \forall u_i \in \text{SNPs} : p(u_i = y_i) > \alpha$$

*Information surprisal*, or self-information, quantifies how much information is contained in a specific outcome of a random variable [Chen et al. 2013]. In the context of genomics, the outcome is the true value of a SNP, and the information content is 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.

$$priv_{IS} = -\log_2 p(x_i = y_i)$$

*Mutual information* measures how much information is shared between two random variables  $X$  and  $Y$  [Lin et al. 2002]. As before,  $X$  can be chosen as the true SNP value and  $Y$  as the adversary's estimate.

$$priv_{MI} = I(X_i; Y_i) = H(X_i) - H(X_i|Y_i)$$

Normalized mutual information can use either Shannon entropy [Zhu and Bettati 2005] or Hartley entropy [Humbert et al. 2013]. In this paper we use the latter.

$$priv_{NMI} = 1 - \frac{I(X_i; Y_i)}{H_0(X_i)}$$

*Conditional privacy loss* [Andersson and Lundin 2008] is derived from mutual information. While mutual information is interpreted as the bits of information shared between the true value and the adversary's estimate, conditional privacy loss can be interpreted as the number

of binary questions an adversary has to ask to arrive at the true value.

$$priv_{CPL} = 1 - 2^{-I(X_i; Y_i)}$$

The *relative entropy*, or Kullback-Leibler divergence, between two random variables  $Y$  and  $X$  measures the information that is lost when  $X$  is used to approximate  $Y$  [Deng et al. 2007]. In the context of genomics, good choices for  $Y$  and  $X$  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.

$$priv_{RE} = \sum_{k \in \{0,1,2\}} p(y_i = k) \log_2 \frac{p(y_i = k)}{p(x_i = k)}$$

*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 [Meilă 2007]. It describes the distance between two random variables, chosen as the true value and the adversary's estimate.

$$priv_{VI} = H(X_i) + H(Y_i) - 2I(X_i; Y_i)$$

**3.2.4. Metrics Measuring the Adversary's Success Probability.** 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 [Ayday et al. 2013].

$$priv_{ASR} = \frac{1}{|\text{SNPs}|} \sum_{i \in \text{SNPs}} p(x_i = y_i)$$

*User-specified innocence* can be seen as a counterpart to the amount of leaked information, because it counts the number of SNPs that remain private [Chen and Pang 2012]. A SNP is considered private if the adversary's estimate for the true outcome is below the threshold  $\alpha$ . A threshold of 0 means that a SNP is considered private only if the adversary considers it impossible. Many scenarios will therefore adopt a higher threshold.

$$priv_{USI} = |u| \text{ so that } \forall u_i \in \text{SNPs} : p(u_i = y_i) \leq \alpha$$

**3.2.5. Metrics Measuring Similarity/Diversity.** 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 [Kalogridis et al. 2010]. In the context of genomics, the adversary's estimate can be used as statistical model, and the true SNP values represent the data.

$$priv_{R2} = 1 - \frac{SS_E}{SS_R + SS_E}, \text{ where } SS_E = \sum_i (y_i - x_i)^2, SS_R = \sum_i (x_i - \bar{Y})^2$$

**3.2.6. Other Metrics.** *Health privacy* focuses on those SNPs known to contribute to a specific disease. 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 [Humbert et al. 2013]. The weights  $c_i$  should be chosen to reflect how much each SNP contributes to the disease. Base metrics discussed in [Humbert et al. 2013] are the expected estimation error, normalized entropy, and normalized mutual information. We extend this list and also investigate relative entropy, conditional entropy, information surprisal, and min-entropy as base metrics.

$$priv_{HP} = \frac{1}{\sum_{i \in S} c_i} \sum_{i \in S} c_i G_i, \text{ where } G_i \text{ is a per-SNP base metric}$$

## 4. INITIAL EVALUATION

### 4.1. Data Sources

We used two publicly available data sources for our initial evaluation. First, we downloaded genomic data from 1857 individuals from openSNP [Greshake et al. 2014]. This dataset consists of genomic data that users acquired from 23andme<sup>5</sup> and FamilyTreeDNA<sup>6</sup> and published on openSNP. On average, each user has data about 730k SNPs. This data serves as ground truth information for all metrics that rely on it (see Table II, column *Inputs*).

Second, we downloaded minor allele frequencies from the Database of Single Nucleotide Polymorphisms (dbSNP) [Sherry et al. 2001]. 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.

### 4.2. Adversary Models

Our adversary models abstract from the strategies and algorithms a real-world adversary would use, and instead represent the strength of an adversary using probability distributions. For our initial evaluation, we use two different types of adversary estimate. The *reference* model uses the population-wide distribution of minor allele frequencies taken from the dbSNP. Following the Hardy-Weinberg principle and denoting the minor allele frequency as  $q$ , the adversary assigns probabilities depending on the number of minor alleles for each SNP: for two minor alleles  $p(aa) = q^2$ , for two major alleles  $p(AA) = (1 - q)^2$ , and for one each of major and minor allele  $p(Aa) = 2q(1 - q)$ .

The *normal* model uses a series of normal distributions with a small standard deviation ( $\sigma = 0.1$ ), truncated to the  $[0, 1]$  range, to represent the probability that the adversary assigns to the true value. We study six strength levels with mean probabilities  $\mu = 0.1, 0.25, 0.4, 0.6, 0.75, 0.9$ . Figure 2b shows the average probability the *normal* adversary assigns to the true SNP value.

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. 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.

### 4.3. Results

To get a high-level overview of how the 23 metrics behave, we computed their values using all genomes in our dataset, but only 10000 SNPs each<sup>7</sup>. We used fixed parameter values for the three metrics that have parameters. (We evaluate the effect of changing parameter settings in Section 5.5 below.) For health privacy, we used 1000 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. We computed 15 replications to make sure the results are not due to random variations in the adversary estimate, and computed confidence intervals for the mean. The relative errors for these confidence intervals were below 5% in all cases, indicating that we performed enough replications to achieve highly

<sup>5</sup><https://www.23andme.com/>

<sup>6</sup><https://www.familytreedna.com/>

<sup>7</sup>We also evaluated the metrics using fewer genomes, but all SNPs for each. The results were very similar, which is why we report our results using the computationally much less demanding scenario with 10000 SNPs per genome.



precise results. We implemented our computations in Python, using SciPy<sup>8</sup> for the entropy-based metrics, and scikit-learn<sup>9</sup> for metrics based on mutual information.

Figure 1 shows the results. For each privacy metric and adversary strength level, we plot one vertical violin plot [Hintze and Nelson 1998]. 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. The six violins on the left represent the *normal* adversaries with estimates ranging from far to close to the true value. The right-most violin represents the *reference* adversary. 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. To illustrate the statistical distribution of the bulk of metric values, we added the median as well as the first and third quartile to the plot. We fitted cubic splines to the medians and quartiles to emphasize how their values change depending on adversary strength. 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. 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.

The most important requirement we look for in a privacy metric is a consistent representation of the privacy level. Privacy should be high for a weak adversary, and decrease with increasing adversary strength, i.e. from left to right in the plots.

**4.3.1. Metrics Measuring the Adversary’s Error.** The expected estimation error (Figure 1a) 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 (Figure 1b). In this case, the decrease in value is much more pronounced (we investigate different base metrics for health privacy in Section 5.5).

The values for mean error, mean squared error, and percentage incorrectly classified (Figures 1c, 1d, and 1e) are decreasing, but only for the weaker four adversaries. The strongest two adversaries cannot be distinguished.

**4.3.2. Metrics Measuring the Adversary’s Uncertainty.** Looking at the entropy-based metrics in Figures 1f–1m, we can see that these metrics peak for medium-strength adversaries and assign similar values to strong and weak adversaries. To explain this, we recall that entropy measures the uncertainty in a random variable. Because of the way we defined the *normal* adversary model, medium-strength adversaries appear more uncertain than adversaries on either end. 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.

**4.3.3. Metrics Measuring Information Gain/Loss.** Mutual information and the metrics derived from it (conditional privacy loss and variation of information) show a similar behavior to the entropy-based metrics, albeit reversed and less pronounced (observe the horizontal bars for the mean in Figures 1n–1q).

Relative entropy and information surprisal, shown in Figures 1r and 1s, are the only two information theoretic metrics that behave as we would expect. Their values decrease with increasing adversary strength.

The amount of information leaked (Figure 1t) and user-specified innocence (Figure 1u) show the same situation from two different angles: information that is considered leaked versus information that remained private. With the parameter setting we used in this ex-

<sup>8</sup><http://www.scipy.org/>

<sup>9</sup><http://scikit-learn.org>

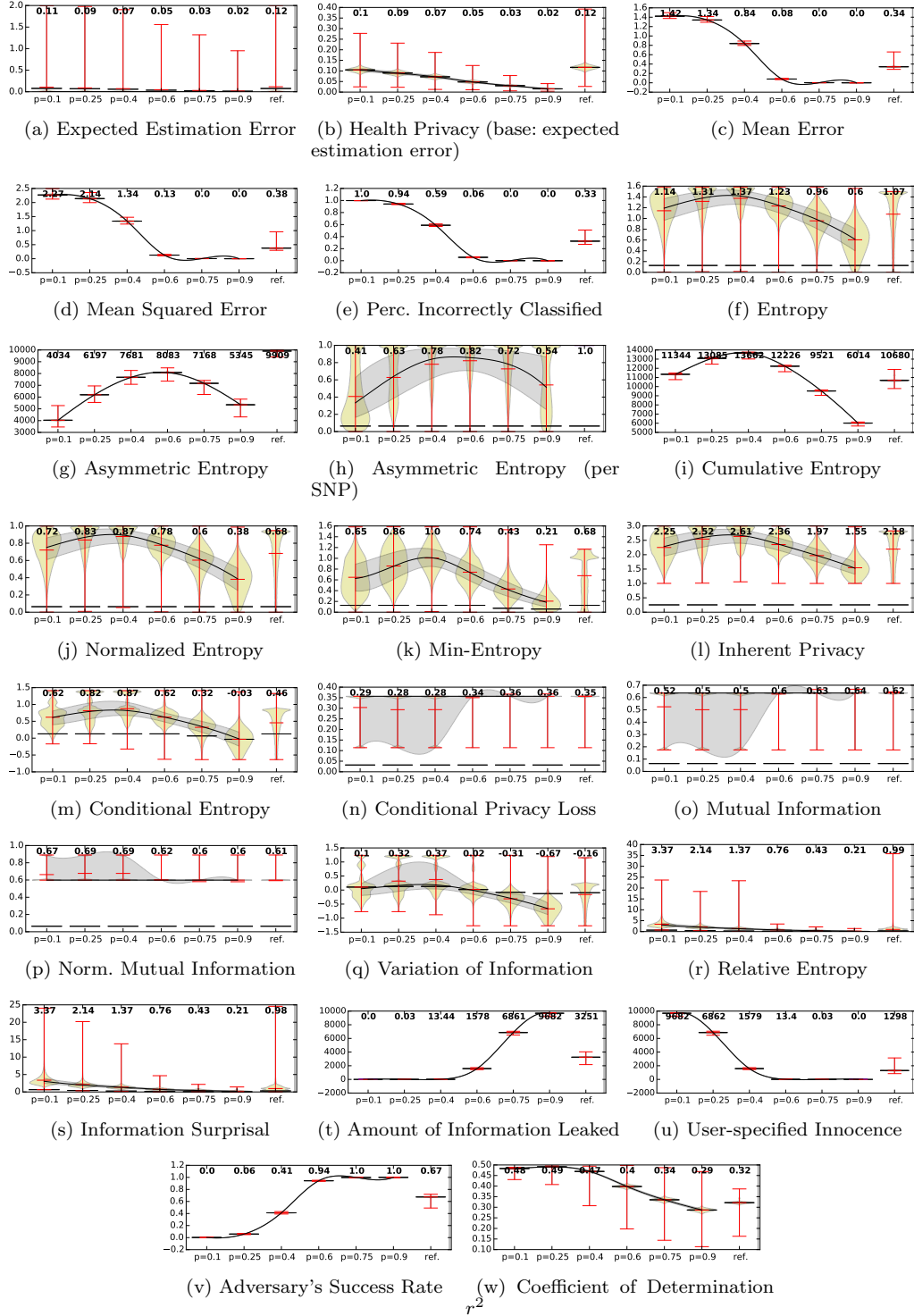


Fig. 1. Privacy metrics evaluated according to adversary strength, ordered weakest to strongest from left to right

periment, 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; the maximum value of 10000 corresponds to the number of SNPs we investigated. It would thus be easy to normalize these metrics to a range of  $[0, 1]$  by dividing by the number of SNPs. The amount of information leaked is the only metric that explicitly counts the number of information items (SNPs) not hidden by a privacy enhancing technology.

**4.3.4. Metrics Measuring the Adversary’s Success Probability.** The adversary’s success rate (Figure 1v) increases with the adversary’s strength, allowing to distinguish five of the six adversaries. The two strongest adversaries cannot be distinguished because we count a success if the estimate with the highest probability corresponds to the true value, regardless of how high this probability is. Because we defined the adversary’s success as the exact opposite of incorrect classification, the percentage incorrectly classified (Figure 1e) is a mirror image of the adversary’s success rate and conveys exactly the same information.

**4.3.5. Metrics Measuring Similarity/Diversity.** The values of the coefficient of determination are decreasing for most adversary strengths, as shown in Figure 1w. However, we expect otherwise: the lowest privacy level – a perfect fit between the adversary’s estimate and the true outcome – should be indicated by higher values of the coefficient of determination. Figure 1w shows the reverse behavior. The coefficient of determination does therefore not give a correct estimation of a user’s privacy level.

Regarding the performance of the reference adversary, we can see that most metrics place it in the middle of our adversary-strength spectrum. Notable exceptions are the expected estimation error and health privacy, which place the reference estimate among the weakest adversaries.

## 5. EXTENDED EVALUATION

We then extended our initial evaluation to address a number of open issues: (1) how do the genomic privacy metrics behave for datasets with different characteristics? (2) how do the genomic privacy metrics behave for different adversary models? (3) how can the results be presented in a more compact and user-friendly way? (4) how do the parameter settings influence the metrics’ behaviors?

### 5.1. Definition of Scenarios

For the extended evaluation, we retained the two datasets from the initial evaluation (openSNP and dbSNP). To study how relationships between individuals influence the strength of privacy metrics, we identified 13 pairs of blood relatives in the openSNP data<sup>10</sup> and added a dataset with verified blood relatives, the CEPH/Utah Pedigree 1463 [Drmanac et al. 2010], or *Utah* for brevity, which contains the genomes of 17 family members from three generations.

We define four scenarios based on the openSNP, dbSNP, and Utah data. In the *kin* scenario, we focus on genomes from related individuals, and evaluate all 23 privacy metrics using all SNPs for 13 pairs of related individuals from openSNP data. In the *utah* scenario,

<sup>10</sup>To identify blood relatives in the openSNP dataset, we first identified pairs of genomes that shared more than 80% of SNP values (statistics from [http://www.isogg.org/wiki/Autosomal\\_DNA\\_statistics](http://www.isogg.org/wiki/Autosomal_DNA_statistics)). We then attempted to verify a potential relationship using the openSNP user names and user profiles. For 10 of these genomes, the relationship degree can be found either by references in the username (e.g., “father of”) or by Google hits on ancestry sites, and another 3 are likely matches judging by the username (same infrequent last name), but with unknown relationship.

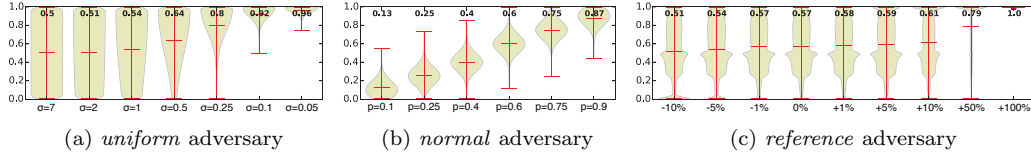


Fig. 2. Average probability the different adversary types assign to the true SNP value

we evaluate all 23 privacy metrics using all SNPs for the 17 related individuals from the Utah data. In the *comparison* scenario, we evaluate all 23 privacy metrics using 10,000 SNPs from all genomes. In the *alzheimer* scenario, the adversary is only interested in an individual’s propensity for Alzheimer’s disease, and therefore we evaluate all privacy metrics on all genomes using three SNPs that are correlated with Alzheimer’s disease.

## 5.2. Adversary Models

For the extended evaluation, we add the *uniform* adversary model and extend the *reference* model. In the *uniform* model, the weakest adversary makes an uninformed guess, represented by a truncated normal distribution that comes close to a uniform distribution. With increasing adversary strength, we skew the distribution towards certainty using increasingly narrow truncated normal distributions (i.e. increasingly smaller  $\sigma$ ). Specifically, we study seven strength levels, setting the mean to  $\mu = 0.99$ , and varying the standard deviation  $\sigma = 7, 2, 1, 0.5, 0.25, 0.1, 0.05$ . Figure 2a shows the average probability the *uniform* adversary assigns to the true SNP value.

Before, we used the *reference* model as an adversary with fixed strength. Now, we vary the adversary strength by assuming that the adversary is uncertain ( $p = 0$ ) or certain ( $p = 1$ ) about some portion of SNPs. Specifically, we study nine strength levels, where the percentages indicate the portion of uncertain resp. certain SNPs:  $-10\%$ ,  $-5\%$ ,  $-1\%$ ,  $0$  (this corresponds to the fixed strength we have used before),  $1\%$ ,  $5\%$ ,  $10\%$ ,  $50\%$ ,  $100\%$ . Figure 2c shows the average probability the *reference* adversary assigns to the true SNP value.

## 5.3. Formalization of the Monotonicity Requirement

The initial evaluation emphasizes that a strong privacy metric should have decreasing privacy levels for increasing adversary strength. In mathematical terms, this means that privacy metrics should be monotonic with increasing adversary strength. Monotonicity is an important requirement, because a nonmonotonic metric can make a privacy-enhancing technology appear stronger than it actually is. If this technology is then used in practice, the use of a weak privacy metric may cause privacy violations.

We formulate an algorithm that evaluates monotonicity based on statistical tests (Algorithm 1). Our algorithm outputs heat maps as a compact and easy-to-understand visualization of the strength of privacy metrics. Because the algorithm relies on statistics, the results are not biased by human judgment. In a monotonic sequence, the differences between successive pairs should all have the same sign. The algorithm therefore awards points for each difference that has the expected sign (positive for metrics where high values indicate high privacy, and negative for metrics where low values indicate high privacy), and penalizes differences that have the wrong sign. Because we have a large number of data points for each adversary strength level, we use statistical tests to evaluate the differences between the means of successive pairs. Welch’s t-test tests the null hypothesis that the metric values for the two adversary strengths have identical means (in contrast to the standard t-test, Welch’s t-test does not assume equal variance in the two samples), and the Wilcoxon rank-sum statistic tests the null hypothesis that the metric values have been drawn from the same distribution. The results of these tests indicate whether the difference between the

**ALGORITHM 1:** Monotonicity Computation for one Privacy Metric**Input:** arrays of metric values for each combination of adversary model and scenario**Output:** heat map visualizing the strength of this privacy metric

tests = [Welch's t-test, Wilcoxon rank-sum statistic]

---

```

foreach combination of adversary model and scenario do
     $m = 0$  ; // holds the monotonicity points value
    foreach test  $\in$  tests do
        prevResult = 0; // holds result for the previous pair
        foreach pair of successive adversary strengths do
            apply test to pair
             $p$  = statistical significance of test
            result = value of test statistic
            if  $p < 0.05$  then // test is statistically significant
                if result  $> 0$  ( $< 0$  for lower-better metrics) then
                     $m = m + 1$ ; // difference in the expected direction
                else if result  $< 0$  ( $> 0$  for lower-better metrics) then
                     $m = m - 1$ ; // difference in the wrong direction
                else
                    ; // result is zero, do nothing
                end
            else // test is not statistically significant
                 $m = m - 0.2$ 
            end
            if result and prevResult have different signs then
                 $m = m - 2$ ; // penalize peaks in the metric value
            end
            prevResult = result; // save result to check next pair for peaks
        end
    end
    end
    normalize  $m$  to  $[-1, 1]$ 
    save  $m$  for plotting
end
plot  $m$  in a heat map (rows = scenarios, columns = adversaries)

```

---

mean metric values is positive, negative, or zero, and whether the difference is statistically significant.

We use heat maps to visualize the resulting point values. Figure 3 shows one heat map for each privacy metric. The rows represent scenarios (kin/utah data, kin/opensnp data, comparison, and alzheimer) and the columns represent adversary models (uniform, normal, and reference). Blue colors indicate a strong metric, green indicates medium strength, and yellow indicates a weak metric. This visualization presents the strengths and weaknesses of a large number of privacy metrics in a compact way and thus helps researchers to select strong metrics. To get a sense of the overall strength of a privacy metric, we aggregate the strengths for all elements in its heat map and show it as a single percentage next to the metric name in Figure 3.

We chose the point values in our algorithm to reflect the monotonicity requirements and to create a high contrast in the visualization, which helps to pinpoint strengths and weaknesses of each metric. We initially assigned points based on the desired behavior of a metric: a change in the right direction (1 point) is better than no change (0 points), which in turn is better than a change in the wrong direction (-1 points). A peak (-2 points) is undesirable because it means that weak adversaries cannot be distinguished from strong adversaries.

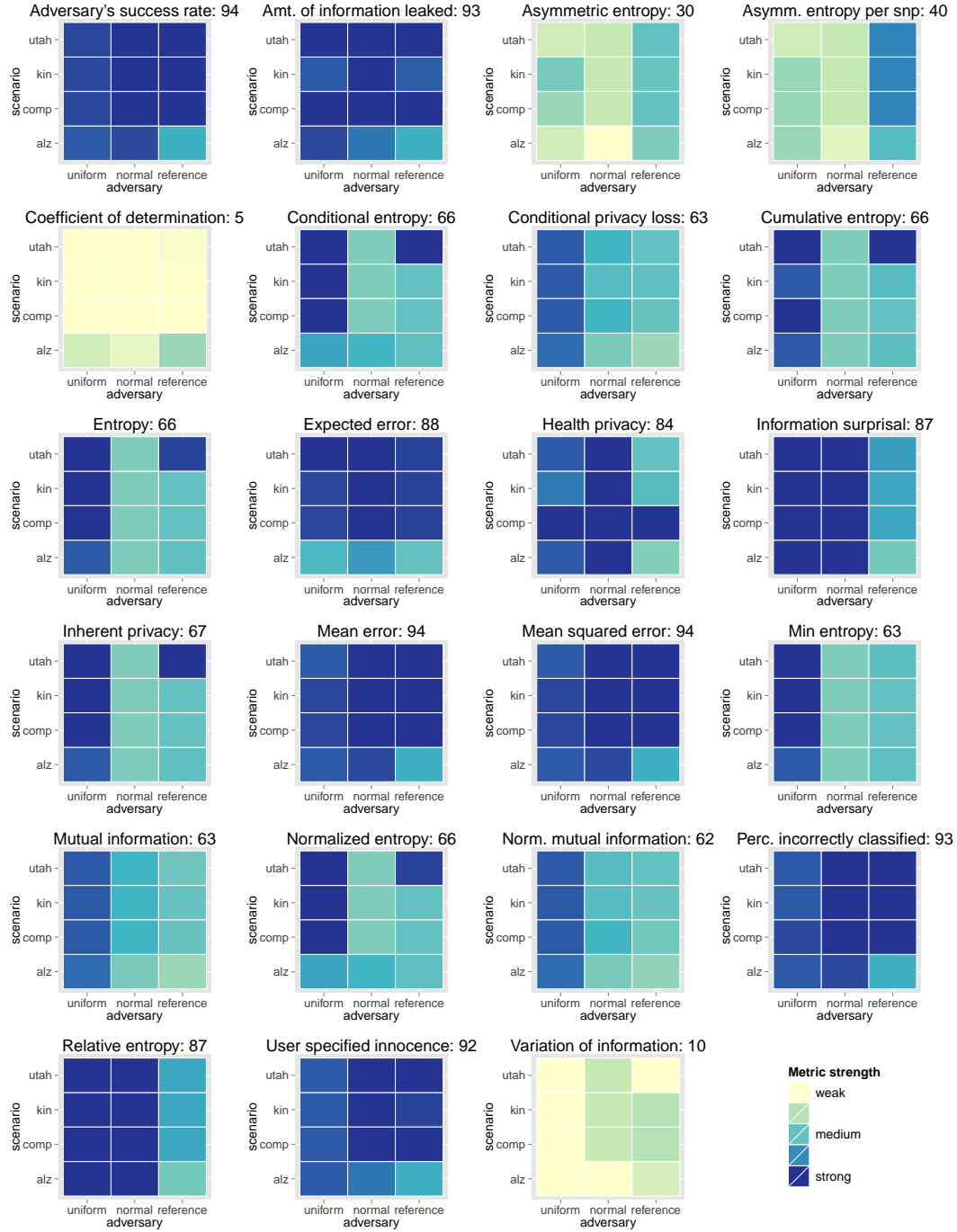


Fig. 3. Strength of 23 genomic privacy metrics shown in heat maps. In each plot, the name of the metric and its overall strength are given in the title, the X axis shows the adversary model, the Y axis shows the scenario, and the colors indicate the strength of the metric (from blue=strong to yellow=weak)

We then conducted a sensitivity analysis to study whether the amount of points influences the final strength value. We analyzed a full factorial design with five values for each of the five potential parameters: change in right/wrong direction, no change in mean, presence of peak, and statistical insignificance. The statistical analysis, using analysis of variance (ANOVA) as well as histograms, indicated that one parameter was statistically not significant (no change in mean), and our algorithm therefore assigns no points to this parameter. All other parameters were significant at  $p < 0.001$ . The specific point values only have a small influence on the final strength level: the mean strength level across all parameter settings and all metrics is within 6% of the mean value resulting from our chosen parameter setting ( $< 2\%$  on average).

## 5.4. Results

Figure 3 shows the heat maps for the strengths of all 23 genomic privacy metrics, obtained according to Algorithm 1. Most entropy-based metrics (asymmetric entropy, conditional entropy, cumulative entropy, min-entropy, normalized entropy, and inherent privacy) behave similarly to entropy, resulting in similar heat maps. These metrics have clear weaknesses for two adversary types (normal and reference) and should therefore only be used in combination with other metrics. A similar behavior, albeit less pronounced, can also be observed for mutual information, normalized mutual information, and conditional privacy loss.

Relative entropy and information surprisal are the only two information theoretic metrics that produce consistently good results, i.e. they produce consistent measurements regardless of the adversary model and scenario. Other strong metrics are the adversary’s success and error (expected estimation error, mean error, mean squared error, percentage incorrectly classified), as well as metrics measuring the number of SNPs that are leaked or remain private. These metrics can be recommended for use in genomic privacy.

The metrics that performed worst in our tests are the coefficient of determination, variation of information, and asymmetric entropy. These metrics do not produce good measurements in any scenario or for any adversary model, and can therefore not be recommended for use in genomic privacy.

**5.4.1. Results for Kin Privacy.** For the majority of genomic privacy metrics, their strength does not vary when they are applied only to related individuals (*kin* and *utah* scenarios, top two rows of each heat map in Figure 3) as compared to a large sample of unrelated individuals (*comparison* scenario, third row). However, some metrics, for example entropy, exhibit a striking difference between the two kin privacy datasets in the strengths indicated for the reference adversary. To explain this, we first note that both the *kin* and *utah* scenarios consist only of a small number of individuals (13 and 17, respectively). Second, all individuals in the *utah* dataset are related to each other, whereas the relationships in the *openSNP* dataset are between pairs or groups of three. Because the reference adversary is based on population-wide allele frequencies, the individuals in the *utah* dataset (top row) would all tend to have the same deviation from these frequencies, and the deviation in this particular case makes some metrics appear stronger compared to the more random sample of individuals in the *openSNP* dataset (second row). This difference in strength occurs only for medium-strength metrics, but does not occur in metrics that are strong across all adversaries and scenarios. This emphasizes the need to select strong privacy metrics.

**5.4.2. Results for Aggregation and Normalization.** Normalization aims to bring the metric values for different scenarios into a common value range to allow comparisons. As the heat maps for normalized entropy and normalized mutual information in Figure 1 show, normalization does not change the strength of a privacy metric with regard to monotonicity.

Aggregation aims to combine the metric values for all SNPs belonging to one individual, effectively reducing the amount of data to analyze. Figure 4 shows the strengths of four aggregated metrics, using two different aggregation methods: an arithmetic mean and a

mean weighted with population-wide minor allele frequencies (denoted *maf* in the Figure). As the comparison between the base metrics in Figure 3 and the aggregated metrics in Figure 4 shows, aggregation does not affect the strength of privacy metrics, regardless of the aggregation method used.

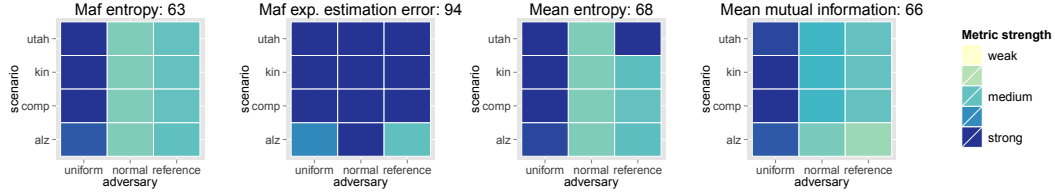


Fig. 4. Strength of four aggregated privacy metrics shown in heat maps. In each plot, the name of the metric and its overall strength are given in the title, the X axis shows the adversary model, the Y axis shows the scenario, and the colors indicate the strength of the metric (from blue=strong to yellow=weak)

## 5.5. Influence of Parameter Settings

**5.5.1. Parameter Settings for 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 the magnitude of the final value. The value of health privacy therefore depends mostly on the value of the base metric.

Figure 5 shows health privacy for five base metrics: normalized entropy, normalized mutual information, expected estimation error, information surprisal, and relative entropy. In every case, health privacy behaves very similar to its base metric. All base metrics perform well for the *uniform* adversary. For the *normal* adversary, entropy and normalized mutual information (Figures 5a and 5b) have their highest values in the middle of the adversary spectrum and thus do not give a good indication of the achieved privacy level. Expected estimation error, relative entropy, and information surprisal (Figures 5c – 5e) have strictly decreasing values for increasing adversary strengths and are thus useful to quantify the privacy level. For the *reference* adversary, most metrics are of average strength because they cannot distinguish some of the adversary strengths.

Figure 6 shows the strengths of health privacy with different base metrics in a heat map. We can see that the strength of health privacy corresponds to the strength of the base metric (compare with the third row in the heat maps in Figure 3). This means that health privacy is a useful way of aggregating per-SNP metrics into a single per-individual metric, provided that the base metric is appropriate.

**5.5.2. Parameter Settings for Amount of Information Leaked and User-Specified Innocence.** Both the amount of information leaked and user-specified innocence have a threshold parameter that indicates when a SNP value is considered leaked resp. private. We found that setting the threshold for leaked information close to 1 resulted in zero leaked SNPs for weak adversaries, and 100% leaked SNPs for strong adversaries. The reverse is true for user-specified innocence, when its threshold is set to 0. This setting therefore doesn't allow to distinguish adversaries of different strengths. We found that thresholds of 0.7 for the amount of information leaked and 0.3 for user-specified innocence allow to distinguish most of the adversary strength levels. Combining the two metrics reveals additional information, because in addition to the number of leaked and private SNPs, the combination also shows for how many SNPs the leakage status is uncertain.



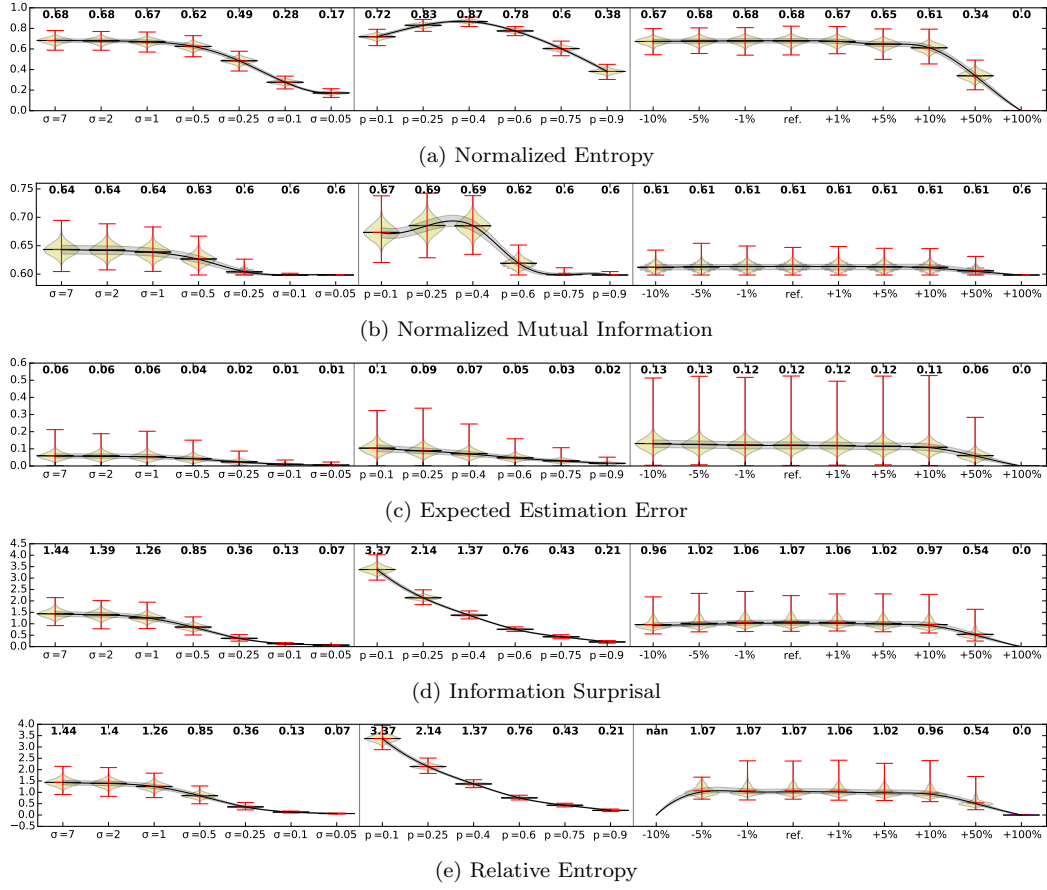


Fig. 5. Health Privacy with different base metrics, based on 100 equally weighted SNPs, evaluated using three adversary models, from left to right: *uniform*, *normal*, and *reference*. Adversary strengths for each model are ordered weakest to strongest from left to right.

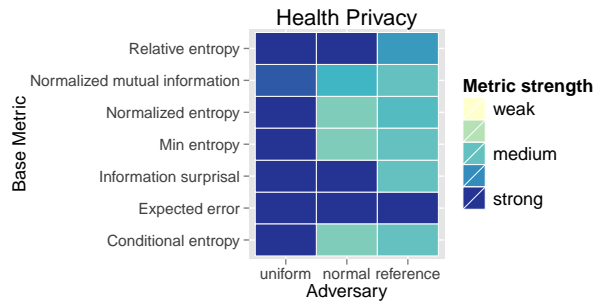


Fig. 6. Strength of health privacy with different base metrics shown in a heat map. The X axis shows the adversary model, the Y axis shows the base metric, and the colors indicate the strength of the metric (from blue=strong to yellow=weak)

## 6. CASE STUDY: ALZHEIMER'S DISEASE

We applied the privacy metrics to study privacy with respect to late-onset Alzheimer's disease. We use the case study to illustrate the process of selecting privacy metrics for a real scenario, and to 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 genomic privacy scenario.

### 6.1. Choice of SNPs for Alzheimer's Disease

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 three SNPs that are present in 695 genomes in the openSNP dataset: rs7412, rs429358 [Bertram et al. 2007], and rs75932628 [Guerreiro et al. 2013]. New associations with Alzheimer's are discovered frequently (see references in [Bertram et al. 2007]), but because these are not available for most individuals in the openSNP dataset, we use only the three SNPs mentioned above.

### 6.2. Selection of Privacy Metrics

The second task is the selection of privacy metrics. Following the process described in [Wagner and Eckhoff 2015], we use eight questions to guide the selection:

- (1) **Output measures.** Wagner and Eckhoff [2015] propose eight categories of output measures and suggest that metrics from as many categories as possible should be selected. Our study includes metrics from only five categories (uncertainty, information gain/loss, similarity/diversity, adversary's success probability, and error). However, the only metric belonging to the similarity/diversity category is the coefficient of determination which, as we have shown above, is not a suitable metric for genomic privacy. We will therefore select metrics from each of the other four categories.
- (2) **Adversary models.** All metrics in our study are computed using the adversary's estimate. This question therefore does not influence our choice of metrics directly.
- (3) **Data source** refers to the data adversaries would use to perform their attack. In our scenario, data could be either observable or published data. Neither data source restricts our choice of metrics.
- (4) **Availability of input data.** In this study, we have access to all input data required by different metrics, including knowledge of the adversary estimate, the true outcome, and parameter settings. This question does therefore not influence our choice of metrics.
- (5) **Target audience.** Even though this paper is targeted at academics, some target audiences may require metrics that can be interpreted easily. We therefore discuss below how each metric can be interpreted.
- (6) **Related work** in genomic privacy has used entropy, expected estimation error, adversary's success rate, and health privacy. We should therefore consider including these four metrics.
- (7) **Strength of metrics.** We can refer to the heat maps in Figure 3 for results about the strength of privacy metrics. The bottom row of each heat map indicates the results specific to the Alzheimer scenario we are interested in here. We list the strongest metrics in the *strong metrics* column of Table III.
- (8) **Implementation of metrics.** We have relied on generic implementations of entropy and mutual information available in Python packages. To the best of our knowledge, validated implementations of specific privacy metrics are not available, and therefore this question does not influence our choice of metric.

Considering our answers to the eight questions, we see that we need to select strong metrics from four categories and include the four metrics considered in related work. Table III shows how strong metrics and metrics from related work fit into the four categories,

with our choice of metrics highlighted in *italics*. In total, we select seven metrics: the four related work metrics, two strong metrics to add to the information gain/loss category, and one strong metric to add to the error category.

Table III. Metric Selection

Category	Strong metrics	Related work metrics
Adversary's success probability	<i>adversary's success rate</i> user-specified innocence	<i>adversary's success rate</i>
Error	expected estimation error <i>mean squared error</i> health privacy (error)	<i>expected estimation error</i> <i>health privacy (error)</i>
Information gain/loss	percentage incorrectly classified <i>amount of leaked information</i> <i>relative entropy</i> information surprisal health privacy (inf. surprisal)	
Uncertainty		<i>entropy</i>

### 6.3. Choice of Metric Parameters

The third task is to choose parameter settings for the selected metrics. Health privacy uses weights for individual SNPs, ideally chosen to reflect how much each SNP contributes to the overall disease risk. This would usually be done using scientific studies or tables released by insurance companies [Ayday et al. 2013; Humbert et al. 2013]. For the sake of our case study, we chose equal weights and severities for the three SNPs.

The amount of information leaked uses a parameter for the threshold probability which depends on the privacy preferences of individual users. For our case study, we chose a threshold of 70%, which means that SNPs are leaked if the adversary's estimate of the true value is above 70%.

### 6.4. Interpretation of Results

After conducting the privacy measurement, the fourth and final task is to plot and interpret the results.

**6.4.1. Interpreting the Values of Privacy Metrics.** To interpret what the values of each privacy metric mean, we show violin plots for the seven selected metrics in Figure 7. (For completeness, we show the results for the remaining metrics in the appendix.) For each metric, the plots show three groups of violins, corresponding to the three adversary models. The distributions of the adversary's success rate (Figure 7a) have up to four peaks at 0,  $\frac{1}{3}$ ,  $\frac{2}{3}$  and 1, showing that the adversary can infer either 0, 1, 2, or 3 SNP values. Within each adversary group, the values are monotonic (non-decreasing), but not all strength levels can be distinguished. The amount of information leaked behaves similarly to the adversary's success rate, showing how often 0, 1, 2, or 3 SNP values have leaked (Figure 7b). Looking at the left-most adversary group, we can see that the amount of information leaked increases at a higher adversary strength level than the adversary's success rate (compare with Figure 7a). This is because the threshold parameter for the amount of information leaked is 70%, whereas the success rate can count successes with lower probabilities.

Entropy (Figure 7c) measures the adversary's uncertainty and therefore cannot reliably indicate the user's privacy level. Entropy performs worst for the *normal* adversary model, because it peaks at a medium adversary strength. This is consistent with the findings from Figure 3. The values of entropy indicate how many bits of information are contained in the adversary's estimate, with high values indicating more uncertainty.

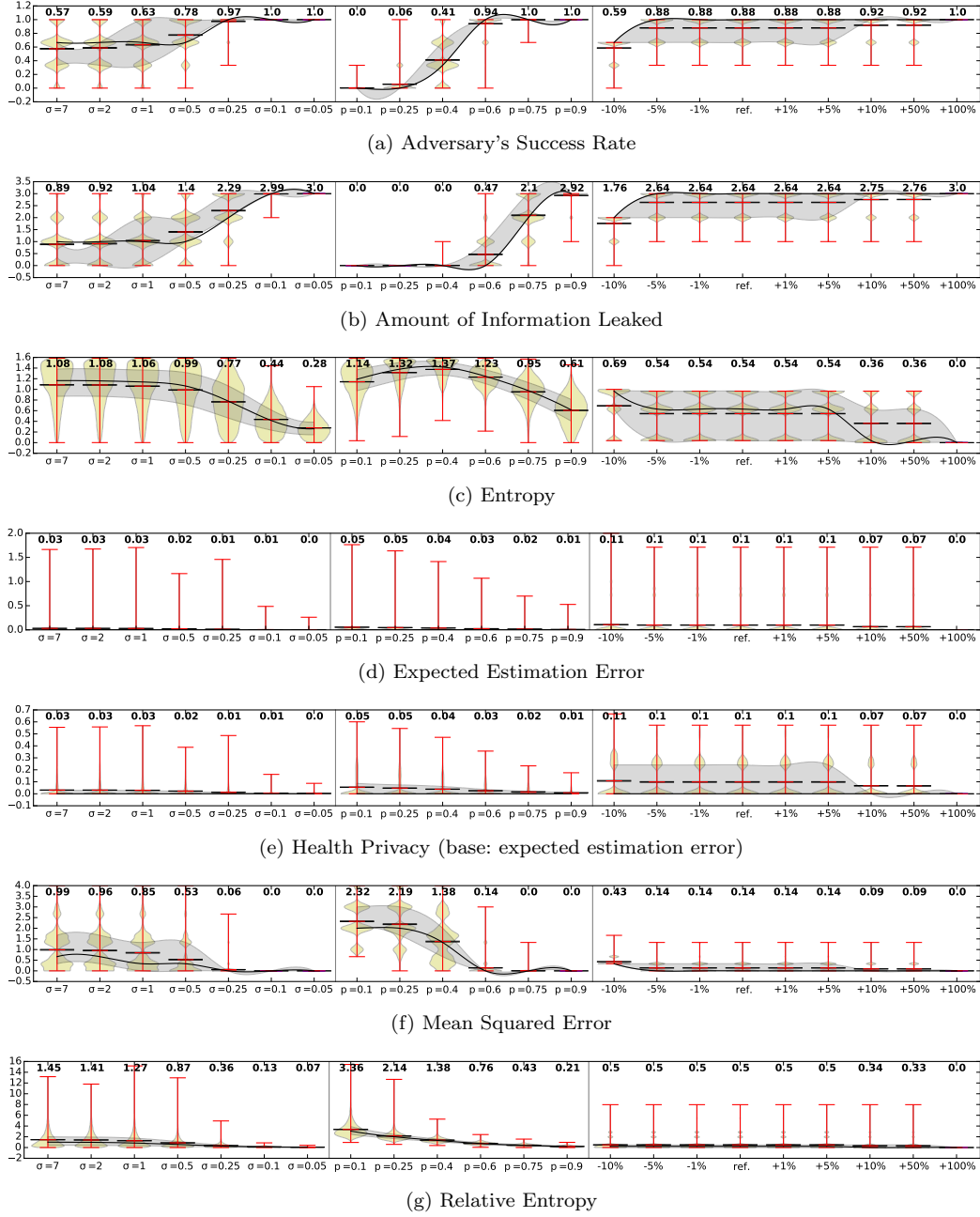


Fig. 7. Strong privacy metrics for the Alzheimer's disease scenario, evaluated using three adversary models, from left to right: *uniform*, *normal*, and *reference*. Adversary strengths for each model are ordered weakest to strongest from left to right.

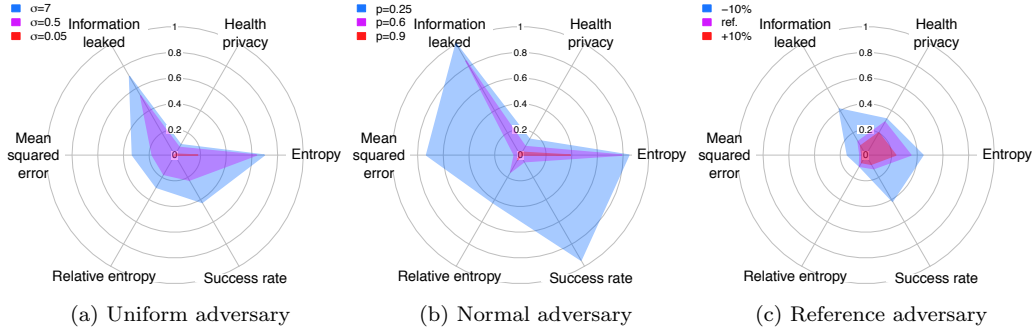


Fig. 8. Radar plots visualizing the privacy level of six privacy metrics for three strength levels of each adversary type

The expected estimation error (Figure 7d) and health privacy with the expected estimation error as base metric (Figure 7e) behave similarly to the initial evaluation. However, both metrics show higher values for the reference adversary than for the other adversary types. This would indicate that the reference adversary performs worse (i.e., has a higher error) even than the adversary whose estimate is furthest from the true value. This differs from what the other metrics show for the reference adversary.

The mean squared error (Figure 7f) 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 meaning of the values is not intuitively clear.

Relative entropy (Figure 7g) indicates how much additional information, measured in bits, the adversary needs to reconstruct the true values. This amount of information is similar to the amount of surprise the adversary will experience upon learning the true value, i.e. the information surprisal metric (see Figure 9b in the appendix).

Based on these findings in this and the previous sections, we rated each metric based on how easy it is to understand what its values mean, and how easily it can be interpreted. We summarize the ratings in Table II (column *Intuitiveness*).

**6.4.2. Interpreting the Overall Privacy Level.** The violin plots presented in Figure 7 are comprehensive, but they make it hard to tell what an individual’s overall privacy level is against a specific adversary. We propose radar plots to visualize the overall privacy level indicated by a combination of metrics.

Figure 8 shows one radar plot for each adversary type. To keep the plots clean, we plot only three strength levels per adversary type. We excluded the expected estimation error because of its similarity to health privacy with the expected estimation error as base metric. The values for each metric have been normalized to the  $[0, 1]$  value range using the 10th and 90th percentile of values across all adversary strengths. In addition, we inverted the values for lower-better metrics. As a result, a larger area in the plots directly corresponds to a higher privacy level.

The plots allow comparisons between the different adversary types, for example clearly showing that the weakest *normal* adversary (Figure 8b, light blue color) is much weaker than the weakest *uniform* adversary in Figure 8a, because the privacy area for the *normal* adversary is much larger. The plots also confirm our expectation that the reference adversaries are comparatively strong because their privacy areas are smaller than for the other two adversary types. A real-world evaluation of privacy may not vary the adversary strengths as we have done, but instead vary parameters of a new privacy-enhancing technology. Since the strength of the adversary and the strength of a privacy enhancing technology are essentially two sides of the same coin, we expect that radar plots will be able to highlight

differences between privacy enhancing technologies in the same way as differences between adversaries.

## 7. CONCLUSIONS AND FUTURE WORK

We measured the strengths of 23 published genomic privacy metrics. We introduced monotonicity as the key indicator of a metric's strength, i.e. metrics should show decreasing privacy for increasing adversary strength. We tested each of the 23 metrics in three different scenarios, for adversaries of different strengths, and found that only 7 out of 23 metrics were strong across scenarios and adversary types. The 7 strong metrics were the adversary's success rate, the amount of information leaked, health privacy (with information surprisal or relative entropy as base metric), information surprisal, percentage incorrectly classified, relative entropy, and user-specified innocence. Furthermore, we found that none of the metrics we tested were sufficiently reliable when used in isolation. Therefore, we recommend that several strong metrics that measure different outputs should be used together. Finally, we introduced two visualization methods previously not used in the privacy field – heat maps and radar plots. Our systematic comparison of genomic privacy metrics will enable researchers to make informed and consistent decisions about the selection of privacy metrics and privacy enhancing technologies.

In future work, we will measure the strength of privacy metrics in other application domains, e.g., vehicular networking and smart metering. Future work also needs to study whether there are additional requirements for privacy metrics aside from monotonicity, and whether privacy metrics should satisfy the conditions for metrics in a mathematical sense.

## ACKNOWLEDGMENTS

This work used the ARCHER UK National Supercomputing Service (<http://www.archer.ac.uk>).

## REFERENCES

- Dakshi Agrawal and Charu C. Aggarwal. 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS 2001)*. ACM, Santa Barbara, CA, USA, 247–255.
- James Alexander and Jonathan Smith. 2003. Engineering Privacy in Public: Confounding Face Recognition. In *Proc. 3rd Int. Workshop on Privacy Enhancing Technologies (PET 2003) (LNCS 2760)*. Springer, Dresden, Germany, 88–106.
- Christer Andersson and Reine Lundin. 2008. On the Fundamentals of Anonymity Metrics. In *Proc. 3rd IFIP Int. Summer School on The Future of Identity in the Information Society*. Springer, Karlstad, Sweden, 325–341.
- Erman Ayday, Jean Louis Raisaro, Urs Hengartner, Adam Molyneaux, and Jean-Pierre Hubaux. 2014. Privacy-preserving processing of raw genomic data. In *Data Privacy Management and Autonomous Spontaneous Security*. Springer, 133–147.
- Erman Ayday, Jean Louis Raisaro, and Jean-Pierre Hubaux. 2013. Personal Use of the Genomic Data: Privacy vs. Storage Cost. In *Proc. IEEE Global Communications Conf. (GLOBECOM 2013)*. IEEE, Atlanta, GA, USA, 2723–2729.
- Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. 2013. Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine. In *Proc. 12th ACM Workshop on Workshop on Privacy in the Electronic Society (WPES'13)*. ACM, Berlin, Germany, 95–106. DOI:<http://dx.doi.org/10.1145/2517840.2517843>
- Elisa Bertino, Dan Lin, and Wei Jiang. 2008. A Survey of Quantification of Privacy Preserving Data Mining Algorithms. In *Privacy-Preserving Data Mining: Models and Algorithms*. Number 34 in Advances in Database Systems. Springer, Chapter 8, 183–205.
- Lars Bertram, Matthew B. McQueen, Kristina Mullin, Deborah Blacker, and Rudolph E. Tanzi. 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics* 39, 1 (January 2007), 17–23.
- Terence Chen, Abdelberi Chaabane, Pierre Ugo Tournoux, Mohamed-Ali Kaafar, and Roksana Boreli. 2013. How Much Is Too Much? Leveraging Ads Audience Estimation to Evaluate Public Profile Uniqueness.

- In *Proc. 13th Int. Symp. on Privacy Enhancing Technologies (PETS 2013) (LNCS 7981)*. Springer, Bloomington, IN, USA, 225–244.
- Xihui Chen and Jun Pang. 2012. Measuring Query Privacy in Location-based Services. In *Proc. 2nd ACM Conf. on Data and Application Security and Privacy (CODASPY'12)*. ACM, San Antonio, TX, USA, 49–60. DOI:<http://dx.doi.org/10.1145/2133601.2133608>
- Sebastian Clauß and Stefan Schiffner. 2006. Structuring Anonymity Metrics. In *Proc. 13th ACM Conf. on Computer and Communications Security 2006 (CCS'06): 2nd ACM Workshop on Digital Identity Management (DIM'06)*. ACM, Alexandria, VA, USA, 55–62. DOI:<http://dx.doi.org/10.1145/1179529.1179539>
- Yuxin Deng, Jun Pang, and Peng Wu. 2007. Measuring Anonymity with Relative Entropy. In *Proc. 8th Int. Workshop on Formal Aspects in Security and Trust (FAST 2011)*. Springer, Leuven, Belgium, 65–79.
- Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. 2003. Towards Measuring Anonymity. In *Privacy Enhancing Technologies*. 54–68.
- Claudia Diaz, Carmela Troncoso, and George Danezis. 2007. Does Additional Information Always Reduce Anonymity?. In *Proc. 6th ACM Workshop on Privacy in Electronic Society (WPES '07)*. ACM, Alexandria, VA, USA, 72–75. DOI:<http://dx.doi.org/10.1145/1314333.1314347>
- Radoje Drmanac, Andrew B. Sparks, Matthew J. Callow, Aaron L. Halpern, Norman L. Burns, Bahram G. Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B. Nilsen, George Yeung, Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P. Pant, Jonathan Baccash, Adam P. Borcharding, Anushka Brownley, Ryan Ceden, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C. Ebert, Coleen R. Hacker, Robert Hartlage, Brian Hauser, Steve Huang, Yuan Jiang, Vitali Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E. McBride, Matt Morenzoni, Robert E. Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock A. Peters, Joe Peterson, Charit L. Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M. Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanovich, Karen W. Shannon, Conrad G. Sheppy, Michel Sun, Joseph V. Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R. Oliphant, William C. Banyai, Bruce Martin, Dennis G. Ballinger, George M. Church, and Clifford A. Reid. 2010. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* 327, 5961 (Jan. 2010), 78–81. DOI:<http://dx.doi.org/10.1126/science.1181498>
- Cynthia Dwork. 2006. Differential Privacy. In *Proc. 33rd Int. Colloq. on Automata, Languages and Programming (ICALP 2006) (LNCS 4052)*. Springer, Venice, Italy, 1–12.
- Julien Freudiger, Maxim Raya, Márk Félegyházi, Panos Papadimitratos, and Jean-Pierre Hubaux. 2007. Mix-Zones for Location Privacy in Vehicular Networks. In *Proc. 1st Int. Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS 2007)*. ICST, Vancouver, Canada.
- Michael T. Goodrich. 2009. The Mastermind Attack on Genomic Data. In *30th IEEE Symposium on Security and Privacy*. 204–218.
- Bastian Greshake, Philipp E. Bayer, Helge Rausch, and Julia Reda. 2014. openSNP—A Crowdsourced Web Resource for Personal Genomics. *PLoS ONE* 9, 3 (March 2014). DOI:<http://dx.doi.org/10.1371/journal.pone.0089204>
- Rita Guerreiro, Aleksandra Wojtas, Jose Bras, Minerva Carrasquillo, Ekaterina Rogaeva, Elisa Majounie, Carlos Cruchaga, Celeste Sassi, John S.K. Kauwe, Steven Younkin, Lilinaz Hazrati, John Collinge, Jennifer Pocock, Tammayn Lashley, Julie Williams, Jean-Charles Lambert, Philippe Amouyel, Alison Goate, Rosa Rademakers, Kevin Morgan, John Powell, Peter St. George-Hyslop, Andrew Singleton, and John Hardy. 2013. TREM2 Variants in Alzheimer's Disease. *New England Journal of Medicine* 368, 2 (January 2013), 117–127.
- Daojing He, S. Chan, and M. Guizani. 2015. Privacy and incentive mechanisms in people-centric sensing networks. *IEEE Communications Magazine* 53, 10 (2015), 200–206. DOI:<http://dx.doi.org/10.1109/MCOM.2015.7295484>
- Jerry L. Hintze and Ray D. Nelson. 1998. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician* 52, 2 (May 1998), 181–184.
- Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* 4, 8 (August 2008), e1000167.
- Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2013. Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy. In *Proc. 20th ACM Conf. on Computer and Communications Security (CCS'13)*. ACM, Berlin, Germany, 1141–1152. DOI:<http://dx.doi.org/10.1145/2508859.2516707>

- Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2014. Reconciling Utility with Privacy in Genomics. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES '14)*. ACM, Scottsdale, AZ, USA, 11–20.
- Mathias Humbert, Kévin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux. 2015. De-anonymizing Genomic Databases Using Phenotypic Traits. DOI: <http://dx.doi.org/10.1515/popets-2015-0020>
- Georgios Kalogridis, Costas Efthymiou, Stojan Z. Denic, Tim A. Lewis, and Rafael Cepeda. 2010. Privacy for Smart Meters: Towards Undetectable Appliance Load Signatures. In *Proc. 1st Int. Conf. on Smart Grid Communications (SmartGridComm 2010)*. IEEE, Gaithersburg, MD, USA, 232–237.
- Zhen Lin, Michael Hewett, and Russ B. Altman. 2002. Using Binning to Maintain Confidentiality of Medical Data. In *Proc. AMIA Symp. (AMIA 2002)*. San Antonio, TX, USA, 454–458.
- Changchang Liu and Prateek Mittal. 2016. LinkMirage: Enabling Privacy-preserving Analytics on Social Relationships. In *NDSS*.
- Bradley A. Malin. 2005. Protecting DNA sequence anonymity with generalization lattices. *Methods of Information in Medicine* 44, 5 (2005), 687–692.
- Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98, 5 (May 2007), 873–895.
- Steven J. Murdoch. 2014. Quantifying and Measuring Anonymity. In *Data Privacy Management and Autonomous Spontaneous Security*. Springer Berlin Heidelberg, 3–13.
- Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing Social Networks. In *30th IEEE Symposium on Security and Privacy*. 173–187. DOI: <http://dx.doi.org/10.1109/SP.2009.22>
- Simon Oya, Carmela Troncoso, and Fernando Pérez-González. 2014. Do Dummies Pay Off? Limits of Dummy Traffic Protection in Anonymous Communications. In *Proc. 14th Int. Symp. on Privacy Enhancing Technologies (PETS 2014) (LNCS 8555)*. Springer, Amsterdam, Netherlands, 204–223.
- Ravi Sachidanandam, David Weissman, Steven C. Schmidt, Jerzy M. Kakol, Lincoln D. Stein, Gabor Marth, Steve Sherry, James C. Mullikin, Beverley J. Mortimore, David L. Willey, Sarah E. Hunt, Charlotte G. Cole, Penny C. Coggill, Catherine M. Rice, Zemin Ning, Jane Rogers, David R. Bentley, Pui-Yan Kwok, Elaine R. Mardis, Raymond T. Yeh, Brian Schultz, Lisa Cook, Ruth Davenport, Michael Dante, Lucinda Fulton, LaDeana Hillier, Robert H. Waterston, John D. McPherson, Brian Gilman, Stephen Schaffner, William J. Van Etten, David Reich, John Higgins, Mark J. Daly, Brendan Blumenstiel, Jennifer Baldwin, Nicole Stange-Thomann, Michael C. Zody, Lauren Linton, Eric S. Lander, and David Altshuler. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 6822 (Feb. 2001), 928–933. DOI: <http://dx.doi.org/10.1038/35057149>
- Sahel Samani, Zhicong Huang, Erman Ayday, Mark Elliot, Jacques Fellay, Jean-Pierre Hubaux, and Zoltán Kutalik. 2015. Quantifying Genomic Privacy via Inference Attack with High-Order SNV Correlations. In *2015 IEEE Security and Privacy Workshops (SPW)*. 32–40.
- Andrei Serjantov and George Danezis. 2002. Towards an Information Theoretic Metric for Anonymity. In *Proc. 2nd Int. Symp. on Privacy Enhancing Technologies (PETS 2002) (LNCS 2482)*. Springer, San Francisco, CA, USA, 41–53.
- S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29, 1 (January 2001), 308–311. DOI: <http://dx.doi.org/10.1093/nar/29.1.308>
- Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying Location Privacy. In *Proc. 2011 32nd IEEE Symp. on Security and Privacy (S&P 2011)*. IEEE, Oakland, CA, USA, 247–262. DOI: <http://dx.doi.org/10.1109/SP.2011.18>
- Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570. DOI: <http://dx.doi.org/10.1142/S0218488502001648>
- Paul Syverson. 2013. Why I’m Not an Entropist. In *Proc. 17th Int. Workshop on Security Protocols (LNCS 7028)*. Springer, Cambridge, UK, 213–230.
- Sarah A. Tishkoff and Kenneth K. Kidd. 2004. Implications of biogeography of human populations for ‘race’ and medicine. *Nature Genetics* 36 (Oct. 2004), S21–S27. DOI: <http://dx.doi.org/10.1038/ng1438>
- Isabel Wagner. 2015. Genomic Privacy Metrics: A Systematic Comparison. In *2015 IEEE Security and Privacy Workshops (SPW)*. 50–59. DOI: <http://dx.doi.org/10.1109/SPW.2015.15>
- Isabel Wagner and David Eckhoff. 2015. Technical Privacy Metrics: a Systematic Survey. *arXiv:1512.00327 [cs, math]* (Dec. 2015). <http://arxiv.org/abs/1512.00327>
- Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 2009. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study. In *Pro-*



- ceedings of the 16th ACM Conference on Computer and Communications Security (CCS '09). ACM, Chicago, IL, USA, 534–544.
- Kris Wetterstrand. 2016. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). (2016). <https://www.genome.gov/sequencingcostsdata>
- Ye Zhu and Riccardo Bettati. 2005. Anonymity vs. information leakage in anonymity systems. In *Proc. 25th IEEE Int. Conf. on Distributed Computing Systems (ICDCS 2005)*. IEEE, Columbus, Ohio, USA, 514–524.

## APPENDIX

### A. SUPPLEMENTARY FIGURES

Figures 9, 10 and 11 show strong, average strength, and weak privacy metrics for the alzheimer scenario which were not selected in Section 6.

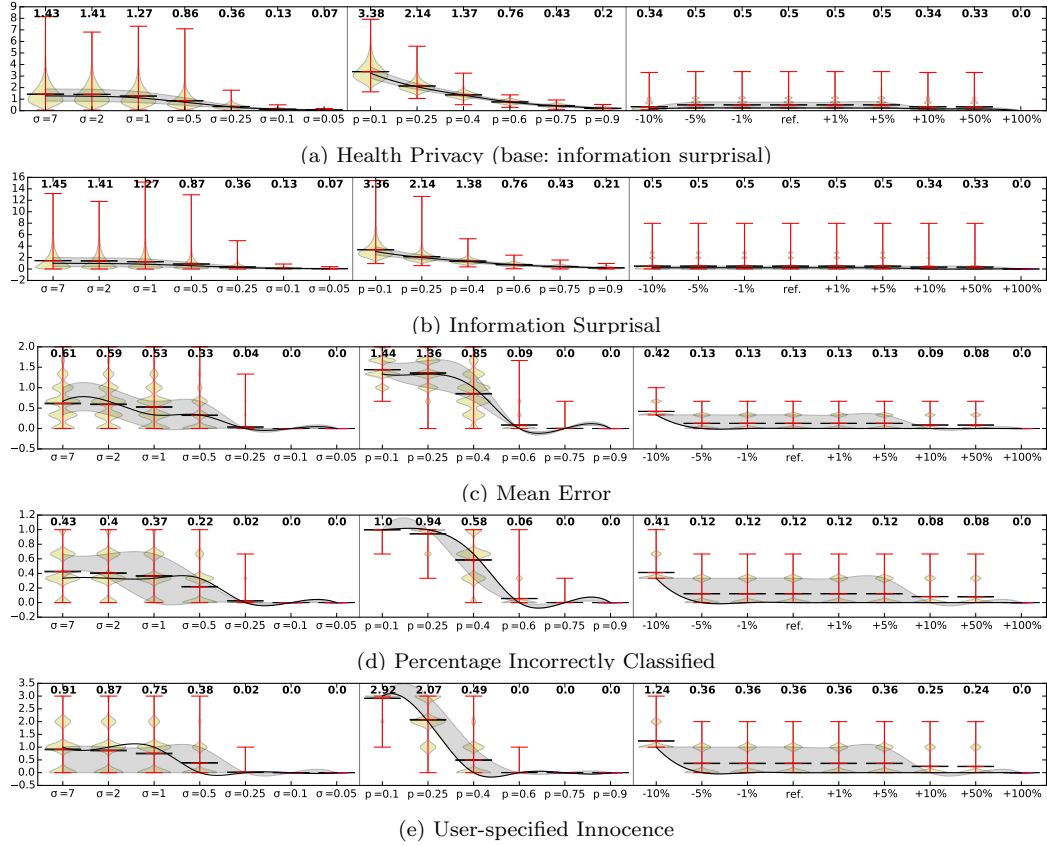


Fig. 9. Strong privacy metrics for the Alzheimer's disease scenario, evaluated according to three adversary models. Adversary strengths for each model are ordered weakest to strongest from left to right.

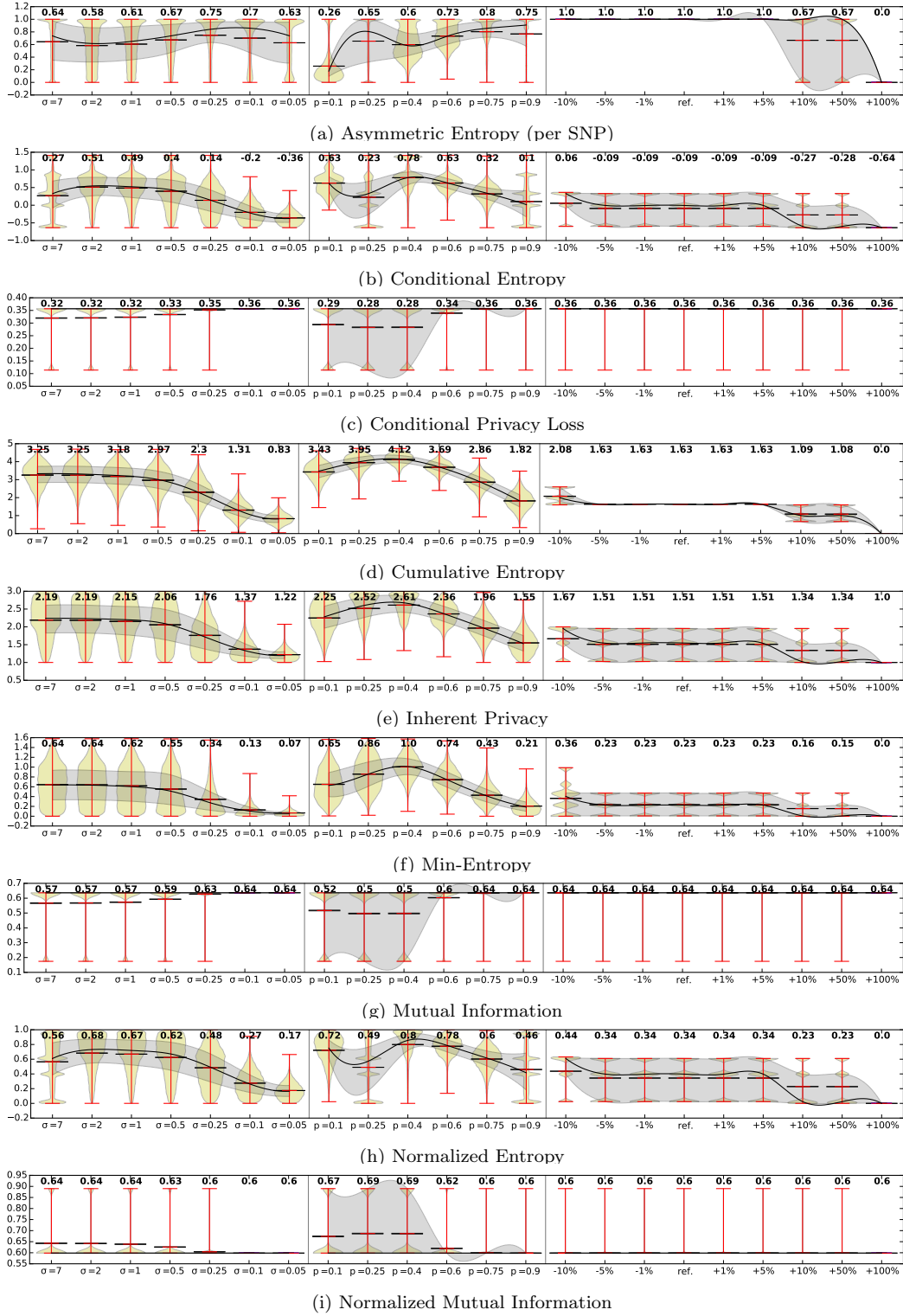


Fig. 10. Average strength privacy metrics for the Alzheimer's disease scenario, evaluated according to adversary strength, ordered weakest to strongest from left to right

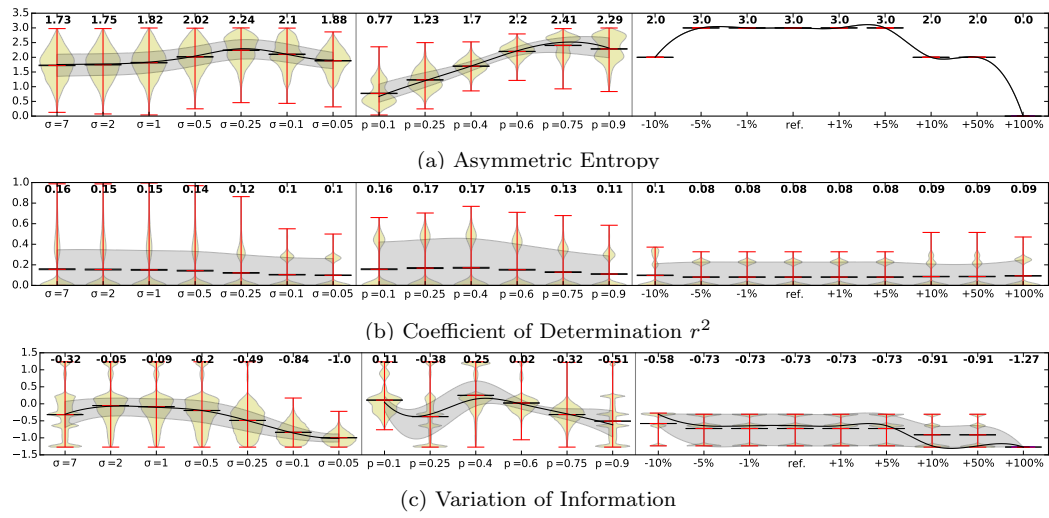


Fig. 11. Weak privacy metrics for the Alzheimer's disease scenario, evaluated according to adversary strength, ordered weakest to strongest from left to right