Int. J. Inf. Secur. (2013) 12:129–149
DOI 10.1007/s10207-012-0182-5

REGULAR CONTRIBUTION

On the measurement of privacy as an attacker's estimation error

David Rebollo-Monedero · Javier Parra-Arnau ·
Claudia Diaz · Jordi Forné

Published online: 11 December 2012

Abstract A wide variety of privacy metrics have been proposed in the literature to evaluate the level of protection offered by privacy-enhancing technologies. Most of these metrics are specific to concrete systems and adversarial models and are difficult to generalize or translate to other contexts. Furthermore, a better understanding of the relationships between the different privacy metrics is needed to enable more grounded and systematic approach to measuring privacy, as well as to assist system designers in selecting the most appropriate metric for a given application. In this work, we propose a theoretical framework for privacy-preserving systems, endowed with a general definition of privacy in terms of the estimation error incurred by an attacker who aims to disclose the private information that the system is designed to conceal. We show that our framework permits interpreting and comparing a number of well-known metrics under a common perspective. The arguments behind these interpretations are based on fundamental results related to the theories of information, probability, and Bayes decision.

Keywords Privacy · Criteria · Metrics · Estimation · Bayes decision theory · Statistical disclosure control · Anonymous communication systems · Location-based services

D. Rebollo-Monedero · J. Parra-Arnau · J. Forné (✉)
Department of Telematics Engineering, Universitat Politècnica de Catalunya, C. Jordi Girona 1-3, 08034 Barcelona, Catalonia, Spain
e-mail: jforne@entel.upc.edu

D. Rebollo-Monedero
e-mail: david.rebollo@entel.upc.edu

J. Parra-Arnau
e-mail: javier.parra@entel.upc.edu

C. Diaz
Katholieke Universiteit Leuven, ESAT/SCD/IBBT-COSIC,
Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium
e-mail: claudia.diaz@esat.kuleuven.be

1 Introduction

The widespread use of information and communication technologies to conduct all kinds of activities has in recent years raised privacy concerns. There is a wide diversity of applications with a potential privacy impact, from social networking platforms to e-commerce or mobile phone applications.

A variety of privacy-enhancing technologies (PETs) have been proposed to enable the provision of new services and functionalities while mitigating potential privacy threats. The privacy concerns arising in different applications are diverse, and so are the corresponding privacy-enhanced solutions that address these concerns. Similarly, various ad hoc privacy metrics have been proposed in the literature to evaluate the effectiveness of PETs. The relationships between these different metrics have, however, not been investigated in depth, what leads to a fragmentation in the understanding of how privacy properties can be measured.

In this paper, we consider a general, theoretical framework for privacy-preserving systems and propose using the attacker's estimation error as privacy metric. We show that the most widely used privacy metrics, such as k-anonymity, l-diversity, t-closeness, ϵ-differential privacy, as well as information-theoretic metrics such as Shannon's entropy, min-entropy, or mutual information, may be construed as particular cases of the estimation error.

Privacy metrics, accompanied with utility metrics, provide a quantitative means of comparing the suitability of two or more privacy-enhancing mechanisms, in terms of the privacy-utility trade-off posed. Ultimately, such metrics will enable us to systematically build privacy-aware information systems by formulating design decisions as optimization problems, solvable theoretically or numerically, capitalizing on a rich variety of mature ideas and powerful techniques from the wide field of optimization engineering.

We illustrate how the general model can be instantiated in three very different areas of application, namely statistical disclosure control, anonymous communications, and location-based services (LBSs). Statistical disclosure control (SDC) [1] is the research area that deals with the inherent compromise between protecting the privacy of the individuals in a microdata set and ensuring that those data are still useful for researchers. Traditionally, institutes and governmental statistical agencies have systematically gathered information about individuals with the aim of distributing those data to the research community [2]. However, the distribution of this information should not compromise respondents' privacy in the sense of revealing information about specific individuals. Motivated by this, considerable research effort has been devoted to the development of privacy-protecting mechanisms [3–7] to be applied to the microdata sets before their release. In essence, these mechanisms rely upon some form of perturbation that permits enhancing privacy to a certain extent, at the cost of losing some of the data utility with respect to the unperturbed version.

With the aim of assessing the effectiveness of such mechanisms, numerous privacy metrics have been investigated. Probably, the best-known privacy metric is *k-anonymity*, which was first proposed in [8,9]. In an attempt to address the limitations of this proposal, various extensions and enhancements were introduced later in [10–15]. While all these proposals have contributed to some extent to the understanding of the privacy requirements of this field, the SDC research community would undoubtedly benefit from the existence of a rule that could help them decide which privacy metric is the most appropriate for a particular application. In other words, there is a need for the establishment of a framework that enables us to compare those metrics and to formulate them by using a common, general definition of privacy.

In anonymous communications, one of the goals is to conceal who talks to whom against an adversary who observes the inputs and outputs of the anonymous communication channel. Mixes [16–18] are a basic building block for implementing anonymous communication channels. Mixes perform cryptographic operations on messages such that it is not possible to correlate their inputs and outputs based on their bit patterns. In addition, mixes delay and reorder messages to hinder the linking of inputs and outputs based on timing information. Delaying messages has an impact on the usability of the system and therefore imposes a cost on the system. On the other hand, higher delays allow for stronger levels of privacy. There is thus a trade-off between delay (cost) and anonymity (privacy), and optimizing the level of anonymity for a given expected delay is interesting to extract as much protection as possible from the anonymous channel at the lower possible cost.

In the end, we approach the particularly rich, important example of LBSs, where users submit queries along with the location to which those queries refer. An example would be the query "Where is the nearest Italian restaurant?", accompanied by the geographic coordinates of the user's current location. In this scenario, a wide range of approaches have been proposed, many of them based on an intelligent perturbation of the user coordinates submitted to the provider [19]. Essentially, users may contact an *untrusted* LBS provider directly, perturbing their location information so as to hinder providers in their efforts to compromise user privacy in terms of location, although clearly not in terms of query contents and activity, and at the cost of an inaccurate answer. In a nutshell, this approach presents again the inherent trade-off between data utility and privacy common to any perturbative privacy method.

The survey of privacy metrics, the detailed analysis of their connection with information theory, and the mathematical unification as an attacker's estimation error presented in this paper shed new light on the understanding of those metrics and their suitability when it comes to applying them to specific scenarios. In regard to this aspect, two sections are devoted to the classification of several privacy metrics, showing the relationships with our proposal and the correspondence with assumptions on the attacker's strategy. While the former section approaches this from a theoretical perspective, the latter illustrates the applicability of our framework to help system designers choose the appropriate metrics, without having to delve into the mathematical details. We also hope to illustrate the riveting intersection between the fields of information privacy and information theory, in an attempt toward bridging the gap between the respective communities. Moreover, the fact that our metric boils down to an estimation error opens the possibility of applying notions and results from the mature, vast field of estimation theory [20].

2 Related work

In this section, we provide an overview of privacy metrics with an emphasis on those used in the three applications under study: anonymous communications, location-based services, and statistical disclosure control.

2.1 Anonymous communication systems and location-based services

Mixes were proposed by Chaum [16] in 1981 and are a basic building block for implementing high-latency anonymous communications. A mix takes a number of input messages and outputs them in such a way that it is infeasible to link an output to its corresponding input. In order to achieve this goal, the mix changes the appearance (by encrypting and padding messages) and the flow of messages (by delaying and reordering them). Mixmaster [17] and Mixminion [18] are more advanced versions of the Chaumian mix [16], and they have been deployed to provide anonymous email services.

Several metrics have been proposed in the literature to assess the level of anonymity provided by anonymous communication systems (ACSs). Reiter and Rubin [21] define the degree of anonymity as a probability $1-p$, where p is the probability assigned by an attacker to the potential initiators of a communication. In this model, users are more anonymous as they appear (toward a certain adversary) to be less likely of having sent a message, and the metric is thus computed individually for each user and for each communication. Berthold et al. [22] on the other hand define the degree of anonymity as the binary logarithm of the number of users of the system, which may be regarded as a Hartley entropy. This metric only depends on the number of users of the system and does not take into account that some users might appear as more likely senders of a message than others.

Information theoretic anonymity metrics were independently proposed in two papers. The metric proposed by Serjantov and Danezis [23] uses Shannon's entropy as measure of the effective anonymity set size. The metric proposed by Diaz et al. [24] normalizes Shannon's entropy to obtain a degree of anonymity on a scale from 0 to 1.

Toth et al. [25] argue that Shannon entropy may not provide relevant information to some users, as it considers the average instead of the worst-case scenario for a particular user. They suggest using instead a local anonymity measure computed from min-entropy and max-entropy. Clauss and Schiffner [26] proposed Rényi entropy as a generalization of Shannon, min- and max-entropy-based anonymity metrics.

Other anonymity metrics in the literature include possibilistic (instead of probabilistic) approaches, such as those proposed by Syverson and Stubblebine [27], Mauw et al. [28], or Feigenbaum et al. [29]. According to these metrics, subjects are considered anonymous if the adversary cannot determine their actions with absolute certainty. Finally, Edman et al. [30] propose a combinatorial anonymity metric that measures the amount of information needed to reveal the full set of relationships between the inputs and the outputs of a mix. Some extensions of this model were proposed by Gierlichs et al. [31] and by Bagai et al. [32].

Having examined the most relevant metrics in the field of anonymous communications, now we briefly touch upon some of the proposals intended for the scenario of LBS. Particularly, the issue of quantifying privacy in this scenario has been explored in [33] and revisited shortly afterward in [34]. At a conceptual level, we encounter the same underlying principle proposed here, in the sense that the authors propose to measure privacy as the adversary's expected estimation error for that particular context. We shall discuss later in Sect. 6.1 that their specific metric for LBS may be construed as an illustrative special case of our own work and describe notable differences with respect to our generic framework.

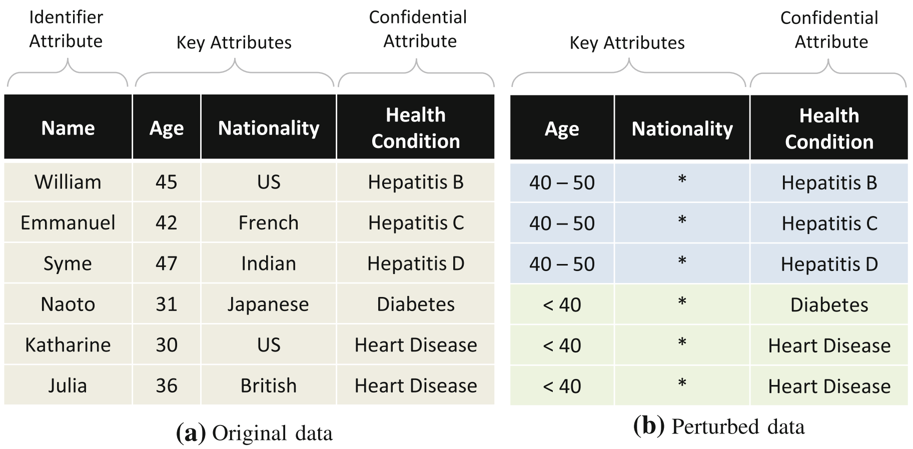
2.2 Privacy criteria in statistical disclosure control

In SDC terminology, a microdata set is a database whose records contain information at the level of individual respondents. In those databases, each row corresponds to an individual and each column, to an attribute. According to the nature of attributes, we may classify them into *identifiers*, *key attributes* or *quasi-identifiers*, or *confidential attributes*. On the one hand, identifiers allow to unequivocally identify individuals. For example, it would be the case of social security numbers or full names, which would be removed before the publication of the microdata set. On the other hand, key attributes are those attributes that, in combination, may be linked with external information to reidentify the respondent to whom the records in the microdata set refer. Last but not least, confidential attributes contain sensitive information on the respondents, such as health condition, political affiliation, religion, or salary.

k-Anonymity [8,9] is the requirement that each tuple of key attribute values be shared by at least k records in the database. This condition is illustrated in Fig. 1, where a microdata set is k-anonymized before publishing it. Particularly, this privacy criterion is enforced by using generalization and suppression, two mechanisms by which key attribute values are, respectively, coarsened and eliminated. As a result, all key attribute values within each group are replaced by a common tuple, and thus, a record cannot be unambiguously linked to any public database containing identifiers. Consequently, k-anonymity is said to protect microdata against *linking attacks*.

Unfortunately, while this criterion prevents identity disclosure, it may fail against the disclosure of the confidential attribute. Concretely, suppose that a privacy attacker knows Emmanuel's key attribute values. If the attacker learns that he is included in the released table depicted in Fig. 1b, then the attacker may conclude that this patient suffers from hepatitis even though the attacker is unable to ascertain which record belongs to this individual. This is known as *similarity attack*, meaning that values of confidential attributes may still be semantically similar. More generally, the *skewness attack* exploits the difference between the prior distribution of confidential attributes in the whole data set and the posterior distribution of those attributes within a specific group.

All these vulnerabilities motivated the appearance of a number of proposals, some of which we now overview. An enhancement of k-anonymity called p-sensitive k-anonymity [10] incorporates the additional restriction that there be at least p distinct values for each confidential attribute within each k-anonymous group. With the aim of



Identifier Attribute	Key Attributes		Confidential Attribute
Name	**Age**	**Nationality**	**Health Condition**
William	45	US	Hepatitis B
Emmanuel	42	French	Hepatitis C
Syme	47	Indian	Hepatitis D
Naoto	31	Japanese	Diabetes
Katharine	30	US	Heart Disease
Julia	36	British	Heart Disease

(a) Original data

Key Attributes		Confidential Attribute
Age	**Nationality**	**Health Condition**
40 – 50	*	Hepatitis B
40 – 50	*	Hepatitis C
40 – 50	*	Hepatitis D
< 40	*	Diabetes
< 40	*	Heart Disease
< 40	*	Heart Disease

(b) Perturbed data

Fig. 1 We apply the mechanisms of generalization and suppression to the key attributes "age" and "nationality," respectively, in such a manner that the requirement of 3-anonymity is satisfied. The upshot of such perturbation is that each tuple of key attributes in the released table **b** is shared by at least 3 records. This means that an attacker who knows the key attribute values of a particular respondent cannot ascertain the record of this respondent beyond a subgroup of 3 records in any public database with identifier attributes

addressing the data utility loss incurred by large values of p, l- diversity [11] proposes instead that there be at least l "well-represented" values for each confidential attribute. Unfortunately, both proposals are still vulnerable to similarity attacks and skewness attacks.

In an attempt to overcome all these deficiencies, *t-closeness* [12] was proposed. A microdata set satisfies t-closeness if, for each group of records with the same tuple of perturbed key attribute values, a measure of discrepancy between the posterior and prior distributions does not exceed a threshold t. Inspired by this measure, Rebollo-Monedero et al. [15] define an (*average*) *privacy risk* as the conditional *Kullback- Leibler (KL) divergence* between the posterior and the prior distributions, a measure that may be regarded as an averaged version of t-closeness. Further, this average privacy risk is shown to be equal to the mutual information between the confidential attributes and the observed, perturbed key attributes, and, finally, a connection is established with Shannon's rate-distortion theory. A related criterion named *δ-disclosure* is proposed in [13], a yet stricter version that measures the maximum absolute log ratio between the prior and the posterior distributions. Lastly, Dwork [14] analyzes privacy for interactive databases, where a randomized perturbation rule is applied to a true answer to a query, before returning it to the user. Consider two databases that differ only by one record, but are subject to a common perturbation rule. Conceptually, the randomized perturbation rule is said to satisfy the *ϵ-differential privacy* criterion if the two corresponding probability distributions of the perturbed answers are similar, according to a certain inequality. Later in Sect. 5.2 we provide further details about these privacy criteria and relate them in terms of our formulation.

3 Preliminaries

In this section, we shall present our convention regarding random variables (r.v.'s) and probability distributions. Next, we shall introduce some elementary concepts for those readers who are not familiar with Bayes decision theory (BDT).

Throughout this paper, we shall follow the convention of using uppercase letters to denote r.v.'s, and lowercase letters to the particular values they take on. We shall call *alphabet* the values an r.v. takes on. Probability mass functions (PMFs) are denoted by p, subindexed by the corresponding r.v. Accordingly, $p_X(x)$ denotes the value of the function p_X at x. We use the notations $p_{X|Y}$ and $p_{X|Y}(x|y)$ equivalently. In addition, we shall follow the notation in [35] to specify that two sequences a_k and b_k are approximately equal in the exponent if $\lim_{k\to\infty} \frac{1}{k} \log \frac{a_k}{b_k} = 0$. As an example to illustrate this notation, consider the sequences $a_k = 2^{3k+\sqrt{k}}$ and $b_k = 2^{3k}$ and check that $\lim_{k\to\infty} \frac{1}{k} \log \frac{a_k}{b_k} = \lim_{k\to\infty} \frac{1}{\sqrt{k}} = 0$, what implies that they agree to first order in the exponent. Further, throughout this work we shall denote the uniform distribution by u. Last but not least, we shall use the notation x^n to denote a sequence $x_1, x_2, \ldots, x_n$.

Having adopted these conventions, now we recall the basics on BDT. Namely, BDT is a statistical method that, fundamentally, uses a probabilistic model to analyze the making of decisions under uncertainty and the costs associated with those decisions [36,37]. In general, Bayes decision principles may be formulated in the following terms. Consider the uncertainty refers to an *unknown* parameter modeled by an r.v. X. In decision-theoretic terminology, this is also known as *state of nature*. Let Y be another r.v. modeling an *observation* or measurement on the state of nature. Suppose that, given a particular observation y, we are required to make a decision on the unknown. Let $\hat{x}$ denote the estimator of X, that is, the rule that provides a decision or estimate $\hat{x}(y)$ for every possible observation y. Clearly, any decision will be accompanied by a cost. This is captured by the *loss function* $d\colon (x, \hat{x}) \mapsto d(x, \hat{x})$, which measures how costly the decision $\hat{x} = \hat{x}(y)$ will be when the unknown is x. However, since the actual loss incurred by a decision cannot be calculated with absolute certainty at the time the decision is made, BDT contemplates the average loss associated with this decision. Concretely, the *Bayes conditional risk* for an estimator $\hat{x}$ is defined in the discrete case as

$$\mathcal{R}(y) = E[d(X, \hat{x}(y))|y] = \sum_x p_{X|Y}(x|y)\, d(x, \hat{x}(y)),$$

where the expectation is taken over the *posterior* probability distribution $p_{X|Y}$. According to this, the *Bayes risk* associated with that estimator is defined as the average of the Bayes conditional risk over all possible observations y, that is,

$$\mathcal{R} = EE[d(X, \hat{x}(Y))|Y] = \sum_{x,y} p_{X\,Y}(x, y)\, d(x, \hat{x}(y)),$$

where the expectation is additionally taken over the probability distribution of Y. Based on this definition, an estimator is called *Bayes estimator* or *Bayes decision rule*, if it minimizes the Bayes risk among all possible estimators. It turns out that this optimal estimator is precisely

$$\hat{x}_{\text{Bayes}}(y) = \arg\min_{\hat{x}} E[d(X, \hat{x})|y],$$

for all y, i.e., the Bayes estimator is the one that minimizes the Bayes conditional risk for every observation.

Once some of the basic elements in Bayes analysis have been examined, we would like to establish a connection between maximum a posteriori (MAP) estimator and Bayes estimator. With this aim, first recall that a MAP estimator, as the name implies, is the estimator that maximizes the posterior distribution. Now consider the loss function d to be the Hamming distance between x and $\hat{x}$, which is an indicator function, and recall that the expectation of an indicator r.v. is the probability of the event it is based on. Mathematically,

$$E[d_{\text{Hamming}}(X, \hat{x})|y] = P\{X \neq \hat{x}|y\},$$

and consequently,

$$\begin{aligned}\hat{x}_{\text{MAP}}(y) &= \arg\max_{\hat{x}} P\{X \neq \hat{x}|y\} \\ &= \arg\max_{\hat{x}} P\{X = \hat{x}|y\}. \qquad (1)\end{aligned}$$

In conclusion, Bayes and MAP estimators coincide when the loss function is Hamming distance.

Table 1 Simplified representation of our notation

	Unknown	Input	Decision
Attacker	X	Y	$\hat{X}$
System		X'	Y'

4 Measuring privacy as an attacker's estimation error

This section presents our first contribution, a general framework that lays the foundation for the establishment of a unified measurement of privacy.

However, it is not until Sect. 5 where we shall show that a number of privacy criteria may be regarded as particular cases of our proposal. Previously, Sect. 4.1 introduces our notation. Next, Sect. 4.2 describes the adversarial model. In Sect. 4.3, we present our privacy metric, and finally, in Sect. 4.4, we illustrate the proposed formulation with a simple but insightful example.

4.1 Mathematical assumptions and notation

In this section, we provide the notation that we shall use throughout this work. To this end, we first introduce the key actors of the proposed framework:

- a *user*, who wishes to protect their privacy;
- a (trusted) *system*, to which each user entrusts their private data for its protection; is to guarantee the privacy of the user, the unique purpose of this entity is to guarantee the privacy of the user, and with this aim, the system may use any privacy-preserving mechanism at its disposal;
- and an *attacker*, who strives to disclose private information about this user.

To clarify the elements involved in our framework, consider a conceptually simple approach to anonymous Web browsing, consisting in a TTP acting as an intermediary between Internet users and Web servers. From the perspective of our model, the users would be those subscribed to the anonymous proxy; the system would be this proxy; and the attackers those servers that attempt to compromise users' privacy from their Web browsing activity.

In the following, the term r.v. is used with full generality to include categorical or numerical data, vectors, tuples or sequences of mixed components, but for mathematical simplicity we shall henceforth assume that all r.v.'s in the paper have finite alphabets.

- The *attacker's unknown* or *uncertainty* is denoted by the r.v. X, which models the private information about a user that the attacker wishes to ascertain.

- The *system's input* is represented by the r.v. X' and refers to user's data required by the system to make a decision.
- The *systems's decision* is modeled by the r.v. Y' and denotes disclosed information, perhaps part of X', or a perturbation.
- The *attacker's input* is denoted by the r.v. Y and captures any evidence or measurement the attacker has about the unknown. As its name indicates, this variable models the information that serves as input for the adversary to ascertain X. In some cases, Y may be directly the information revealed by the system, i.e., $Y = Y'$. That is, the only information available to the attacker is exactly that disclosed by the system. In other circumstances, the attacker may observe a perturbed version of Y', maybe together with background knowledge about the unknown. In such cases, we have $Y \neq Y'$. Since the attacker's input is, in fact, the information *observed* by the attacker, directly from the system or indirectly from other sources, throughout this work we shall use the terms *attacker's input* and *attacker's observation* indistinguishably to refer to the variable Y.
- The *attacker's decision* is modeled by the r.v. $\hat{X}$ and represents the attacker's estimate of X from Y.

In order to clarify this notation, we provide an example in which the above variables are put in the context of SDC. In this scenario, the data publisher plays the role of the system. Concretely, X may represent identifying or confidential attribute values the attacker endeavors to ascertain with regard to an individual appearing in a released table. The individuals contained in this table are what we call users. The system's input becomes now the key attribute values that the publisher has about the individuals. On the other hand, Y' is the perturbed version of those values, which jointly with the (unperturbed) confidential attribute values constitute the released table. Furthermore, the attacker's input consists of the released table and, possibly, background knowledge the privacy attacker may have. For example, this could be the case of a voter registration list. In the end, the attacker's decision is the estimate of X. All this information is shown in Table 2.

Similarly, now we specify the variables of our framework in the special case of a mix. Under this scenario, the mix represents the system, whose objective is to hide the correspondence between the incoming and outgoing messages. Precisely, the attacker's uncertainty is this correspondence. The system's input and system's decision are the arrival and departure times of the messages, respectively. On the other hand, the information available to the attacker, i.e., the attacker's observation Y, consists of X', Y' and the design parameters of the mix. Finally, $\hat{X}$ is the attacker's decision on the correspondence between the messages. This is depicted in Fig. 2 and summarized in Table 3.

Table 2 Description of the variables used in our notation in the special case of SDC. Often, $X = X'$ and $Y = Y'$

	Unknown	Input	Decision
Attacker	Identifier or confidential attributes	Perturbed table, possibly with background knowledge	Estimate of identifier or confidential attributes
System		Key attributes	Perturbed key attributes

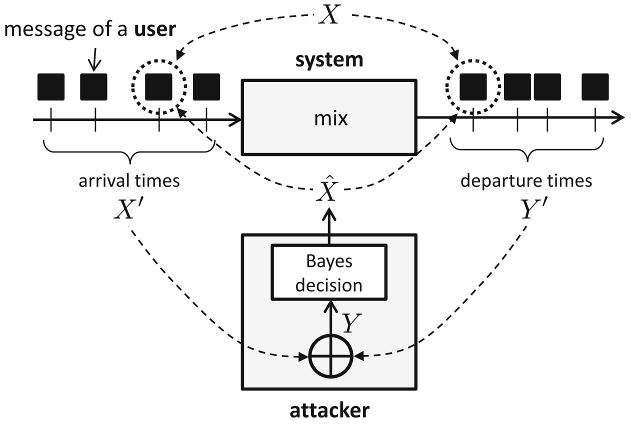


Fig. 2 Our framework is put in the context of mixes

Table 3 Description of the variables used in our notation in the special case of mixes

	Unknown	Input	Decision
Attacker	Correspondence between incoming and outgoing messages	Arrival and departure times of the messages, mix design parameters and maybe background knowledge	Estimate of correspondence between incoming and outgoing messages
System		Arrival times of the messages	departure times of the messages

4.2 Adversarial model

The consideration of a framework that encompasses a variety of privacy criteria necessarily requires the formalization of the attacker's model. In this spirit, we now proceed to present the parameters that characterize this model.

Firstly, we shall contemplate an adversarial model in which the attacker uses a Bayes (best) decision rule. Conceptually, this corresponds to the estimation made by an attacker that uses optimally the available information, as we formally argued in Sect. 3. Namely, for every possible decision of the system resulting in an observation y, the attacker will make a

Bayes decision $\hat{x}(y)$ on X. With regard to this attacker's decision rule, we would like to remark the fact that, whereas it is a deterministic estimator, the system's decision is assumed to be a *randomized* perturbation rule given by $p_{Y'|X'}$. As a consequence of this, it is clear that the system does not leak any private information when deciding Y', provided that Y' and X' are statistically independent.

Secondly, as explained in Sect. 3, we shall require to evaluate the cost of each decision made by the attacker. For this purpose, we consider the *attacker's distortion function* $d_{\mathrm{A}}\colon (x, \hat{x}) \mapsto d_{\mathrm{A}}(x, \hat{x})$, which measures the degree of dissatisfaction that the attacker experiences when $X = x$ and $\hat{X} = \hat{x}(y)$. Similarly, we contemplate the *system's distortion function* $d_{\mathrm{S}}\colon (x', y') \mapsto d_{\mathrm{S}}(x', y')$, which reflects the extent to which the system, and therefore the user, is discontent when $Y' = y'$ and $X' = x'$.

A crucial distinction in the type of attacker's distortion function d_{A} considered will be whether it captures a sort of geometry over the symbols of the alphabet, or not. The most evident example of distortion function that does not take into account this geometry is the Hamming function, which we already introduced at the end of Sect. 3. Concretely, this binary metric just indicates whether x and $\hat{x}$ coincide and provides no more information about the discrepancy between them. On the other hand, the squared error loss $d_{\mathrm{A}}(x, \hat{x}) = (x - \hat{x})^2$ and the absolute error loss $d_{\mathrm{A}}(x, \hat{x}) = |x - \hat{x}|$ are just two commonly used examples of distortion functions that do rely or induce a certain geometry.

4.3 Definition of our privacy criterion

Bearing in mind the above considerations, and consistently with Sect. 3, we define *conditional privacy* as

$$\mathcal{P}(y) = E[d_{\mathrm{A}}(X, \hat{x}(y))|y], \tag{2}$$

which is the estimation error incurred by the attacker, conditioned on the observation y. Based on this definition, we contemplate two possible measures of privacy. In particular, we define *worst-case privacy* as

$$\mathcal{P}_{\min} = \min_{y} \mathcal{P}(y). \tag{3}$$

On the other hand, we define *average privacy* as

$$\mathcal{P}_{\mathrm{avg}} = E\mathcal{P}(Y) = E d_{\mathrm{A}}(X, \hat{x}(Y)), \tag{4}$$

which is the average of the conditional privacy over all possible observations y.

In order to measure the utility loss caused by the perturbation of the original data, we define the *average distortion* as

$$\mathcal{D} = E d_{\mathrm{S}}(X', Y'). \tag{5}$$

According to these definitions, a privacy-protecting system and an attacker would adopt the following strategies. Namely, the system would select the decision rule $p_{Y'|X'}$ that maximizes either the average privacy or the worst-case privacy, while not allowing the average distortion to exceed a certain threshold. On the other hand, the attacker would choose the Bayes estimator, which would lead to the minimization of *both* measures of privacy. The reason behind this is that the Bayes estimator also minimizes the conditional privacy, as stated in Sect. 3.

On a different note, we would like to remark that a privacy risk $\mathcal{R}$ in lieu of $\mathcal{P}$ could be defined for $-d_{\mathrm{A}}(x, \hat{x}(y))$ instead of $d_{\mathrm{A}}(x, \hat{x}(y))$. An analogous argument justifies the use of utility instead of distortion.

Last but not least, we would also like to note that, in the special case when the unknown variable X models the identity of a user, our measure of privacy may be regarded, in fact, as a measure of anonymity.

4.4 Example

Next, we present a simple example that sheds some light on the formulation introduced in the previous sections.

For the sake of simplicity, consider $X' = X$, that is, the system's input is the confidential information that needs to be protected. Suppose that X is a binary r.v. with $P\{X = 0\} = P\{X = 1\} = \frac{1}{2}$. In order to hinder privacy attackers in their efforts to ascertain X, for each possible outcome x, the system will disclose a perturbed version y'. Namely, with probability p the system will decide to reveal the complementary value of x, whereas with probability $1 - p$ no perturbation will be applied, i.e., $y' = x$. Note that, in this example, the system's decision rule is completely determined by p, for which we conveniently impose the condition $0 \leqslant p < \frac{1}{2}$.

At this point, we shall assume that the attacker only has access to the disclosed information Y', and therefore, the attacker's input Y boils down to it. We anticipate that, throughout this work, this supposition will be usual. In addition, we shall consider the attacker's distortion function to be the Hamming distance. However, as commented on in Sect. 3, this implies that the Bayes estimator matches the MAP estimator. According to this observation, it is easy to demonstrate that the attacker's best decision is $\hat{X} = Y$. Therefore, the average privacy (4) becomes

$$\mathcal{P}_{\mathrm{avg}} = P\{X \neq \hat{X}\} = P\{X \neq Y\} = P\{X \neq Y'\} = p.$$

On the other hand, if we suppose that the system's distortion function is also the Hamming distance, from (5), it follows that

$$\mathcal{D} = P\{X' \neq Y'\} = P\{X \neq Y'\} = p.$$

Based on these two results, we now proceed to describe the strategy that the attacker would follow. To this end, we

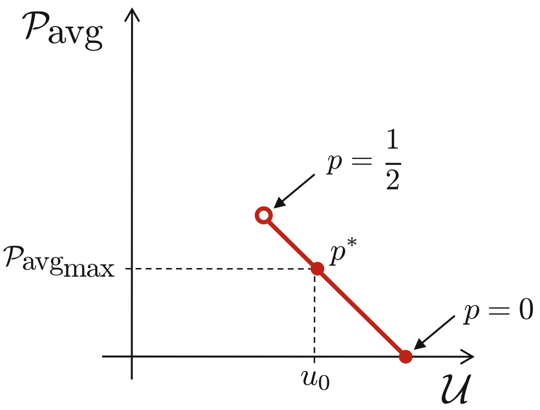


Fig. 3 Representation of the trade-off curve between privacy and utility for the example

define the *average utility* $\mathcal{U}$ as $1 - \mathcal{D}$. According to this, the system would strive to maximize the average privacy with respect to p, subject to the constraint $\mathcal{U} \geq u_0$. Figure 3 illustrates this simple optimization problem by showing the trade-off curve between privacy and utility. In this example, it is straightforward to verify that the optimal value of average privacy is $\mathcal{P}_{\text{avg}_{\max}} = 1 - u_0$, for $\frac{1}{2} < u_0 \leq 1$.

5 Theoretical analysis

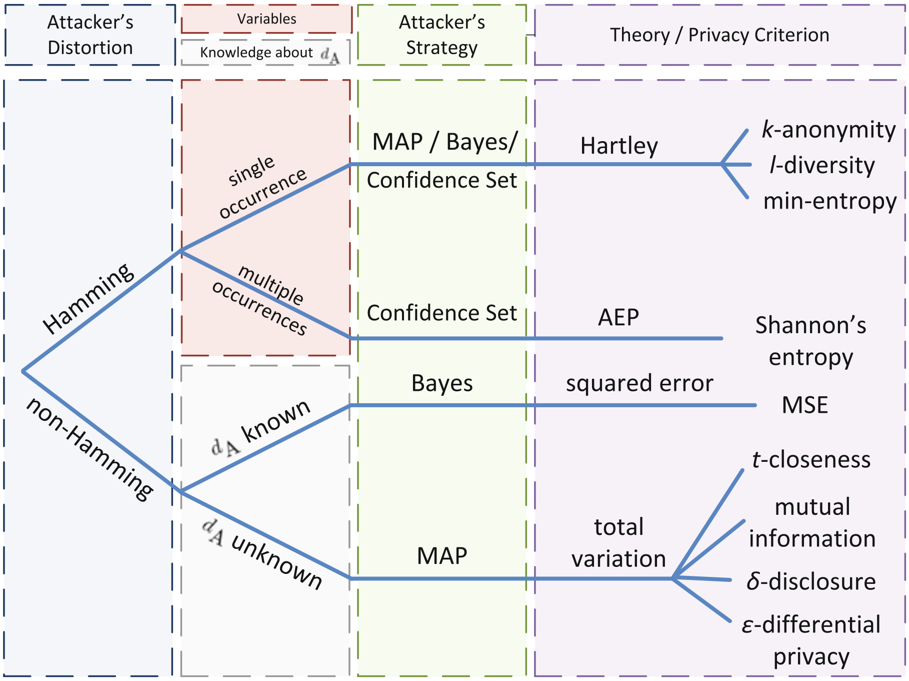
In this section, we present our second contribution, namely, the interpretation of several well-known privacy criteria as particular cases of our more general definition of privacy. The arguments behind the justification of these privacy metrics as a particularization of our criterion are based on numerous concepts from the fields of information theory, probability theory and BDT. In this section, we therefore approach this issue from a theoretical perspective; however, we refer those readers not particularly interested in the mathematical details to Sect. 7.

For a comprehensive exposition of these arguments, the underlying assumptions and concepts will be expounded in a systematic manner, following the points sketched in Fig. 4.

As mentioned in Sect. 4.2 and illustrated by the first branch of the tree depicted in this figure, our starting point makes the significant distinction between attacker's distortion measures based on the Hamming distance and the rest, according to whether we wish to capture a certain, gradual measure of distance between alphabet values beyond sheer symbol equality. It is important to recall from Sect. 3 that in the case of a Hamming distortion measure, expected distortion boils down to probability of error, yielding a different class of estimation problems.

Bearing in mind the above remark, in Sect. 5.1 we shall contemplate the case when the attacker's distortion function is the Hamming distance, whereas in Sect. 5.2 we shall deal with the more general case in which d_{A} can be any other distortion function. In the special case of Hamming distance, we consider two alternatives for the variables in Table 1: single-occurrence and multiple-occurrence data. The former case considers the variables to be tuples of a small number of components, and the latter case assumes that these variables are sequences of data. In the scenario of single-occurrence data, we shall establish a connection between Hartley's entropy and our privacy metric, which will allow us to interpret

Fig. 4 The arguments that lead to the interpretation of several privacy metrics as particular cases of our definition of privacy are conceptually organized in the above points. As can be observed, these arguments clearly depend on the attacker's distortion function, namely on the geometry of this function (Hamming or non-Hamming) and on the knowledge the user has about it, i.e., it is known or unknown to the user. Other parameters include the nature of the variables of our framework, and obviously the attacker's strategy



k-anonymity, l-diversity and min-entropy criteria as particular cases of our framework. The arguments that will enable us to justify this connection stem from MAP estimation, BDT and the concept of confidence set. On the other hand, when we consider multiple- occurrence data, we shall use the asymptotic equipartition property (AEP) to argue that the Shannon entropy, as a measure of privacy, is a characterization of the cardinality of a high-confidence set of sequences.

In the more general case in which the attacker's distortion function is not the Hamming distance, we shall explore two possible scenarios. On the one hand, we shall consider the case where this function is known to the system. Under the assumption of a Bayes attacker's strategy, we shall use BDT to justify the system's best decision rule. On the other hand, we shall contemplate the case in which the attacker's distortion function is unknown to the system. Specifically, this scenario will allow us to connect our framework to several privacy criteria through the concept of total variation, provided that the attacker uses MAP estimation.

5.1 Hamming distortion

In this section, we shall analyze the special case when the attacker's distortion function is the Hamming distance, commented on in Sects. 3 and 4.2. In addition, we shall contemplate two cases for the variables of our framework: single- occurrence and multiple-occurrence data.

5.1.1 Single occurrence

This section considers the scenario in which the variables defined in Sect. 4.1 are tuples of a relatively small number of components, including both categorical and numerical data, defined on a finite alphabet. In order to establish a connection between some of the most popular privacy metrics and our criterion, first we shall introduce the concept of confidence set and briefly recall a riveting generalization of Shannon's entropy.

Consider an r.v. X taking on values in the alphabet $\mathcal{X}$. A *confidence set* $\mathscr{C}$ with confidence p is defined as a subset of $\mathcal{X}$ such that $P\{X \in \mathscr{C}\} = p$. In the case of continuous-valued random scalars, confidence sets commonly take the form of intervals. In these terms, it is clear that a privacy attacker aimed at ascertaining X will benefit the most from those confidence sets whose cardinality is reduced substantially with respect to the original alphabet size, with high confidence. To connect the concept of confidence set to our interpretation of privacy as an attacker's estimation error, consider an attacker model where the attacker only takes into account the shape of the PMF of the unknown X to identify a confidence set $\mathscr{C}$ for some desired confidence p, and beyond that, assumes all the included members equally relevant. This last assumption may be interpreted as an investigation on a tractable list of potential identities, carried out in parallel. MAP estimation within that set, considering it uniformly distributed, leads to an estimation error of $1 - \frac{1}{|\mathscr{C}|}$, that is, a bijection of its cardinality.

In our interpretations, we further use the Rényi entropy, a family of functionals widely used in information theory as a measure of uncertainty. More specifically, Rényi's entropy of order α is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_X(x_i)^\alpha,$$

where p_X is the PMF of an r.v. X that takes on values in the alphabet $\mathcal{X} = \{x_1, \ldots, x_n\}$. In the important case when α is 0, Rényi's entropy is essentially given by the support set of p_X, that is,

$$H_0(X) = \log |\{x \in \mathcal{X} : p_X(x) > 0\}| .$$

In this particular case, Rényi's entropy is referred to as Hartley's entropy. Evidently, when p_X is strictly positive, the support set becomes the alphabet and $H_0(X) = \log n$. Under this assumption, the Hartley entropy can be understood as a confidence set with $p = 100\%$. On the other hand, in the limit when α approaches 1, Rényi's entropy reduces to Shannon's

$$H_1(X) = -\sum_i p_X(x_i) \log p_X(x_i).$$

Lastly, in the limit as α goes to ∞, the Rényi entropy approaches the *min-entropy*

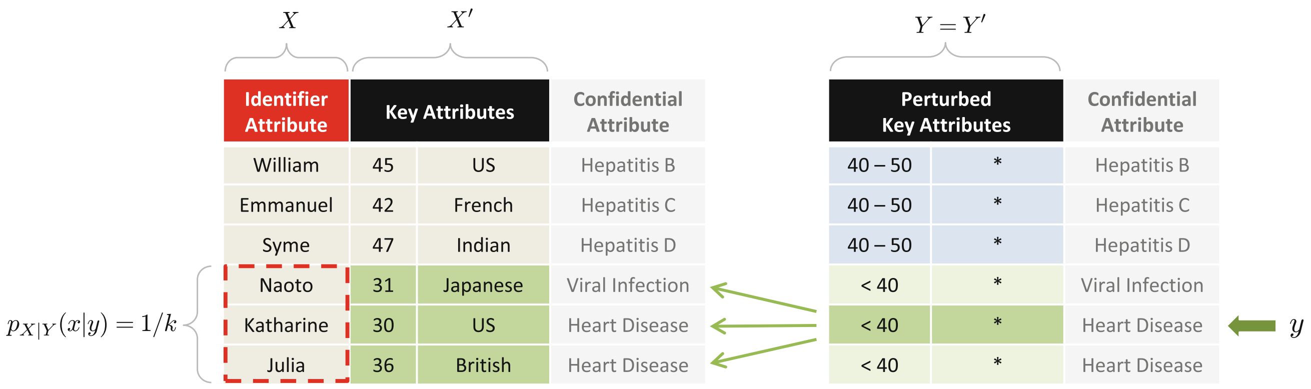
$$H_\infty(X) = \min_i - \log p_X(x_i) = - \log \max_i p_X(x_i).$$

We shall shortly interpret min-entropy, Shannon's entropy, and Hartley's entropy within our general framework of privacy as an attacker estimation error, when Hamming distance is used as a distortion measure, first for single occurrences of a target information, and later for multiple occurrences. For now, we could loosely consider an attacker striving to ascertain the outcome of the finite-alphabet r.v. X and the effect of the dispersion of its PMF on such task. Conceptually, we could then regard these three types of entropies simply as worst-case, average-case, and best-case measurements of privacy, respectively, on account of the fact that

$$H_\infty(X) \leq H_1(X) \leq H_0(X), \tag{6}$$

with equality if, and only if, X is uniformly distributed. More specifically, the min-entropy $H_\infty(X)$ is the minimum of the *surprisal* or *self-information* $-\log p_X(x_i)$, whereas the Shannon entropy $H_1(X)$ is a weighted average of such logarithms, and finally, the Hartley entropy $H_0(X)$ optimistically measures the cardinality of the entire set of possible values of X regardless of their likelihood.

After showing the Hartley, Shannon and min entropies are particular cases of Rényi's entropy, now we go on to



Identifier Attribute	Key Attributes		Confidential Attribute
William	45	US	Hepatitis B
Emmanuel	42	French	Hepatitis C
Syme	47	Indian	Hepatitis D
Naoto	31	Japanese	Viral Infection
Katharine	30	US	Heart Disease
Julia	36	British	Heart Disease

(a) Original data

Perturbed Key Attributes		Confidential Attribute
40 – 50	*	Hepatitis B
40 – 50	*	Hepatitis C
40 – 50	*	Hepatitis D
< 40	*	Viral Infection
< 40	*	Heart Disease
< 40	*	Heart Disease

(b) Perturbed data

Fig. 5 A data publisher plans to release a 3-anonymized microdata set. To this end, the publisher must enforce that, for a given tuple of key attribute values in **b**, the probability of ascertain the identifier value of the corresponding record in **a** must be at most $\frac{1}{3}$

describe a scenario that will allow us to relate our privacy metric to an extensively used criterion. Specifically, we focus on the important case of SDC, where the data publisher plays the system's role. In this scenario, a data publisher wishes to release a microdata set and, before distributing it, the publisher applies some algorithm [10–15] to enforce the k-anonymity requirement [8,9]. As mentioned in Sect. 2.2, the objective of a linking attack is to unveil the identity of the individuals appearing in a released table by linking the records in this table to any public data set including identifiers. Since k-anonymity is aimed at protecting the data against this attack, in our scenario the attacker's unknown X becomes the user identity. The other variables shown in Table 2 are as follows: X' are the key attribute values, Y' are the perturbed key attribute values, the attacker's observation Y is assumed to be Y', and finally, $\hat{X}$ is an estimate of the identity of a user. Figure 5 illustrates this particular case.

In order to protect the data set from identity disclosure, the algorithm must ensure that, for any observation y consisting in a tuple of perturbed key attribute values in the released table, the identifier value of the corresponding record in the original table cannot be ascertained beyond a subgroup of at least k records. As we shall see next, this requirement will be reflected mathematically by assuming that the probability distribution $p_{X|Y}(\cdot|y)$ of the identifier value, conditioned on the observation y, is the uniform distribution on a set of at least k individuals. Lastly, we consider the more general case in which Y consists of Y' and any background knowledge.

That said, our adversarial model contemplates an attacker who uses a MAP estimator, which, as shown in Sect. 3, is equivalent to the Bayes estimator. Under this model, given an observation y, the conditional privacy (2) becomes

$$\mathcal{P}(y) = P\{X \neq \hat{x}(y)|y\} = 1 - \max_x p_{X|Y}(x|y), \tag{7}$$

which precisely is the MAP error ε_{MAP}, conditioned on that observation y; in terms of min-entropy, we may recast our metric as

$$\mathcal{P}(y) = \varepsilon_{\text{MAP}} = 1 - 2^{-H_\infty(X|y)},$$

which shows that the concept of min-entropy is intimately related to MAP decoding. If we finally apply the aforementioned uniformity condition of $p_{X|Y}(\cdot|y)$ and assume that this PMF is the uniform distribution on a group of exactly k individuals, that is, $u_i = 1/k$ for all $i = 1, \dots, k$, then

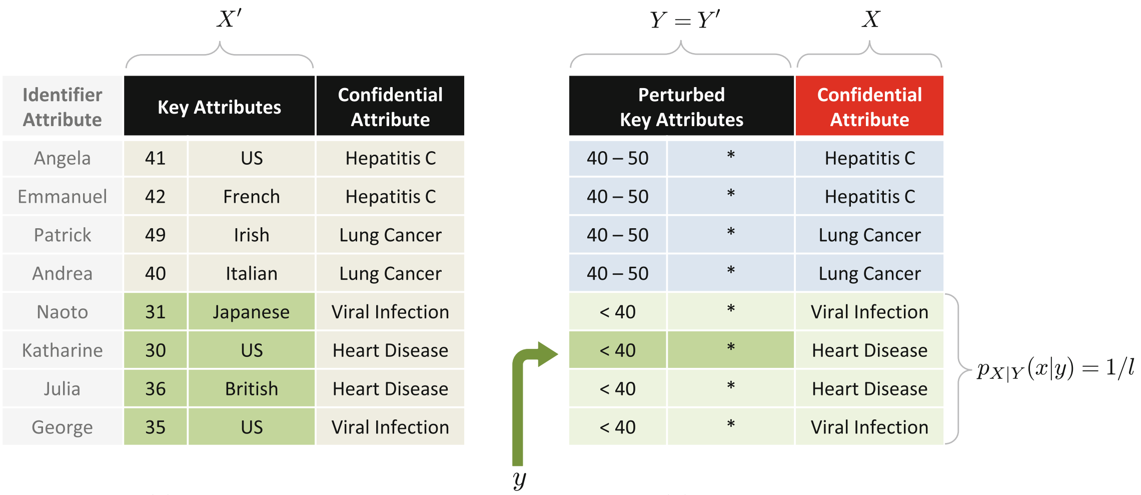
$$\mathcal{P}(y) = 1 - 1/k = 1 - 2^{-H_0(X|y)},$$

which expresses the conditional privacy in terms of Hartley's entropy. In a nutshell, the k-anonymity criterion may be interpreted as a special case of our privacy measure, determined by this Rényi's entropy.

After examining this first interpretation, next we shall explore an enhancement of k-anonymity. As argued in Sect. 3, this criterion does not protect against confidential attribute disclosure. In an effort to address this limitation, several privacy metrics were proposed. In the remainder of this section, we shall focus on one of these approaches. In particular, we shall consider the l-diversity metric [11], which builds on the k- anonymity principle and aims at overcoming the attribute disclosure problem.

As mentioned in Sect. 2.2, a microdata set satisfies l-diversity if, for each group of records sharing a tuple of key attribute values in the perturbed table, there are at least l "well-represented" values for each confidential attribute. Depending on the definition of well-represented, this criterion can reduce to distinct l-diversity, which is equivalent to l-sensitive k-anonymity, or be more restrictive. Concretely, a microdata is said to meet the entropy l-diversity requirement if, for each group of records with the same tuple of perturbed key attribute values, the entropy of each confidential attribute is at least $\log l$.

In our new scenario, a data publisher, still playing the system's role, applies an algorithm on the microdata set to


X'
$Y = Y'$
X
y
$p_{X|Y}(x|y) = 1/l$

Identifier Attribute	Key Attributes		Confidential Attribute
Angela	41	US	Hepatitis C
Emmanuel	42	French	Hepatitis C
Patrick	49	Irish	Lung Cancer
Andrea	40	Italian	Lung Cancer
Naoto	31	Japanese	Viral Infection
Katharine	30	US	Heart Disease
Julia	36	British	Heart Disease
George	35	US	Viral Infection

(a) Original data

Perturbed Key Attributes		Confidential Attribute
40 – 50	*	Hepatitis C
40 – 50	*	Hepatitis C
40 – 50	*	Lung Cancer
40 – 50	*	Lung Cancer
< 40	*	Viral Infection
< 40	*	Heart Disease
< 40	*	Heart Disease
< 40	*	Viral Infection

(b) Perturbed data

Fig. 6 In this example, the 2-diversity principle is applied to a microdata set. In order to meet this requirement, we assume that, for each group of records with the same tuple of perturbed key attribute values, the probability distribution of the confidential attribute value in **b** is the uniform distribution on a set of at least 2 values

enforce the l- diversity principle. Since the aim of this criterion is to protect the data against attribute disclosure, we consider that the attacker's unknown X refers to the confidential attribute. The other variables remain the same as in our previous interpretation.

Having said that, we shall make the assumption that the l- diversity requirement is met by enforcing that, for a given tuple y of perturbed key attribute values, the probability distribution $p_{X|Y}(\cdot|y)$ of the confidential attribute within the group of records sharing this tuple is the uniform distribution on a set of at least l values. This is depicted in Fig. 6. Note that this assumption entails that the data fulfill both the distinct and entropy l-diversity requirements. Lastly, we shall suppose again that the attacker uses MAP estimator.

As mentioned before, under the premise of a MAP attacker, our measure of conditional privacy boils down to the MAP error (7). If we also apply the assumption above about the uniformity of $p_{X|Y}(\cdot|y)$ and suppose that this distribution is uniform on a group of l individuals, then the conditional privacy yields

$$\mathcal{P}(y) = 1 - 1/l = 1 - 2^{-H_0(X|y)},$$

which expresses our privacy metric again in terms of Hartley's entropy. In short, the l-diversity criterion lends itself to be interpreted as a particular case of our more general privacy measure.

5.1.2 *Multiple occurrences*

In this section, we shall consider the case when the variables shown in Table 1 are sequences of categorical and numerical data but in a finite alphabet. Recall from Sect. 3 that we use the notation X^k to denote a sequence $X_1, \ldots, X_k$.

The special case that we contemplate now could perfectly model the scenario in which a user interacts with an LBS provider, through an intermediate system protecting the user's location privacy. In this scenario, a user would submit queries along with their locations to the trusted system. An example would be the query "Where is the nearest parking garage?", accompanied by the geographic coordinates of the user's current location. As many approaches suggest in the literature of private LBSs, the system would perturb the user coordinates and submit them to the LBS provider. Concordantly, we may choose Euclidean distance as the natural attacker's distortion measure. Alternatively, if the attacker's interest lies in whether the user is at home, at work, shopping for groceries or at the movies, in order to profile their behavior, or more simply, whether the user is at a given sensitive location or not, then the appropriate model for the location space becomes discrete, and Hamming distance is more suited.

In this context, the consideration of sequences of discrete r.v.'s in our notation makes sense. Specifically, an attacker would endeavor to ascertain the sequence X^k of k unknown locations visited by the user, from the sequence Y'^k of k perturbed locations that the system would submit to the LBS. Put differently, the attacker's unknown would be the location data the user conveys to the system, i.e., $X^k = X'^k$, and the information available to the adversary the perturbed version of this data, that is, $Y^k = Y'^k$.

Having motivated the case of sequences of data, in this section, we shall establish a connection between our metric and Shannon's entropy as a measure of privacy. But in order to emphasize this connection, first, we briefly recall one of the pillars of information theory: the asymptotic equipartition property [35], which derives from the weak law of large numbers and results in important consequences in this field.

Consider a sequence X^k of k independent, identically distributed (i.i.d.) r.v.'s, drawn according to p_X, with alphabet size n. Loosely speaking, the AEP states that among all possible n^k sequences, there exists a *typical subset* $\mathscr{T}_\epsilon^k$ of sequences almost certain to occur. More precisely, for any $\epsilon > 0$, there exists a k sufficiently large such that $P\{\mathscr{T}_\epsilon^k\} > 1 - \epsilon$, and $|\mathscr{T}_\epsilon^k| \leq 2^{k(H_1(X)+\epsilon)}$. A similar argument called joint AEP [35] also holds for the i.i.d. sequences (X^k, Y^k) of length k drawn according to $\prod_{i=1}^k p_{X\,Y}(x_i, y_i)$. Another information-theoretic result is related to those sequences x^k that are jointly typical with a given typical sequence y^k. Namely, the set of all these sequences x^k is referred to as the *conditionally typical set* $\mathscr{T}_\epsilon^{X^k|y^k}$ and satisfies, on the one hand, that $P\{\mathscr{T}_\epsilon^{X^k|y^k}\} > 1 - \epsilon$ for large k and, on the other, that its cardinality is bounded by Shannon's conditional entropy, $|\mathscr{T}_\epsilon^k| \leq 2^{k(H_1(X|Y)+\epsilon)}$. Further, it turns out that these conditionally typical sequences are equally likely, with probability $2^{-kH_1(X|Y)}$, approximately in the exponent. While the most likely sequence may in fact *not* belong to the typical set, the set of typical sequences encompasses a sufficiently large number of sequences that amount to a probability arbitrarily close to certainty.

Next, we proceed to interpret, under the perspective of our framework, the Shannon entropy as a measure of privacy. To this end, consider the scenario in which a privacy attacker observes a typical Y^k and strives to estimate the unknown X^k. Conveniently, we assume $X^k = X'^k$ and $Y^k = Y'^k$, which models the LBS example described before, provided that the attacker ignores any spatial–temporal constraint. In other words, we model a scenario without memory and hence suppose that (X_i, Y_i) are i.i.d. drawn according to $p_{X\,Y}$. We would like to stress that the consideration of this simplified model is just for the purpose of providing a simple, clear example that illustrates the application of our framework. Having said this, in the terms above we may regard $\mathscr{T}_\epsilon^{X^k|y^k}$ as a set of arbitrarily high confidence with cardinality $2^{kH_1(X|Y)}$, approximately in the exponent.

The upshot is that the Shannon (conditional) entropy of an unknown r.v. (given an observed r.v.) is an approximate measure of the size of a high-confidence set, measure suitable for attacker models based on the estimation of sequences, rather than individual samples. Moreover, within this confidence set, sequences are equally likely, approximately in the exponent, concordantly with the interpretation of confidence-set cardinality as a measure of privacy made in Sect. 5.1.1 on single occurrences. Even though for simplicity our argument focused on memoryless sequences, the Shannon–McMillan–Breiman theorem is a generalization of the AEP to stationary ergodic sequences, in terms of entropy rates [38].

Finally, we mentioned that the most likely sequence may in fact be atypical, and thus, Shannon entropy is not directly applicable to MAP estimation over the entire set of sequences. Nevertheless, because the most likely memoryless sequence is simply a repetition of the most likely symbol, MAP estimation on sequences is a trivial extension of the argument on min-entropy presented in Sect. 5.1.1.

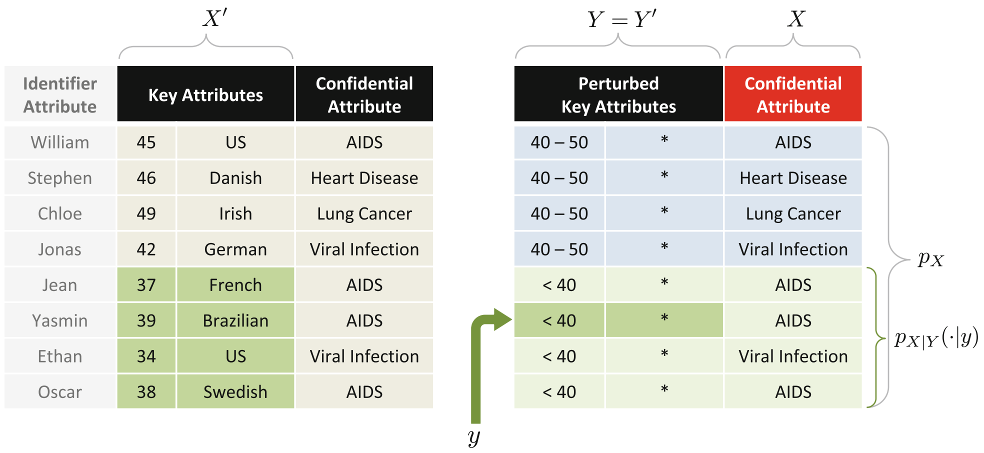
5.2 Non-hamming distortion

This section investigates the complementary case described in Sect. 5 in which the attacker's distortion function is not the Hamming distance. Particularly, in this section, we turn our attention to the scenario of SDC and contemplate two possible alternatives regarding the system's knowledge on the function d_{A}—first, when this function is known to the data publisher, and secondly, when it is unknown. Under the former assumption, the system would definitely use BDT to find the decision rule $p_{Y'|X'}$ which maximizes either the worst-case privacy (3) or the average privacy (4) and satisfies a constraint on average distortion. The latter assumption, however, describes a more general and realistic scenario. The remainder of this subsection precisely interprets several privacy criteria under this assumption. The only piece of information which is though known to the publisher is $d_{\max} = \max_{x,\hat{x}} d_{\mathrm{A}}(x, \hat{x})$, that is, the *maximum value* attained by said function.

Bearing in mind the above consideration, in our new scenario a privacy attacker endeavors to guess the confidential attribute value of a particular respondent in the released table. Initially, the attacker has a prior belief given by p_X, that is, the distribution of that confidential attribute value in the whole table. Later, the attacker observes that the user belongs to a group of records sharing a tuple of perturbed key attribute values y, which is supposed to coincide with the system's decision y'. Based on this observation, the attacker updates their prior belief and obtains the posterior distribution $p_{X|Y}(\cdot|y)$. This situation is illustrated in Fig. 7. A fundamental question that arises in this context is how much privacy the released table leaks as a result of that observation. In the remainder of this section, we elaborate on this question and provide an upper bound on the reduction in privacy incurred by the disclosure of that information.

5.2.1 Total variation and t-closeness

For notational simplicity, we occasionally rename the posterior and the prior distributions $p_{X|Y}(\cdot|y)$ and p_X simply with the symbols p and q, respectively, but bear in mind that p is a PMF of x parametrized by y. In addition, we shall assume that the attacker adopts a MAP strategy. More


X'
$Y = Y'$
X
p_X
$p_{X|Y}(\cdot|y)$
y

Identifier Attribute	Key Attributes		Confidential Attribute
William	45	US	AIDS
Stephen	46	Danish	Heart Disease
Chloe	49	Irish	Lung Cancer
Jonas	42	German	Viral Infection
Jean	37	French	AIDS
Yasmin	39	Brazilian	AIDS
Ethan	34	US	Viral Infection
Oscar	38	Swedish	AIDS

(a) Original data

Perturbed Key Attributes		Confidential Attribute
40 – 50	*	AIDS
40 – 50	*	Heart Disease
40 – 50	*	Lung Cancer
40 – 50	*	Viral Infection
< 40	*	AIDS
< 40	*	AIDS
< 40	*	Viral Infection
< 40	*	AIDS

(b) Perturbed data

Fig. 7 At first, an attacker believes that the probability that a user appearing in **b** suffer from AIDS is 1/2. However, after observing that the user's record is one of the last four records, this probability becomes 3/4

precisely, $\hat{x}_p$ and $\hat{x}_q$ will denote the attacker's estimate when using the distributions p and q. Under these assumptions, the *reduction* (prior minus posterior) in conditional privacy can be expressed as

$$\begin{aligned}\Delta\mathcal{P}(y) &= E_p d_{\mathrm{A}}(X,\hat{x}_q) - E_p d_{\mathrm{A}}(X,\hat{x}_p)\\ &= E_p d_{\mathrm{A}}(X,\hat{x}_q) - E_q d_{\mathrm{A}}(X,\hat{x}_q) + E_q d_{\mathrm{A}}(X,\hat{x}_q)\\ &\qquad - E_q d_{\mathrm{A}}(X,\hat{x}_p) + E_q d_{\mathrm{A}}(X,\hat{x}_p) - E_p d_{\mathrm{A}}(X,\hat{x}_p),\end{aligned}$$

where E_p and E_q denotes that the expectation is taken over the posterior and the prior distributions, respectively, as PMFs of x.

In this expression, the first two terms can be upper bounded by $d_{\max}\sum_x |p_x - q_x|$, since $\sum_x (p_x - q_x) \leqslant \sum_x |p_x - q_x|$. Clearly, this same bound applies to the last two terms. On the other hand, the remaining terms $E_q d_{\mathrm{A}}(X,\hat{x}_q) - E_q d_{\mathrm{A}}(X,\hat{x}_p)$ are upper bounded by 0, since the error incurred by $\hat{x}_q$ is smaller than or equal to that of $\hat{x}_p$. In the end, we obtain that

$$\Delta\mathcal{P}(y) \leqslant 2\, d_{\max} \sum_x |p_x - q_x|\,.$$

At this point, we shall briefly review the concept of *total variation*. For this purpose, consider P and Q to be two PMFs over $\mathcal{X}$. In probability theory, the total variation distance between P and Q is

$$\mathrm{TV}(P \,\|\, Q) = \frac{1}{2} \sum_{x\in\mathcal{X}} |P(x) - Q(x)|\,.$$

Furthermore, recall that, in information theory, *Pinsker's inequality* relates the total variation distance with the KL divergence. Particularly, $\mathrm{TV}(P \,\|\, Q) \leqslant \frac{\sqrt{2}}{2}\sqrt{\mathrm{D}(P \,\|\, Q)}$. Having stated this result, now the total variation distance permits writing the upper bound on $\Delta\mathcal{P}(y)$ in terms of the KL divergence:

$$\Delta\mathcal{P}(y) \leqslant 4\, d_{\max}\mathrm{TV}(p \,\|\, q) \leqslant 2\sqrt{2}\, d_{\max} \sqrt{D(p \,\|\, q)},$$

where the last inequality follows from Pinsker's inequality. Returning to the notation of prior and posterior distributions,

$$\begin{aligned}\Delta\mathcal{P}(y) &\leqslant 4\, d_{\max}\mathrm{TV}(p_{X|Y}(\cdot|y) \,\|\, p_X)\\ &\leqslant 2\sqrt{2}\, d_{\max} \sqrt{D(p_{X|Y}(\cdot|y) \,\|\, p_X)}.\end{aligned} \tag{8}$$

This upper bound allows to establish a connection between our privacy criterion and t-closeness [12]. The latter criterion boils down to defining a maximum discrepancy between the posterior and prior distributions,

$$t = \max_y \mathrm{D}(p_{X|Y}(\cdot|y) \,\|\, p_X).$$

Under this definition and on account of (8),

$$\Delta\mathcal{P}(y) \leqslant 2\sqrt{2}\, d_{\max} \sqrt{t}.$$

Therefore, t-closeness is essentially equivalent to bounding the decrease in conditional privacy.

On a different note, we would like to make a comment on an issue of a purely technical nature. Clearly, in light of inequality (8), the minimization of either the total variation distance or the KL divergence leads to the minimization of an upper bound on $\Delta\mathcal{P}(y)$. However, the fact that the KL divergence imposes a worse upper bound suggests us considering it when the resulting mathematical model be more tractable than the one built upon the total variation distance.

5.2.2 Mutual information and rate-distortion theory

The privacy criterion proposed in [15], called *(average) privacy risk* $\mathcal{R}$, is the average-case version of t-closeness. Formally, $\mathcal{R}$ is a conditional KL divergence, the average discrepancy between the posterior and the prior distributions, which turns out to coincide with the mutual information between the confidential data X and the observation Y:

$$\begin{aligned}\mathcal{R} &= E_Y\, \mathrm{D}(p_{X|Y}(\cdot|Y) \,\|\, p_X) \\ &= E_Y E_{X|Y}\left[\log \frac{p_{X|Y}(X|Y)}{p_X(X)} \middle| Y\right] \\ &= E \log \frac{p_{X|Y}(X|Y)}{p_X(X)} = \mathrm{I}(X;Y).\end{aligned}$$

Directly from their definition, $\mathcal{R} \leqslant t$, meaning that t- closeness is a stricter measure of privacy risk. Because the KL divergence is itself an average, $\mathcal{R}$ is clearly an average-case privacy criterion, but t-closeness is technically a maximum of an expectation, a hybrid between average case and worst case. The next subsection will comment on a third, purely worst-case criterion. When choosing a privacy criterion, it is important to keep in mind that optimizing a privacy mechanism for the best worst-case scenario will in general yield a worse average case, and viceversa.

Further, we conveniently rewrite inequality (8) as

$$\frac{1}{8\, d_{\max}^2} \Delta\mathcal{P}(y)^2 \leqslant \mathrm{D}(p_{X|Y}(\cdot|y) \,\|\, p_X).$$

By averaging over all possible observation y, the right-hand side of this inequality becomes the privacy risk $\mathcal{R}$, which we showed to be equal to the mutual information. This leads to a bound on the privacy reduction in terms of mutual information,

$$\frac{1}{8\, d_{\max}^2} E[\Delta\mathcal{P}(Y)^2] \leqslant \mathrm{I}(X;Y).$$

Based on this observation, it is clear that the minimization of the mutual information contributes to the minimization of an upper bound on $\Delta\mathcal{P}(y)$. With this in mind, we now consider the more general scenario in which Y' and Y need not necessarily coincide and contemplate the case of a data publisher. Concretely, from the perspective of a publisher, we would choose a randomized perturbation rule $p_{Y'|X'}$ with the aim of minimizing the mutual information between X and Y and consequently protecting user privacy. Evidently, the publisher would also need to guarantee the utility of the data to a certain extent and thus impose a constraint on the average distortion. In conclusion, the data publisher would strive to solve the optimization problem

$$\min_{\substack{p_{Y'|X'} \\ E d_{\mathrm{U}}(X',Y') \leqslant \mathcal{D}}} \mathrm{I}(X;Y), \qquad (9)$$

which surprisingly bears a strong resemblance with the rate-distortion problem in the field of information theory.

More specifically, the above optimization problem is a generalization of a well-known, extensively studied information- theoretic problem with more than half a century of maturity. Namely, the problem of lossy compression of source data with a distortion criterion, first proposed by Shannon in 1959 [39].

The importance of this lies in the fact that some of the information-theoretic results and methods for the rate-distortion problem can be extended to the problem (9). For example, in the special case when $X = X'$ and $Y = Y'$, our more general problem boils down to Shannon's rate-distortion and, interestingly, can be computed with the Blahut-Arimoto algorithm [35].

Bear in mind that the very same metric, or conceptually equivalent variations thereof, may in fact be interpreted under different perspectives. Recall, for instance, that mutual information is the difference between an unconditional entropy and a conditional entropy, effectively the posterior uncertainty modeled simply by the Shannon entropy, normalized with respect to its prior correspondence. Under this perspective, mutual information might also be connected to the branch of the tree in Fig. 4 leading to Shannon's entropy.

5.2.3 δ-Disclosure and differential privacy

Finally, we quickly remark on the connection of δ-disclosure and ϵ-differential privacy with our theoretical framework. *δ-disclosure* [13] is an even stricter privacy criterion than t-closeness, and hence much stricter than that average privacy risk $\mathcal{R}$ or mutual information, discussed in the previous subsection. The definition of δ-disclosure may be rewritten in terms of our notation as

$$\delta = \max_{x,y} \left| \log \frac{p_{X|Y}(x|y)}{p_X(x)} \right|,$$

and understood as a worst-case privacy criterion. In fact,

$$\mathcal{R} \leqslant t \leqslant \delta.$$

We mentioned in the background section that [14] analyzes the case of the randomized perturbation Y of a true answer X to a query in a private information retrieval system, before returning it to the user. Consider two databases d and d' that differ only by one record, but are subject to a common perturbation rule $p_{Y|X}$, and let p_Y and p'_Y be the two probability distributions of perturbed answers induced. After a slight manipulation of the definition given in the work cited, but faithfully to its spirit, we may say that a randomized perturbation rule provides ϵ-differential privacy when

$$\epsilon = \max_{y,d,d'} \log \frac{p_Y(y)}{p'_Y(y)}.$$

Even though it is clear that this formulation does not quite match the problem in terms of prior and posterior distributions described thus far, this manipulation enables us to still establish a loose relation with δ-disclosure, in the sense that the latter privacy criterion is a slightly stricter measure of discrepancy between PMFs, also based on a maximum (absolute) log ratio. We note, however, that although there is a formal similarity between the metrics, there are substantial differences between them in terms of their assumptions, objectives, models, and privacy guarantees.

6 Numerical example

This section provides two simple albeit insightful examples that illustrate the measurement of privacy as an attacker's estimation error. Specifically, we quantify the level of privacy provided, first, by a privacy-enhancing mechanism that perturbs location information in the scenario of LBS, and secondly, by an anonymous communication protocol largely based on Crowds [21].

6.1 Data perturbation in location-based services

Our first example contemplates a user who wishes to access an LBS provider. For instance, this could be the case of a user who wants to find the closest Italian restaurant to their current location. For this purpose, the user would inevitably have to submit their GPS coordinates to the (untrusted) provider. To avoid revealing their exact location, however, the user itself could perturb their location information by adding, for example, Gaussian noise. Alternatively, we could consider a user delegating this task to a (trusted) intermediary entity, as described in Sect. 5.1.2. In any case, data perturbation would enhance user privacy in terms of location, although clearly at the cost of data utility. Simply put, perturbative privacy methods present the inherent trade-off between data utility and privacy.

Under the former strategy, and in accordance with the notation defined in Sect. 4.1, the user becomes the system—it is the user who is responsible for protecting their location data. Playing the role of the system, the user decides then to perturb their location data X on an individual basis for each query. In other words, we do not contemplate the case of sequences of data X^k, as Sect. 5.1.2 does.

A key element of our framework is the attacker's distortion function. In our example we assume the squared error between the actual location x and the attacker's estimate $\hat{x}$, that is, $d_{\mathrm{A}}(x, \hat{x}) = \|x - \hat{x}\|^2$. Unlike Hamming distance, note that the squared error does quantify how much the estimate differs from the unknown. As for the other variables of our model, we contemplate that the attacker's input Y is directly the location data perturbed by the user, Y', as illustrated in Fig. 8. Put differently, the attacker, assumed to be the service provider, has no more information than that disclosed by the user. Under all these assumptions, the average privacy (4) is

$$\mathcal{P}_{\mathrm{avg}} = E[\|X - \hat{X}\|^2],$$

that is, the mean squared error (MSE).

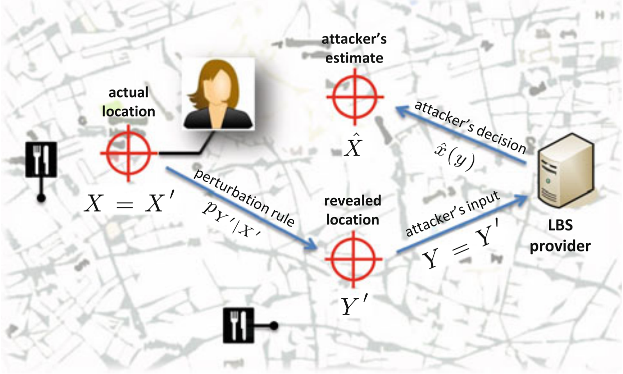


Fig. 8 A user looking for a nearby Italian restaurant accesses an LBS provider. The user decides to perturb their actual location before querying the provider. In doing so, the user hinders the provider itself and any attacker capable of capturing their query, in their efforts to compromise user privacy in terms of location. In this example, we contemplate that the user is solely responsible for protecting their private data. In terms of our notation, this allows us to regard the user as the system. Notice that the user's actual location is, on the one hand, the attacker's unknown and, on the other hand, the information that the user (system) takes as input to generate the location that will be finally revealed. Thus, we conclude that $X = X'$. Then, according to some randomized perturbation rule $p_{Y'|X'}$, the user discloses, for each location data x', a perturbed version y'. This perturbed location is submitted to the provider, which only has access to this information, i.e., $Y = Y'$. Lastly, based on this revealed information, the attacker uses a Bayes estimator $\hat{x}(y)$ to ascertain the user's actual location X

As a final remark, we would like to connect our privacy criterion with a metric specifically conceived for the LBS scenario at hand [34]. In this cited work, the authors propose a framework that contemplates different aspects of the adversarial model, captured by means of what they call *certainty*, *accuracy* and *correctness*. The information to be protected by a trusted intermediary system are traces modeling the locations visited by users over a period of time. The system accomplishes this task by hiding certain locations, reducing the accuracy of such locations or adding noise. As a result, the attacker observes a perturbed version of the traces and, together with certain mobility profiles of these users, attempts to deduce some information of interest X about the actual traces. In terms of our notation, the observed trajectories and the mobility patterns constitute the attacker's observation Y.

More accurately, given a particular observation y, the attacker strives to calculate the posterior distribution $p_{X|Y}$. However, since the adversary may have a limited number of resources, they may have to content themselves with an estimate $\hat{p}_{X|Y}$. The authors then use Shannon's entropy to measure the *uncertainty* of X and define *accuracy* as the discrepancy between $p_{X|Y}$ and $\hat{p}_{X|Y}$. Finally, they refer to location privacy as *correctness* and measure it as

$$\mathrm{E}_{\hat{p}_{X|Y}}[d_{\mathrm{S}}(X, x_t)|y],$$

where x_t is the true outcome of X, d_S a distance function specified by the system, and the expectation is taken over the estimate of the posterior distribution.

The most notable difference between [34] and our own work is that the authors limit the scope of their metric to the specific scenario of location-based services, whereas here we attempt to provide a general overview. Besides, their proposal is a measure of privacy in an average-case sense. Another important distinction between the cited work and ours is that the former arrives to the conclusion that entropy and k-anonymity are not appropriate metrics for quantifying privacy in the context of LBS. Our work, however, does *not* argue against the use of entropy, k-anonymity and any of the other privacy metrics examined in Sect. 5. In fact, we regard these metrics as particular cases of the attacker's estimation error under certain assumptions on the adversarial model, the attacker's strategy and a number of different considerations explored in that section. Lastly, their implementation of estimation strategies using the forward–backward [40] and the Metropolis-Hastings [41] algorithms are undoubtedly of great interest, but the focus of the present work is on metrics.

6.2 Crowds-like protocol for anonymous communications

In Sect. 2.1 we mentioned Chaum's mixes as a building block to implement anonymous communications networks. A different approach to communication anonymity is based on collaborative, peer-to-peer architectures. An example of collaborative approach is Crowds [21], in which users form a "crowd" to provide anonymity for each other.

In Crowds, a user who wants to browse a Web site forwards the request to another member of his crowd chosen uniformly at random. This crowd member decides with probability p to send the request to the Web site, and with probability $1 - p$ to send it to another randomly chosen crowd member, who in turn repeats the process. For the purposes of illustration, we consider a variation of the Crowds protocol. The main difference with respect to the original Crowds is that we do not introduce a mandatory initial forwarding step. We note that this variation provides worse anonymity than the original protocol, while also reducing the cost (in terms of delay and bandwidth) with respect to Crowds. Further, we assume that the users participating in the protocol are honest; i.e., we only consider the Web site receiving the request as possible adversary.

More formally, consider n users indexed by $i = 1, \ldots, n$, wishing to communicate with an untrusted server. In order to attain a certain degree of anonymity, each user submits the message directly to said server with probability $p \in (0, 1)$ and forwards it to any of the other users, including themselves, with probability $1 - p$. In the case of forwarding, the recipient performs exactly the same probabilistic decision until the message arrives at the server. Figure 9 shows the operation of this protocol.

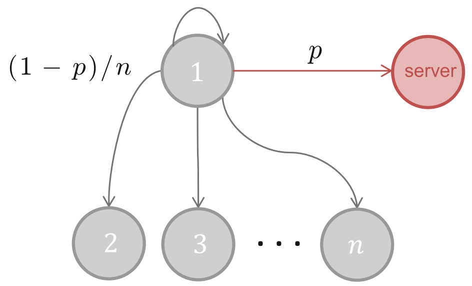


Fig. 9 Anonymous communication protocol inspired by Crowds. In our second numerical example, we contemplate a scenario where users send messages to a common, untrusted server, who aims at compromising sender anonymity. In response to this privacy threat, users decide to adhere to a modification of the Crowds protocol, whose operation is as follows: each user flips a biased coin and depending on the outcome chooses to submit the message to the server or else to another user, who is asked to perform the same process. The probability that a user forward the message to the server is denoted by p, whereas the probability of sending it to any other peer, including themselves, is $(1 - p)/n$

In our protocol, we assume that the server attempts to guess the identity of the author of a given message, represented by the r.v. X, knowing only the user who last forwarded it, represented by the r.v. Y, consistently with the notation defined in Sect. 4.1. The other variables of our framework are as follows. Since the set of users involved in the protocol collaborate to frustrate the efforts of the server, they are in fact the system. The information that then serves as input to this system is simply the identity of the user who initiates the forwarding protocol, X. That is, the attacker's uncertainty and the system's input coincide, $X' = X$. Then again, the assumption that the server just knows the last sender in the forwarding chain leads to $Y = Y'$.

Under this model, and under the assumption of a uniform message-generation rate, that is, $p_X(x) = 1/n$ for all x, it can be proven that the conditional PMF of X given $Y = y$ is

$$p_{X|Y}(x|y) = \begin{cases} p + (1-p)/n, & x = y \\ (1-p)/n, & x \neq y \end{cases}. \qquad (10)$$

Figure 10 shows this conditional probability in the particular case when $x = 1$, i.e., the probability that the originator of a message be user 1, conditioned to the observation that the last sender is user y. Note that, because of the symmetry of our model, it would be straightforward to derive a PMF analogous to the one plotted in this figure, but for other originators of the message, namely $x = 2, \ldots, n$.

That said, assume that the attacker chooses Hamming distance as distortion function. Under this assumption, the conditional privacy (2) yields

$$\mathcal{P}(y) = P\{X \neq \hat{x}(y)|y\},$$

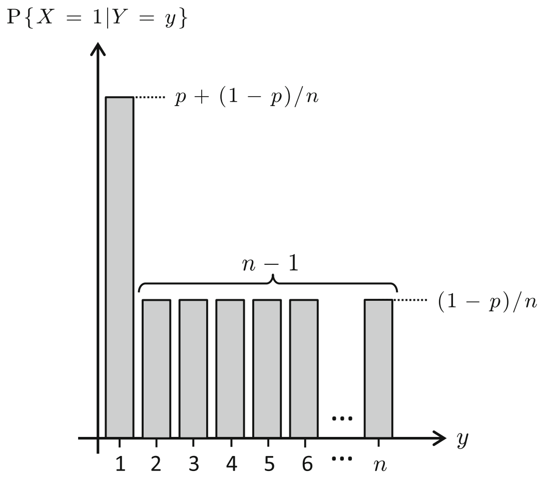


Fig. 10 Probability that the original sender of a given message be the user 1, conditioned to the observation that the last sender in the forwarding path is user y. From this figure, we observe the PMF attains its maximum value when this last sender is precisely the user 1

that is, the MAP error conditioned on the observation y. Because Hamming distance implies, by virtue of (1), that Bayes estimation is equivalent to MAP estimation, it follows that the attacker's (best) decision rule is $\hat{x}(y) = y$. Leveraging on this observation, we obtain that the privacy level provided by this variant of Crowds is

$$\mathcal{P}(y) = \varepsilon_{\mathrm{MAP}} = 1 - P\{X = y|y\} = (1-p)(1-1/n),$$

from which it follows an entirely expected result—the lower the probability p of forwarding a message directly to the server, the higher the privacy provided by the protocol, but the higher the delay in the delivery of said message.

In the following, we consider the measurement of the privacy protection offered by this protocol, in terms of the three Rényi's entropies introduced in Sect. 5.1, namely the min-entropy $H_\infty(X|y)$, the Shannon entropy $H_1(X|y)$ and the Hartley entropy $H_0(X|y)$ of the r.v. X, modeling the actual sender of a given message (the privacy attacker's target), given the observation of the user who last forwarded it, y. Specifically, we connect the interpretations described in Sect. 5.1 to the example at hand.

But first we would like to recall from Sect. 5.1.1 that $H_\infty(X|y)$, $H_1(X|y)$ and $H_0(X|y)$ may be considered, from the point of view of the user, as a worst-case, average-case, and best-case measurements of privacy, respectively, in the sense that

$$H_\infty(X|y) \leqslant H_1(X|y) \leqslant H_0(X|y),$$

owing to (6), with equality if and only if the conditional PMF of X given $Y = y$ is uniform. Revisiting the interpretations given in that section, recall that the min-entropy $H_\infty(X|y)$ is directly connected with the maximum probability, in our case $\max_{x_i} p_{X|Y}(x_i|y) = p+(1-p)/n$, on account of (10). More concretely, and in the context of our example, min-entropy reflects the model in which a privacy attacker makes a single guess of the originator of a message, specifically the most likely one, which corresponds to $x = y$.

At the other extreme, the Hartley entropy $H_0(X|y)$ is a possibilistic rather than probabilistic measure, as it corresponds to the assumption that a privacy attacker would not content themselves with discarding all but the most likely sender, but consider instead all possible users. More accurately, measuring privacy as a Hartley's entropy essentially boils down to the cardinality of the set of all possible originators of a message, namely $H_0(X|y) = \log n$.

On a middle ground lies Shannon's entropy, which was interpreted in Sect. 5.1.2 by means of the AEP, specifically in terms of the effective cardinality of the set of typical sequences of i.i.d. samples of an r.v. Put in the context of our Crowds-like protocol, however, Shannon's entropy may be deemed as an average-case metric that considers the entire PMF of X given $Y = y$, and not merely its maximum value or its support set.

7 Guide for designers of SDC and ACSs

The purpose of this section is to show the applicability of our framework to those designers of SDC and ACSs who, wishing to quantify the level of protection offered by their systems, do not want to delve into the mathematical details set forth in Sect. 5. In order to assist such designers in the selection of the privacy metric most appropriate for their requirements, this section revises the application scenarios of SDC and anonymous communications, and classifies some of the metrics used in these fields in terms of worst case, average case, and best case, from the perspective of the user.

Before proceeding any further, we would like to briefly recall the distinction precisely between worst-case, average-case, and best- case measurements of privacy. To this end, consider the scenario of ACSs in general and mixes in particular. In this specific scenario, the knowledge of the privacy attacker may be modeled by a probability distribution on the possible senders of a given message. A clear example of best-case privacy metric is Hartley's entropy, which measures the degree of anonymity attained by the mere cardinality of the set of candidate senders, or equivalently, by the logarithm of such cardinality. Loosely speaking, Hartley's entropy may be regarded as a best-case metric (from the point of view of users), in the sense that it represents a privacy attacker's thorough effort in considering any and all possibilities, regardless of their likelihood.

On the opposite extreme, min-entropy may be understood as the MAP estimation error where the attacker simply guesses the most likely outcome. This information-theoretic quantity may be construed as a worst-case metric, in the sense that the attacker is concerned with the most vulnerable statistical link between senders and messages. Finally, Shannon's entropy takes into account the underlying probability distribution in its entirety, between the extremes posed by the previous two metrics, yielding a quantity bounded according to (6). For this reason, one may think of it as an average-case metric.

Next, we elaborate on the distinction between Hamming and non-Hamming distortion functions, between whether these functions are known or unknown to the system, and finally between single and multiple-occurrence data. The reason is that the understanding of these concepts is fundamental for a system designer who, following the arguments sketched in Fig. 4, wants to choose the suitable metrics for their field of application. With this purpose, next we illustrate these concepts by means of a couple of simple albeit insightful examples.

The first consideration a system designer should take into account when applying our framework refers to the geometry of the attacker's distortion function d_{A}, namely whether it is a *Hamming* or a *non-Hamming* function. To illustrate this key point, consider a set of users in a social network. A Hamming function taking as inputs the users u_1 and u_2 would model an attacker who contemplates *only* their identities when comparing them, and ignores any other information such as the relationship between them within the social network, their profile similarity or their common interests. On the contrary, a more sophisticated adversary could represent said network by a graph, modeling users and relationships among them as nodes and edges, respectively. Leveraging on this graph, the attacker could use a non-Hamming function to compute the number of hops separating these two users and, accordingly, lead to the conclusion that they are, for example, close friends since $d_{\mathrm{A}}(u_1, u_2) = 1$.

The second consideration builds on the assumption of a non-Hamming attacker's distortion function. Under this premise, we contemplate two possible cases—when the function is *known* to the system and when *not*. The former case is illustrated, for instance, in the context of LBSs—in this application scenario, an adversary will probably use the Euclidean distance to measure how their estimated location differs from the user's actual location. The latter case, i.e., when the measure of distortion used by the attacker is unknown to the system, would undoubtedly model a more general and realistic scenario. As an example of this case, consider a system perturbing the queries that a user wants to submit to a database, and an attacker wishing to ascertain the actual queries of this user. Suppose that these queries are one-word queries and that the perturbation mechanism replaces them with synonyms or semantically similar words. Under these assumptions, our attacker could opt for a non-Hamming distortion function and measure the distance between the actual query and the estimate as the number of edges in a given ontology graph. Although the system could be aware of this fact, the specific ontology used by the attacker could not be available to the system, and consequently the distortion function would remain unknown.

Our last consideration is related to the nature of the variables of our framework, summarized in Table 1. Specifically, we contemplate two possible cases—*single* and *multiple-occurrence* data. The former case considers such variables to be tuples of a small number of components, and the latter assumes that these variables are sequences of data. An LBS attacker who observes the disclosed, possibly perturbed location of a user and makes a single guess about their actual location is an example of single-occurrence data. To illustrate the case of multi-occurrence data, consider a set of users exchanging messages through a mix system. Recall that such systems delay and reorder messages with the aim of concealing who is communicating with whom. Among the multiple attacks these systems are vulnerable to, the statistical disclosure attack [42] is a good example for our purposes of illustration, since it assumes an adversary who observes a large number or *sequence* of messages coming out of the mix, with the aim of tracing back their originators.

Having examined these key aspects of our framework, now we turn our attention, first, to the application scenario of SDC, and secondly, to the case of ACSs. In the former scenario, a data publisher aims at protecting the privacy of the individuals appearing in a microdata set. Depending on the privacy requirements, the publisher may want to prevent an attacker from ascertaining the confidential attribute value of any respondent in the released table. Under this requirement, t-closeness and mutual information appear as acceptable measures of privacy, since both criteria protect against confidential *attribute disclosure*. Recall that the assumptions on which they are based are a prior belief about the value of the confidential attribute in the table, and a posterior belief of said value given by the observation that the user belongs to a particular group of this table. Building on these premises, t-closeness may be regarded as a *worst-case* measurement of privacy, in the sense that it identifies the group of users whose distribution of the confidential attribute deviates the most from the distribution of this same attribute in the entire table. In this regard, we would like to note that a worst-case metric from the point of view of the user is a best-case measure from the standpoint of the attacker, and vice versa.

Although t-closeness overcomes the similarity and skewness attacks mentioned in Sect. 2.2, its main limitation is that no computational procedure to reach this criterion has been specified. An alternative is the mutual information between the confidential attributes and the observation, an average-case version of t-closeness that leads to a looser measure of privacy. In any of these two metrics, it is assumed the more general case in which the attacker's distortion function is not the Hamming distance. Specifically, this assumption models an adversary who does not content themselves with finding out whether the estimate and the unknown match, but wishes to quantify how much they diverge.

Another distinct privacy requirement is that of *identity disclosure*, whereby a publisher wishes to protect the released table against a linking attack. In this attack, the adversary's aim is to uncover the identity of the individuals in the released table by linking the records in this table to a public data set including identifier attributes. Under this requirement and under the assumption that the attacker regards each respondent within a particular group as equally likely, k-anonymity may be deemed as a *best-case* measure of privacy, determined by Hartley's entropy. We refer to this criterion as a best-case metric precisely due to the naive assumption of a uniform distribution of the identifier attribute. In other words, the underlying adversarial model does not contemplate that an attacker may have background knowledge that allows them to consider certain users as more likely than others. In the end, we may also regard the l-diversity criterion as a best-case metric, since it assumes a uniform distribution of the confidential attribute on a set of at least l values. Put another way, this rudimentary adversarial model does not contemplate, for example, the fact that certain values of the confidential attribute may be semantically similar.

In the scenario of anonymous communication systems, there exists a wide variety of approaches. Among them, a popular anonymous communication protocol is Crowds. Although in this section we limit the discussion of the privacy provided by such systems to a variant of this protocol, we would like to stress that the conclusions drawn here may be extended to other anonymous systems. Having said this, recall that in the original Crowds protocol, a system designer makes available to users a collaborative protocol that helps them enhance the anonymity of the messages sent to a common, untrusted Web server. The design parameters are the number of users participating in the protocol and the probability of forwarding a message directly to the server.

In our variant of this protocol, however, we contemplate an attacker who strives to guess the identity of the sender of a given message, based on the knowledge of the last user in the forwarding path. Under this adversarial model, we may regard min-entropy, Shannon's entropy or Hartley's entropy as particular cases of our measure of privacy, depending on the specific strategy of the attacker. For example, under an adversary who uses MAP estimation and, accordingly, opts for the last sender, min-entropy may be interpreted as a worst-case privacy metric. Alternatively, we may assume an attacker that considers the entire probability distribution of possible senders, and not only the most likely candidate. In this case, Shannon's entropy may be deemed as an average-case measure. Finally, under a rudimentary attacker who takes into account just the number of potential originators of the message, Hartley's entropy may be regarded as a best-case measurement of privacy. The discussion in this section is summarized in Table 4.

8 Conclusion

A wide variety of privacy metrics have been proposed in the literature. Most of these metrics have been conceived for specific applications, adversarial models, and privacy threats and thus are difficult to generalize. Even for specific applications, we often find that various privacy metrics are available. For example, to measure the anonymity provided by anonymous communication networks, several flavors of entropy (Shannon, Hartley, min-entropy) can be found in the literature, while no guidelines exist that explain the relationship between the different proposals, and provide an understanding of how to interpret and put in context the results provided by each of them.

In the scenario of SDC, numerous approaches attempt to capture, to a greater or lesser degree, the private information leaked as a result of the dissemination of microdata sets. In this spirit, k-anonymity is possibly the best-known privacy measure, mainly due to its mathematical tractability. However, numerous extensions and enhancements were introduced later with the aim of overcoming its limitations. While all these metrics have provided further insight into our understanding of privacy, the research community would benefit from a framework embracing all those metrics and making it possible to compare them, and to evaluate any privacy-protecting mechanism by the same yardstick.

In this work, we propose a privacy measure intended to tackle the above issues. Our approach starts with the definition and modeling of the variables of a general framework. Then, we proceed with a mathematical formulation of privacy, which essentially emerges from BDT. Specifically, we define privacy as the estimation error incurred by an attacker. We first propose what we refer to as conditional privacy, meaning that our measure is conditioned on an attacker's particular observation. Accordingly, we define the terms of average privacy and worst-case privacy.

The formulation is then investigated theoretically. Namely, we interpret a number of well-known privacy criteria as particular cases of our more general metric. The arguments behind these justifications are based on fundamental results related to the fields of information theory, probability theory and BDT. More accurately, we interpret our privacy criterion as k-anonymity and l-diversity principles by connecting them to Rényi's entropy and the concept of confidence set. Under certain assumptions, a conditional version of the AEP allows us to interpret Shannon's entropy as an arbitrarily high-confidence set. Then, the total variation distance and Pinsker's inequality justify t-closeness requirement and the criterion proposed in [15] as particular instances of our

Table 4 Guide for designers of SDC and ACSs

	Worst case	Average case	Best case
Statistical disclosure control	t-Closeness	Mutual information	k-Anonymity l-diversity
Anonymous communication systems	Min-entropy	Shannon's entropy	Hartley's entropy

This table classifies several privacy metrics depending, first, on whether they are regarded as worst-case, average-case, and best-case measures, and secondly on their application domain

measure of privacy. In the course of this interpretation, we find that our formulation bears a strong resemblance with the rate-distortion problem in information theory.

Our survey of privacy metrics, our detailed analysis of their connection with information theory, and our mathematical unification as an attacker's estimation error, shed new light on the understanding of those metrics and their suitability when it comes to applying them to specific scenarios. In regard to this aspect, two sections are devoted to the classification of several privacy metrics, showing the relationships with our proposal and the correspondence with assumptions on the attacker's strategy. While the former section approaches this from a theoretical perspective, the latter shows the applicability of our framework to those designers of SDC and ACSs who do not wish to delve into the mathematical details. It is also our goal to illustrate the riveting interplay between the field of information privacy on the one hand, and on the other the fields of information theory and stochastic estimation, while bridging the gap between the respective communities.

A couple of simple albeit insightful examples are also presented. Our first example quantifies the level of privacy provided by a privacy-enhancing mechanism that perturbs location information in the scenario of LBS. Under certain assumptions on the adversarial model, our measure of privacy becomes the MSE. Then we turn our attention to the scenario of anonymous communication systems and measure the degree of anonymity achieved by a modification of the collaborative protocol Crowds. We contemplate different strategies for the attacker and, accordingly, interpret min-entropy, Shannon's entropy and Hartley's entropy as worst-case, average-case, and best-case privacy metrics.

In closing, we hope that this unified perspective of privacy metrics, drawing upon the principles of information theory and Bayesian estimation, is a helpful, illustrative step toward the systematic modeling of privacy-preserving information systems.

Acknowledgments This work was partly supported by the Spanish Government through projects Consolider Ingenio 2010 CSD2007-00004 "ARES", TEC2010-20572-C02-02 "Consequence" and by the Government of Catalonia under grant 2009 SGR 1362. Additional sources of funding include IWT SBO SPION, GOA TENSE, the IAP Programme P6/26 BCRYPT, and the FWO project "Contextual privacy and the proliferation of location data". D. Rebollo-Monedero is the recipient of a Juan de la Cierva postdoctoral fellowship, JCI-2009-05259, from the Spanish Ministry of Science and Innovation. C. Diaz is funded by an FWO postdoctoral grant.

References

1. Willenborg, L., DeWaal, T.: Elements of Statistical Disclosure Control. Springer, NewYork (2001)
2. Jabine, T.B.: Statistical disclosure limitation practices at united states statistical agencies. J. Off. Stat. **9**(2), 427–454 (1993)
3. Citteur, C.A.W., Willenborg, L.C.R.J.: Public use microdata files: current practices at national statistical bureaus. J. Off. Stat. **9**(4), 783–794 (1993)
4. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. **14**(1), 189–201 (2002)
5. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Min. Knowl. Discov. **11**(2), 195–212 (2005)
6. Solanas, A., Martínez-Ballesté, A., Domingo-Ferrer, J.: VMDAV: a multivariate microaggregation with variable group size. In: Proceedings in Computational Statistics (COMPSTAT), Springer, Rome, Italy (2006)
7. Rebollo-Monedero, D., Forné, J., Soriano, M.: Private location-based information retrieval via k-anonymous clustering. In: Proceedings of the CNIT International Workshop on Digital Communication, Series Lecture Notes in Computer Science (LNCS), Sept. 2009, Springer, Sardinia, Italy, invited paper (2009)
8. Sweeney, L.: k-Anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzz. Knowl. Based Syst. **10**(5), 557–570 (2002)
9. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
10. Truta, T.M., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: Proceedings of the International Workshop on Privacy Data Management (PDM), Atlanta, GA, p. 94 (2006)
11. Machanavajjhala, A., Gehrke, J., Kiefer, D., Venkitasubramanian, M.: l-Diversity: privacy beyond k-anonymity. In: Proceedings of the IEEE International Conference on Data Engineering (ICDE), Atlanta, GA, Apr 2006, p. 24 (2006)
12. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the IEEE International Conference on Data Engineering (ICDE), Istanbul, Turkey, Apr 2007, pp. 106–115 (2007)
13. Brickell, J., Shmatikov, V.: The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas, NV, Aug 2008 (2008)
14. Dwork, C.: Differential privacy. In: Proceedings of the International Colloquium on Automata, Languages and Programming, Springer, pp. 1–12 (2006)
15. Rebollo-Monedero, D., Forné, J., Domingo-Ferrer, J.: From t-closeness-like privacy to postrandomization via information

theory. IEEE Trans. Knowl. Data Eng., **22**(11), 1623–1636, Nov. 2010. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190 (2010)

16. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Commun. ACM **24**(2), 84–88 (1981)
17. Cottrell, L.: Mixmaster and remailer attacks. [Online]. Available: http://obscura.com/loki/remailer/remailer-essay.html (1994)
18. Danezis, G., Dingledine, R., Mathewson, N.: Mixminion: design of a type III anonymous remailer protocol. In: Proceedings of the IEEE Symposium on Security and Privacy (SP), Berkeley, CA, pp. 2–15, May (2003)
19. Duckham, M., Mason, K., Stell, J., Worboys, M.: A formal approach to imperfection in geographic information. Comput. Environ. Urban Syst. **25**(1), 89–103 (2001)
20. Lehmann, E.L.: Theory of Point Estimation. Springer, New York (1983)
21. Reiter, M.K., Rubin, A.D.: Crowds: anonymity for web transactions. ACM Trans. Inform. Syst. Secur. **1**(1), 66–92 (1998)
22. Berthold, O., Pfitzmann, A., Standtke, R.: The disadvantages of free MIX routes and how to overcome them. In: Proceedings of the Designing Privacy Enhancing Technologies: Workshop Design Issues in Anonymity, Unobservability, Series Lecture Notes in Computer Science (LNCS), Springer, Berkeley, CA, July 2000, pp. 30–45 (2000)
23. Serjantov, A., Danezis, G.: Towards an information theoretic metric for anonymity. In: Proceedings of the Workshop on Privacy Enhancing Technologies (PET), vol. 2482, pp. 41–53. Springer (2002)
24. Díaz, C., Seys, S., Claessens, J., Preneel, B.: Towards measuring anonymity. In: Proceedings of the Workshop on Privacy Enhancing Technologies (PET), Series Lecture Notes in Computer Science (LNCS), vol. 2482, pp. 54–68. Springer, Apr 2002 (2002)
25. Tóth, G., Hornák, Z., Vajda, F.: Measuring anonymity revisited. In: Proceedings of the Nordic Workshop on Secure IT Systems, Nov, pp. 85–90 (2004)
26. Clauß, S., Schiffner, S.: Structuring anonymity metrics. In: Proceedings of the ACM Workshop on Digital Identity Management (DIM), ACM, Fairfax, VA, Nov 2006, pp. 55–62 (2006)
27. Syverson, P., Stubblebine, S.: Group principals and the formalization of anonymity. In: Proceedings of the World Congress on Formal Methods, pp. 814–833 (1999)
28. Mauw, S., Verschuren, J., de Vink, E.P.: A formalization of anonymity and onion routing. In: Proceedings of the European Symposium on Research in Computer Security (ESORICS), vol. 3193. Lecture Notes in Computer Science (LNCS), pp. 109–124 (2004)
29. Feigenbaum, J., Johnson, A., Syverson, P.: A model of onion routing with provable anonymity. In: Proceedings of the Financial Cryptography and Data Security (FI), Springer (2007)
30. Edman, M., Sivrikaya, F., Yener, B.: A combinatorial approach to measuring anonymity. IEEE J. Intell. Secur. Inform. 356–363 (2007)
31. Gierlichs, B., Troncoso, C., Díaz, C., Preneel, B., Verbauwhede, I.: Revisiting a combinatorial approach toward measuring anonymity. In: Proceedings of the ACM Workshop on Privacy in the Electronic Society, ACM, pp. 111–116 (2008)
32. Bagai, R., Lu, H., Li, R., Tang, B.: An accurate system-wide anonymity metric for probabilistic attacks. In: Proceedings of the Workshop on Privacy Enhancing Technologies (PET), Series Lecture Notes in Computer Science (LNCS), vol. 6794, pp. 117–133, Springer (2011)
33. Shokri, R., Freudiger, J., Jadliwala, M., Hubaux, J.P.: A distortion-based metric for location privacy. In: Proceedings of the ACM Workshop on Privacy in the Electronic Society (2009)
34. Shokri, R., Theodorakopoulos, G., Boudec, J.Y.L., Hubaux, J.P.: Quantifying location privacy. In: Proceedings of the IEEE Symposium on Security and Privacy (SP), IEEE Comput. Soc., Washington, DC, USA, pp. 247–262 (2011)
35. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, New York (2006)
36. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985)
37. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
38. Algoet, P.H., Cover, T.M.: A sandwich proof of the Shannon-McMillan-Breiman theorem. Ann. Prob. **16**(2), 899–909 (1988)
39. Shannon, C.E.: Coding theorems for a discrete source with a fidelity criterion. In: IRE National Convention Record, vol. 7, Part 4, pp. 142–163 (1959)
40. Reid, D.B.: An algorithm for tracking multiple targets. IEEE Trans. Autom. Control **24**(6), 843–854 (1979)
41. Hastings, W.K.: Monte carlo sampling methods using markov chains and their applications. Biometrika **57**(1), 97–109 (1970)
42. Danezis, G.: Statistical disclosure attacks: traffic confirmation in open environments. In: Proceedings of the Security and Privacy in the Age of Uncertainty (SEC), Athens, Greece, May 2003, pp. 421–426 (2003)