

Beyond k -Anonymity: A Decision Theoretic Framework for Assessing Privacy Risk

Guy Lebanon¹, Monica Scannapieco^{*2}
Mohamed R. Fouad¹, and Elisa Bertino¹

¹ Purdue University, USA, lebanon@stat.purdue.edu,
{mrf,bertino}@cs.purdue.edu

² ISTAT and Università di Roma “La Sapienza”, Italy, scannapi@istat.it

Abstract. An important issue any organization or individual has to face when managing data containing sensitive information, is the risk that can be incurred when releasing such data. Even though data may be sanitized, before being released, it is still possible for an adversary to reconstruct the original data by using additional information that may be available, for example, from other data sources. To date, however, no comprehensive approach exists to quantify such risks. In this paper we develop a framework, based on statistical decision theory, to assess the relationship between the disclosed data and the resulting privacy risk. We relate our framework with the k -anonymity disclosure method; we make the assumptions behind k -anonymity explicit, quantify them, and extend them in several natural directions.

1 Introduction

The problem of data privacy is today a pressing concern for many organizations and individuals. The release of data may have some important advantages in terms of improved services and business, and also for the society at large, such as in the case of homeland security. However, unauthorized data disclosures can lead to violations to the privacy of individuals, can result in financial and business damages, as in the case of data pertaining to enterprises; or can result in threats to national security, as in the case of sensitive GIS data [6]. Preserving the privacy of such data is a complex task driven by various goals and requirements. Two important privacy goals are: (i) preventing identity disclosure, and (ii) preventing sensitive information disclosure. Identity disclosure occurs when the released information makes it possible to identify entities either directly (e.g., by publishing identifiers like SSNs) or indirectly (e.g., by linkage with other sources). Sensitive information typically includes information that must be protected by law, for example medical data, or is required by the subjects described by the data. In the latter case, data sensitivity is a subjective measure and may differ across entities.

* This work was performed while visiting research assistant at CERIAS and Department of Computer Sciences, Purdue University, USA. It is partially supported by the project “ESTEEM” (<http://www.dis.uniroma1.it/~esteem/index.html>)

To date, an important practical requirement for any privacy solution is the ability to quantify the *privacy risk* that can be incurred in the release of certain data. However, most of the work related to data privacy has focused on how to transform the data so that no sensitive information is disclosed or linked to specific entities. Because such techniques are based on *data transformations* that modify the original data with the purpose of preserving privacy, they are mainly focused on the tradeoff between data privacy and data quality (see e.g. [9, 2, 3, 5]). Conversely, few approaches exist to quantify privacy risks and thus to support informed decisions. Duncan et al. [4] describes a framework, called Risk-Utility (R-U) confidentiality map, which addresses the tradeoff between data utility and disclosure. Lakshmanan et al. [8] is an approach to the risk analysis for disclosed anonymized data that models a database as series of transactions and the attacker’s knowledge as a belief function. Our model is fundamentally different from both works; indeed, we deal exactly with relational instances, rather than with generic files or data frequencies; also, we incorporate the concept of data sensitivity into our framework and we consider generic disclosure procedures, not only anonymization like in [8].

The goal of the work presented in this paper is to propose a comprehensive framework for the estimation of privacy risks. The framework is based on statistical decision theory and introduces the notion of a *disclosure rule*, that is a function representing the data disclosure policy. Our framework estimates the *privacy risk* by means of a function that takes into account a given disclosure rule and (possibly) the knowledge that can be exploited by the attacker. It is important to point out that our framework is able to assess privacy risks also when no information is available about the knowledge, referred to as *dictionary*, that the adversary may exploit. The privacy risk function incorporates both identity disclosure and sensitive information disclosure. We introduce and analyze different shapes of the privacy risk function. Specifically, we define the risk in the classical decision theory formulation and in the Bayesian formulation. We prove several interesting results within our framework: we show that, under reasonable hypotheses, the estimated privacy risk is an upper bound for the true privacy risk; we analyze the computational complexity of evaluating the privacy risk function, and we propose an algorithm for efficiently finding the disclosure rule that minimizes the privacy risk. We finally gain insight by showing that the privacy risk is a quantitative framework for exploring the assumptions and consequences of k -anonymity.

2 Privacy Risk Framework

As private information in databases is being disclosed, undesired effects occur such as privacy breaches, and financial loss due to identity theft. To proceed with a quantitative formalism we assume that we obtain a numeric description, referred to as loss, of that undesired effect. The loss may be viewed as a function of (i) whether the disclosed information enables identification and (ii) the sensitivity of the disclosed information. The first argument of the loss function encapsulates whether the disclosed data can be tied to a specific entity or not.

Consider for example the case of a hospital disclosing a list of the ages of patients, together with data indicating whether they are healthy or not. Even though this data is sensitive and if there is a little chance that the disclosed information can be tied to specific individuals, no privacy loss occurs as the data is anonymous. The second argument of the loss function, the sensitivity of the disclosed information, may be high as is often the case for sensitive medical data. On the other hand, other disclosed information such as gender, may be only marginally private or not private at all. It is important to note that a precise quantification of the sensitivity of the disclosed information may depend on the entity to whom the data relates. For example, data such as annual income and past medical history may be very sensitive to some and only marginally sensitive to others.

Let T be a relation with a relational scheme $T(A_1, \dots, A_m)$, where each attribute A_i is defined over the domain $\text{Dom}_i \cup \{\perp, \S\}$, with the only exception of A_1 as detailed later. The relation T stores the records that are considered for disclosure and has some values either missing or suppressed for privacy preservation. Specifically, a null value is denoted by \perp whereas a suppressed value is denoted by \S . Furthermore, we denote the different attribute values of a specific record \mathbf{x} in T using a vector notation (x_1, \dots, x_m) . The first attribute x_1 corresponds to a unique record identifier that can be neither \perp nor \S . The set of all possible records may be written as

$$\mathcal{X} = (\text{Dom}_1) \times (\text{Dom}_2 \cup \{\perp, \S\}) \times \dots \times (\text{Dom}_m \cup \{\perp, \S\}).$$

If T has cardinality n , it can be seen as a subset of \mathcal{X}^n which we may think of as a matrix whose rows are the different records. We refer to the i^{th} record in such a relation as \mathbf{x}_i and its j^{th} attribute as x_{ij} .¹

2.1 Disclosure Rules and Privacy Risk

Statistical decision theory [10] offers a natural framework for measuring the quantitative effect of the information disclosure phenomenon. The uncertainty is encoded by a parameter θ abstractly called “a state of nature” which is typically unknown. However, it is known that θ belongs to a set Θ , usually a finite or infinite subset of \mathbb{R}^l . The decisions are being made based on a sample of observations (x_1, \dots, x_n) , $x_i \in \mathcal{X}$ and are represented via a function $\delta : \mathcal{X}^n \rightarrow \mathcal{A}$ where \mathcal{A} is an abstract action space. The function δ is referred to as a decision policy or decision rule. A key element of statistical decision theory is that the state of nature θ governs the distribution p_θ that generates the observed data.

Instead of decision rules $\delta : \mathcal{X}^n \rightarrow \mathbb{A}$, we introduce *disclosure rules* defined as follows.

Definition 1. A disclosure rule δ is a function $\delta : \mathcal{X} \rightarrow \mathcal{X}$ such that

$$[\delta(\mathbf{z})]_j = \begin{cases} \perp & z_j = \perp \\ \S & \text{the } j^{\text{th}} \text{ attribute is suppressed} \\ z_j & \text{otherwise} \end{cases}$$

¹ Note that throughout the paper, records and vectors are denoted by *bold italic* symbols whereas variables and attributes are denoted by only *italic* symbols.

The state of nature θ that influences the disclosure outcome is the side information used by the attacker in his identification attempt. Such side information θ is often a public data resource composed of identities and their attributes, for example a phone book. The distribution over records p_θ is taken to be the empirical distribution \tilde{p} over the data that is to be disclosed $\mathbf{x}_1, \dots, \mathbf{x}_n$, defined below.

Definition 2. *The empirical distribution \tilde{p} on \mathcal{X} associated with a set of records $\mathbf{x}_1, \dots, \mathbf{x}_n$ is*

$$\tilde{p}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n 1_{\{\mathbf{z}=\mathbf{x}_i\}}$$

where $1_{\{\mathbf{z}=\mathbf{x}_i\}}$ is 1 if $\mathbf{z} = \mathbf{x}_i$ and 0 otherwise.

The empirical distribution is used for defining the risk associated with a disclosure rule δ using the mechanism of expectation. Note that the expectation with respect to \tilde{p} is simply the empirical mean $E_{\tilde{p}}(f(\mathbf{x}, \theta)) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i, \theta)$. The loss and risk functions in the privacy adaptation of statistical decision theory are defined below.

Definition 3. *The loss function $\ell : \mathcal{X} \times \Theta \rightarrow [0, \infty]$ measures the loss incurred by disclosing the data $\delta(\mathbf{z}) \in \mathcal{X}$ due to possible identification based on $\theta \in \Theta$.*

Definition 4. *The risk of the disclosure rule δ in the presence of side information θ is the average loss of disclosing the records $\mathbf{x}_1, \dots, \mathbf{x}_n$: $R(\delta, \theta) = E_{\tilde{p}(\mathbf{z})}(\ell(\delta(\mathbf{z}), \theta)) = \frac{1}{n} \sum_{i=1}^n \ell(\delta(\mathbf{x}_i), \theta)$.*

Definition 5. *The Bayes risk of the disclosure rule δ is $R(\delta) = E_{p(\theta)}(R(\delta, \theta))$ where $p(\theta)$ is a prior probability distribution on Θ .*

It is instructive at this point to consider in detail the identification process and its possible relations to the loss function. We use the term *identification attempt* to refer to the process of trying to identify the entity represented by the record. We refer to the subject performing the identification attempt as the *attacker*. The attacker performs an identification attempt based on the disclosed record $\mathbf{y} = \delta(\mathbf{x}_i)$ and additional side information θ referred to as a *dictionary*. The role of the dictionary is to tie a record \mathbf{y} to a list of possible candidate identities consistent with the record \mathbf{y} , i.e. having the same values on common fields. For example, consider \mathbf{y} being (first-name, surname, phone#) and the dictionary being a phone book. The attacker needs only considering dictionary entities that are consistent with the disclosed record. Recall that some of the attributes (first-name, surname, phone#) may be replaced with \perp or \S symbols due to missing information or due to the disclosure process, respectively. In this example, if all the attribute values are revealed and the available side information is an up-to-date phone book, it is likely that only one entity exists in the dictionary that is consistent with the revealed information. On the other hand, if the attribute value for phone# is suppressed, by replacing it with \S symbol, the phone-book θ may or may not yield a single consistent entity, depending on the popularity of the (first-name, surname) combination. From the attacker's

stand point, missing values are perceived the same way as suppressed values. Thus, in the rest of the paper and for the sake of notational simplicity, both missing and suppressed values will be denoted by the symbol \perp .

Note that the loss function $\ell(\delta(\mathbf{x}_i), \theta)$ measures the loss due to disclosing $\delta(\mathbf{x}_i)$ in the presence of the side information – in this case the dictionary θ . Specifying the loss is typically entity and problem dependent. We can, however, make some progress by decomposing the loss into two parts: (i) the ability to identify the entity represented by $\delta(\mathbf{x}_i)$ based on the side information θ and (ii) the sensitivity of the information in $\delta(\mathbf{x}_i)$. The identification part is formalized by the random variable Z defined as follows.

Definition 6. Let $\rho(\delta(\mathbf{x}_i), \theta)$ denote the set of individuals in the dictionary θ consistent with the record $\delta(\mathbf{x}_i)$. Moreover, let the random variable $Z(\delta(\mathbf{x}_i))$ be a binary variable that takes value 1 if $\delta(\mathbf{x}_i)$ is identified and 0 otherwise.

Assuming a uniform selection of entries in the dictionary by the attacker, we have

$$p_{Z(\delta(\mathbf{x}_i))}(1) = \begin{cases} |\rho(\delta(\mathbf{x}_i), \theta)|^{-1} & \rho(\delta(\mathbf{x}_i), \theta) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

and $p_{Z(\delta(\mathbf{x}_i))}(0) = 1 - p_{Z(\delta(\mathbf{x}_i))}(1)$.

2.2 Sensitivity

The sensitivity of disclosed data is formalized by the following definition.

Definition 7. The sensitivity of a record is measured by a function $\Phi : \mathcal{X} \rightarrow [0, +\infty]$ where higher values indicate higher sensitivity.

We allow Φ to take on the value $+\infty$ in order to model situations where the information in the record is so private that its disclosure is prohibited under any positive identification chance.

The sensitivity $\Phi(\delta(\mathbf{x}_i))$ measures the adverse effect of disclosing the record $\delta(\mathbf{x}_i)$ if the attacker correctly identifies it. We make the assumption (whose relaxation is straightforward) that if the attacker does not correctly identify the disclosed record, there is no adverse effect. The adverse effect is therefore a random variable with two possible outcomes: $\Phi(\delta(\mathbf{x}_i))$ with probability $p_{Z(\delta(\mathbf{x}_i))}(1)$ and 0 with probability $p_{Z(\delta(\mathbf{x}_i))}(0)$. It is therefore natural to account for the uncertainty resulting from possible identification by defining the loss $\ell(\mathbf{y}, \theta)$ as the expectation of the adverse effect resulting from disclosing $\mathbf{y} = \delta(\mathbf{x}_i)$

$$\begin{aligned} \ell(\mathbf{y}, \theta) &= E_{p_{Z(\mathbf{y})}}(\Phi(\mathbf{y})Z(\mathbf{y})) \\ &= p_{Z(\mathbf{y})}(1) \cdot \Phi(\mathbf{y}) + p_{Z(\mathbf{y})}(0) \cdot 0 = \frac{\Phi(\mathbf{y})}{|\rho(\mathbf{y}, \theta)|} \end{aligned}$$

where the last equality holds if the dictionary selection probabilities are uniform and $\rho(\mathbf{y}, \theta) \neq \emptyset$.

The risk $R(\delta, \theta)$ with respect to the distribution \tilde{p} that governs the record set $\mathbf{x}_1, \dots, \mathbf{x}_n$ becomes

$$R(\delta, \theta) = E_{\tilde{p}(\mathbf{x})}(\ell(\delta(\mathbf{z}), \theta)) = \frac{1}{n} \sum_{i=1}^n \frac{\Phi(\delta(\mathbf{x}_i))}{|\rho(\delta(\mathbf{x}_i), \theta)|}$$

and the Bayes risk under the prior $p(\theta)$ becomes (if Θ is discrete replace the integral below by a sum)

$$R(\delta) = E_{p(\theta)}(R(\delta, \theta)) = \frac{1}{n} \sum_{i=1}^n \Phi(\delta(\mathbf{x}_i)) \int_{\Theta} \frac{p(\theta)d\theta}{|\rho(\delta(\mathbf{x}_i), \theta)|}.$$

We now provide more details concerning records \mathbf{x}_i and their space \mathcal{X} , that will be useful in the following. As introduced, a record $\mathbf{x}_i \in \mathcal{X}$ has attribute values (x_{i1}, \dots, x_{im}) where each attribute $x_{ij}, j = 2, \dots, m$ either takes values in a domain Dom_j or is unavailable, in which case we denote it by \perp . The first attribute is $x_{i1} \in \text{Dom}_1$. We assume that $[\delta(\mathbf{x}_i)]_1 = x_{i1}$, i.e. $[\delta(\mathbf{x}_i)]_1$ cannot have \perp values. This assumption is for notational purposes only and in reality the disclosed data should be taken to be $[\delta(\mathbf{x}_i)]_2, \dots, [\delta(\mathbf{x}_i)]_m$. Notice also that the primary key of the relation can be distinct from the introduced record identifier, and can be one or more attributes defined over $\text{Dom}_2, \dots, \text{Dom}_m$. We make the assumption $[\delta(\mathbf{x}_i)]_1 = x_{i1}$ in order to allow a possible dependency of $\Phi(\delta(\mathbf{x}_i))$ on the identifier $x_{i1} = [\delta(\mathbf{x}_i)]_1$ which enables the flexibility needed to treat attribute values related to different entities differently. For example, a certain entity, such as a specific person, may wish to protect certain attributes such as religion or age that may be less private for a different person. Possible expressions for the Φ function are provided in the Appendix.

3 Tradeoff between Disclosure Rules and Privacy Risk

In evaluating disclosure rules δ we have to balance the following tradeoff. On one hand, disclosing private information incurs the privacy risk $R(\delta, \theta)$. On the other hand, disclosing information serves some purpose, or else no information would ever be disclosed. Such disclosure benefit may arise from various reasons such as increased productivity due to the sharing of commercial data.

We choose to represent this tradeoff by specifying a set of disclosure rules Δ that are acceptable in terms of their disclosure benefit. From this set, we seek to choose the rule that incurs the least privacy risk $\delta^* = \arg \min_{\delta \in \Delta} R(\delta, \theta)$. Notice that this framework is not symmetric in its treatment of the disclosure benefit and privacy risk and emphasizes the increased importance of privacy risk in the tradeoff.

It is difficult to provide a convincing example of a set Δ without specifying in detail the domain and the disclosure benefit. Nevertheless, we specify below several sets of rules that serve to illustrate the decision theoretic framework of this paper. The basic principle behind these rules is that the more attribute values are

being disclosed, the greater the disclosure benefit is. The details of the specific application will eventually determine which set of rules is most appropriate.

The three sets of rules below are parameterized by a positive integer k . The set Δ_1 consists of rules that disclose a total of k attribute values for all records combined

$$\Delta_1 = \{\delta : \delta(\mathbf{x}_1), \dots, \delta(\mathbf{x}_n) \text{ contain a total of } k \text{ non } \perp \text{ entries}\}.$$

The second set Δ_2 consists of rules that disclose a certain number of attribute values for each record

$$\Delta_2 = \{\delta : \forall i \delta(\mathbf{x}_i) \text{ contains } k \text{ non } \perp \text{ entries}\}.$$

The third set Δ_3 consists of rules that disclose a certain number of attribute values for each attribute

$$\Delta_3 = \{\delta : \forall j \{[\delta(\mathbf{x}_i)]_j\}_{i=1}^n \text{ contains } k \text{ non } \perp \text{ entries}\}.$$

The set Δ_1 may be applicable in situations where the disclosure benefit is influenced simply by the number of disclosed attribute values. Such a situation may arise if there is a need for computing statistics on the joint space of represented entities-attributes without an emphasis on either dimension. The set Δ_2 may be applicable when the disclosure benefit is tied to per-entity data, for example discovering association rules in grocery store transactions. A rule $\delta \in \Delta_2$ guarantees that there are sufficient attributes disclosed for each entity to obtain meaningful conclusions. Similarly, the set Δ_3 may be useful in cases where there is an emphasis on per-attribute data.

Disclosure rules $\delta \in \Delta$ are evaluated on the basis of the risk functions $R(\delta, \theta), R(\delta)$. In some cases, the attacker’s dictionary is publicly available. We can then treat the “true” side information θ^{true} as known, and the optimal disclosure rule is the minimizer of the risk

$$\delta^* = \arg \min_{\delta \in \Delta} R(\delta, \theta^{\text{true}}). \quad (1)$$

If the attacker’s side information is not known, but we can express a prior belief $p(\theta)$ describing the likelihood of $\theta^{\text{true}} \in \Theta$, we may use the Bayesian approach and select the minimizer of the Bayes risk

$$\delta_B^* = \arg \min_{\delta \in \Delta} E_{p(\theta)}(R(\delta, \theta)). \quad (2)$$

If there is no information concerning θ^{true} whatsoever, a sensible strategy is to select the minimax rule δ_M^* that achieves the least worst risk, i.e. δ_M^* satisfies

$$\sup_{\theta \in \Theta} R(\delta_M^*, \theta) = \inf_{\delta \in \Delta} \sup_{\theta \in \Theta} R(\delta, \theta). \quad (3)$$

Notice that in all cases above we try to pick the best disclosure rule in terms of privacy risk, out of a set Δ of disclosure rules that are acceptable in terms of

the amount of revealed data. The rules δ^* , δ_B^* , δ_M^* are useful, respectively, if we know θ^{true} , we have a prior over it, or we have no knowledge whatsoever.

An alternative situation to the one above is that the database is trying to estimate (or minimize) the privacy risk $R(\delta, \theta^{\text{true}})$ based on side information $\hat{\theta} \neq \theta^{\text{true}}$ available to the database. In such cases we can use $R(\delta, \hat{\theta})$ as an estimate for $R(\delta, \theta^{\text{true}})$ but we need to find a way to connect the two risks above by leveraging on a relation between $\hat{\theta}$ and θ^{true} .

A reasonable assumption is that the database dictionary $\hat{\theta}$ is specific to the database while the attacker's dictionary θ^{true} is a more general-purpose dictionary. We can then say that θ^{true} contains the records in $\hat{\theta}$ as well as additional records. Following the same reasoning we can also assume that for each record that exists in both dictionaries, $\hat{\theta}$ will have more attribute values that are not \perp . For example, consider a database of employee records for some company. $\hat{\theta}$ would be the database dictionary and θ^{true} would be a general-purpose dictionary such as a phone-book. It is natural to assume that θ^{true} will contain additional records over the records in $\hat{\theta}$ and that the non- \perp attributes in θ^{true} (e.g. `first-name,surname,phone#`) will be more limited than the non- \perp attributes in $\hat{\theta}$. After all, some of the record attributes are private and would not be disclosed in order to find their way into the attacker's dictionary (resulting in more \perp symbols in the θ^{true}).

Under the conditions specified above we can show that the true risk is bounded from above by $R(\delta, \hat{\theta})$ and that the chosen rule $\arg \min_{\delta \in \Delta} R(\delta, \hat{\theta})$ has a risk that is guaranteed to bound the true privacy risk. This is formalized below.

We consider dictionaries θ as relational tables, where $\theta_i = (\theta_{i1}, \dots, \theta_{iq})$ is a record of a relation $T_\theta(A_1, \dots, A_q)$, with A_1 corresponding to the record identifier.

Definition 8. We define the relation \preceq between dictionaries $\theta = (\theta_1, \dots, \theta_{l_1})$ and $\eta = (\eta_1, \dots, \eta_{l_2})$ by saying that $\theta \preceq \eta$ if for every θ_i , $\exists \eta_v$ such that $\eta_{v1} = \theta_{i1}$ and $\eta_{vk} \neq \perp \Rightarrow \theta_{ik} = \eta_{vk}$. The relation \preceq constitutes a partial ordering on the set of dictionaries Θ .

Theorem 1. If $\hat{\theta}$ contains records that correspond to $\mathbf{x}_1, \dots, \mathbf{x}_n$ and $\hat{\theta} \preceq \theta^{\text{true}}$, then

$$\forall \delta \quad R(\delta, \theta^{\text{true}}) \leq R(\delta, \hat{\theta}).$$

Proof. For every disclosed record $\delta(\mathbf{x}_i)$ there exists a record in $\hat{\theta}$ that corresponds to it and since $\hat{\theta} \preceq \theta^{\text{true}}$ there is also a record in θ^{true} that corresponds to it. As a result, $\rho(\delta(\mathbf{x}_i), \hat{\theta})$ and $\rho(\delta(\mathbf{x}_i), \theta^{\text{true}})$ are non-empty sets.

For an arbitrary $\mathbf{a} \in \rho(\delta(\mathbf{x}_i), \hat{\theta})$ we have $\mathbf{a} = \hat{\theta}_v$ for some v and since $\hat{\theta} \preceq \theta^{\text{true}}$ there exists a corresponding record θ_k^{true} . The record θ_k^{true} will have the same or more \perp symbols as \mathbf{a} and therefore $\theta_k^{\text{true}} \in \rho(\delta(\mathbf{x}_i), \theta^{\text{true}})$. The same argument can be repeated for every $\mathbf{a} \in \rho(\delta(\mathbf{x}_i), \hat{\theta})$ thus showing that $\rho(\delta(\mathbf{x}_i), \hat{\theta}) \subseteq \rho(\delta(\mathbf{x}_i), \theta^{\text{true}})$ or $|\rho(\delta(\mathbf{x}_i), \theta^{\text{true}})|^{-1} \leq |\rho(\delta(\mathbf{x}_i), \hat{\theta})|^{-1}$.

The probability of identifying $\delta(\mathbf{x}_i)$ by the attacker is thus smaller than the identification probability based on $\hat{\theta}$. It then follows that for all i , $\ell(\delta(\mathbf{x}_i), \theta^{\text{true}}) \leq \ell(\delta(\mathbf{x}_i), \hat{\theta})$ as well as $R(\delta, \theta^{\text{true}}) \leq R(\delta, \hat{\theta})$.

Solving (1) in the general case requires evaluating $R(\delta, \theta^{\text{true}})$ for each $\delta \in \Delta$ and selecting the minimum. The reason is that the dictionary θ controlling the identification distribution $p_{Z(\delta(\mathbf{x}_i))}(1) = |\rho(\delta(\mathbf{x}_i), \theta)|^{-1}$ is of arbitrary shape. A practical assumption, that is often made for high dimensional distributions, is that the distribution underlying θ factorizes into a product form

$$\frac{|\rho(\delta(\mathbf{x}_i), \theta)|}{N} = \prod_j \frac{|\rho_j([\delta(\mathbf{x}_i)]_j, \theta)|}{N} \quad \text{or}$$

$$|\rho(\delta(\mathbf{x}_i), \theta)| = \prod_j \alpha_j([\delta(\mathbf{x}_i)]_j, \theta)$$

for some appropriate functions α_j . In other words, the appearances of y_j for all $j = 1, \dots, m$ in θ are independent random variables. Returning to the phone-book example, the above assumption implies that the popularity of first names does not depend on the popularity of last names, e.g.,

$$\begin{aligned} p(\text{first-name} = \text{Mary} | \text{surname} = \text{Smith}) &= \\ p(\text{first-name} = \text{Mary} | \text{surname} = \text{Johnson}) &= \\ p(\text{first-name} = \text{Mary}). \end{aligned}$$

The independence assumption does not hold in general, as attribute values may be correlated, for instance, by integrity constraints; we plan to relax it in future work.

First we analyze the complexity of evaluating the risk function $R(\delta, \theta)$. This would depend on the complexity of computing Φ , denoted by $C(\Phi)$, and the complexity of computing $|\rho(\delta(\mathbf{x}_i), \theta)|$ which is $O(Nm)$, where N is the number of records in the dictionary θ . Solving $\arg \min_{\delta \in \Delta} R(\delta, \theta)$ by enumeration requires $O(n)(C(\Phi) + O(Nm)) \cdot |\Delta|$ computations.

We have $|\Delta_1| = \binom{nm}{k}$, $|\Delta_2| = \binom{m}{k}^n$, $|\Delta_3| = \binom{n}{k}^m$ and $C(\Phi)$ typically being $O(m)$ for the additive and multiplicative forms. In a typical setting where $k \ll m$ we have for Δ_2 and a linear or multiplicative Φ , a minimization complexity of $O(nNm^{kn+1})$.

The complexities above are computed for the naive enumeration algorithm. A much more efficient algorithm for obtaining $\arg \min_{\delta \in \Delta} R(\delta, \theta)$ for Δ_2 and Φ_5 under the assumption of dictionary independence is described below.

If we define $C_1(\mathbf{y}) = \{j : j > 1, y_j = \perp\}$, $C_2(\mathbf{y}) = \{j : j > 1, y_j \neq \perp\}$ we have

$$\begin{aligned} \ell(\mathbf{y}, \theta) &= \frac{\prod_{j \in C_2(\mathbf{y})} e^{w_j, y_j}}{|\rho(\mathbf{y}, \theta)|} = \frac{\prod_{j \in C_2(\mathbf{y})} e^{w_j, y_j}}{\prod_{k > 1} \alpha_k(y_k, \theta)} \\ &= \prod_{j \in C_2(\mathbf{y})} \frac{e^{w_j, y_j}}{\alpha_j(y_j, \theta)} \cdot \prod_{l \in C_1(\mathbf{y})} \frac{1}{\alpha_l(\perp, \theta)} \\ &= \prod_{j \in C_2(\mathbf{y})} e^{w_j, y_j} \frac{\alpha_j(\perp, \theta)}{\alpha_j(y_j, \theta)} \cdot \prod_{l=2}^m \frac{1}{\alpha_l(\perp, \theta)}. \end{aligned}$$

To select the disclosure of k attributes that minimizes the above loss it remains to select the set $C_2(\mathbf{y})$ of k indices that minimizes the loss. This set corresponds to the k smallest $\{e^{w_{j,y_1}} \frac{\alpha_j(\perp, \theta)}{\alpha_j(y_j, \theta)}\}_{j=2}^m$ and leads to the following algorithm.

Algorithm 1: MinRisk

- (1) **foreach** $i = 1, \dots, n$
- (2) **foreach** $j = 2, \dots, m$
- (3) set $\gamma_j := e^{w_{j,x_{i1}}} \frac{\alpha_j(\perp, \theta)}{\alpha_j(x_{ij}, \theta)}$
- (4) identify the k smallest elements in $\{\gamma_j\}_{j=2}^m$
- (5) set $\delta(\mathbf{x}_i)$ to disclose the attributes corresponding to these k elements

Theorem 2. *The algorithm `MinRisk` for solving $\arg \min_{\delta \in \Delta} R(\delta, \theta)$ requires $O(nNm)$ computations.*

Proof. For each record $\mathbf{y} = \mathbf{x}_i$ we compute the following. The set $\{\gamma_j = e^{w_{j,y_1}} \frac{\alpha_j(\perp, \theta)}{\alpha_j(y_j, \theta)}\}_{j=2}^m$ can be obtained in $O(Nm)$. Moreover, the set corresponding to the k smallest elements in $\{\gamma_j\}_{j=2}^m$ can be obtained in two steps: (i) Get the k^{th} -smallest element in $\{\gamma_j\}_{j=2}^m$, γ' (this requires $O(m)$ computations), then (ii) scan the set $\{\gamma_j\}_{j=2}^m$ for elements $< \gamma'$ (again, this requires $O(m)$ computations). Hence the overall complexity of the above procedure is $O(n)(O(Nm) + O(m)) = O(nNm)$.

4 Privacy Risk and k -Anonymity

k -Anonymity [9] has recently received considerable attention by the research community [11, 1]. Given a relation T , k -anonymity ensures that each disclosed record can be indistinctly matched to at least k individuals in T . It is enforced by considering a subset of the attributes called *quasi-identifiers*, and forcing the disclosed values of these attributes to appear at least k times in the database. k -anonymity uses two operators to accomplish this task: suppression and generalization. We ignore the role of generalization operators in this paper as our privacy framework is cast solely in terms of suppression at attribute-level. However, it is straightforward to extend the privacy risk framework to include generalization operators leading to a more complete analogy with k -anonymity, and we plan to do it in future work.

In its original formulation, k -anonymity does not seem to make any assumptions on the possible external knowledge that could be used for entity identification and does not characterize the privacy loss. However, k -anonymity does make strong implicit assumptions whose absence eliminates any motivation it might possess. Following the formal presentation of k -anonymity in the privacy risk context, we analyze these assumptions and possible relaxations.

Since the k -anonymity requirement is enforced on the relation T , the anonymization algorithm considers the attacker’s dictionary as equal to the

relation $T = \theta$. Representing the k -anonymity rule by δ_k^* we have that the k -anonymity constraints may be written as

$$\forall i \quad |\rho(\delta_k^*(\mathbf{x}_i), T)| \geq k. \quad (4)$$

The sensitivity function is taken to be constant $\Phi \equiv c$ as k -anonymity considers only the constraints (4) and treats all attributes and entities in the same way. As a result, the loss incurred by k -anonymity δ_k^* is bounded by $\ell(\delta_k^*(\mathbf{x}_i), T) \leq c/k$ where equality is achieved if the constraint $|\rho(\delta_k^*(\mathbf{x}_i), T)| = k$ is met. On the other hand, any rule δ_0 that violates the k -anonymity requirement for some \mathbf{x}_i will incur a loss higher (under $\theta = T$ and $\Phi \equiv c$) than the k -anonymity rule

$$\ell(\delta_0(\mathbf{x}_i), T) = \frac{c}{|\rho(\delta_0(\mathbf{x}_i), T)|} \geq \ell(\delta_k^*(\mathbf{x}_i), T).$$

We thus have the following result presenting k -anonymity as optimal in terms of the privacy risk framework.

Theorem 3. *Let δ_k^* be a k -anonymity rule and δ_0 be a rule that violates the k -anonymity constraint, both with respect to $\mathbf{x}_i \in T$. Then*

$$\ell(\delta_k^*(\mathbf{x}_i), T) \leq c/k < \ell(\delta_0(\mathbf{x}_i), T).$$

As the above theorem implies, the k -anonymity rule minimizes the privacy loss per example \mathbf{x}_i and may be seen as $\arg \min_{\delta \in \Delta} R(\delta, T)$ where Δ is a set of rules that includes both k -anonymity rules and rules that violate the k -anonymity constraints. The assumptions underlying k -anonymity, in terms of the privacy risk framework are

1. $\theta^{\text{true}} = T$
2. $\Phi \equiv c$
3. Δ is under-specified.

The first assumption may be taken as an indication that k -anonymity does not assume any additional information regarding the attacker's dictionary. As we showed earlier, the resulting risk $R(\delta_k^*, T) \leq c/k$ may be seen as a bound on the true risk $R(\delta_k^*, \theta^{\text{true}})$ under some assumptions. Alternatively, the privacy framework also introduces the mechanisms of the minimax rule and the Bayes rule if additional information is available such as the set Θ of possible dictionaries or even a prior on Θ . Moreover, the attacker's dictionary θ is often a standard public resource. In such cases the constraints (4) should be taken with respect to θ , rather than T .

The second assumption $\Phi \equiv c$ is somewhat questionable. The privacy risk framework measures the loss as the expectation of the data sensitivity, as measured by Φ , with respect to the identification probability. Taking the sensitivity function Φ to be a constant ignores the role of the sensitivity of the disclosed data in the framework. The loss measured would depend only on the identification probability and not on the types of attributes that are being disclosed. In other words, privacy loss becomes synonymous with identification. This leads to the paradoxical situation where the disclosure of a sensitive attribute such as the type of medical situation diagnosed (e.g. HIV positive) may lead to lower

risk than the disclosure of a less sensitive attribute such as the precise date of the most recent doctor visit (assuming that the precise date of the most recent doctor visit leads to greater identification chance).

The third assumption implies that the set Δ may be specified in several ways. Recall that the risk minimization framework is based on the assumption that there is a tradeoff in disclosing private information. On one hand the disclosed data incurs a privacy loss and on the other hand disclosing data serves some benefit. The risk minimization framework $\arg \min_{\delta \in \Delta} R(\delta, \theta)$ assumes that Δ contains a set of rules acceptable in terms of their disclosure benefit, and from which we select the one incurring the least risk. k -Anonymity ignores this tradeoff and the set of candidate rules Δ may be specified in several ways, for example $\Delta = \Delta_0 \cup \{\delta_k^*\}$ where Δ_0 contains rules that violate the k -anonymity constraints.

In light of the above, k -anonymity may be modified in several directions. If we possess some information concerning the attacker’s dictionary we can do a better job using $\delta^*, \delta_B^*, \delta_M^*$, as well as upper-bound the true risk using $\hat{\theta}$ (see Section 5 for an explanation of these concepts). We can alter Φ to account for the different sensitivities of different attributes, perhaps even allowing Φ to be entity-dependent. Finally, a more careful consideration of the disclosure benefit may lead to a better definition of the rule set Δ allowing the preference of some k -anonymity rule over others. As mentioned earlier, our discussion is in terms of the suppression operator alone. Nevertheless, the same arguments and conclusions apply to k -anonymity using both suppression and generalization operators.

5 Conclusion

In this paper we have described a novel framework for assessing privacy risk in a variety of situations. We consider optimal disclosure rules in the contexts of exact knowledge, partial knowledge, and no knowledge with respect to the attacker’s side information. We discuss several forms for expressing the largely ignored role of data sensitivity in the privacy risk. We have shown that the estimated privacy risk is an upper bound for the true privacy risk, under some reasonable hypotheses on the relationships between the attacker’s dictionary and the database dictionary. We have also provided a computationally efficient algorithm for minimizing the privacy risk under some hypotheses. Finally, we have proved the generality of our framework by showing that k -anonymity is a special case of it, and we have highlighted, in our decision theory based formulation, the particular assumptions underlying k -anonymity.

At first glance it may appear that the privacy risk framework requires knowledge that is typically unavailable or somewhat undesirable assumptions. After all, it seems possible to use k -anonymity without making such compromising assumptions. This is a misleading interpretation as any attempt at forming a sensible privacy policy or characterizing the result of private data disclosure requires such assumptions. In particular, assumptions have to be made concerning the attacker’s resources and the data sensitivity. Existing algorithms such as k -anonymity typically make such assumptions implicitly. However, in order to obtain a coherent view of privacy it is essential to make these assumptions explicit, and discuss their strengths and weaknesses.

Appendix

Possible Expression for the Φ function

In the following, we review several possible expressions for the function Φ , which models the sensitivity of a record that is involved in a disclosure process, if the attacker correctly identifies it. Since Φ is defined on \mathcal{X} , the set of all possible records, defining it by a lookup table is often impractical. We therefore consider several options leading to compact and efficient representation. Given a disclosed record $\mathbf{y} = \delta(\mathbf{x}_i)$, the simplest meaningful form for Φ is a linear combination of non-negative weights w_j over the disclosed attributes

$$\Phi_1(\mathbf{y}) = \sum_{j>1:y_j \neq \perp} w_j$$

where w_j represents the sensitivity associated with the corresponding attribute A_j . A weight of $+\infty$ represents the most sensitive information that may only be disclosed if there is zero chance of it leading to identification (we define $0 \cdot \infty = 0$).

Alternatively, we may assume that attributes sensitivities vary from record to record, but are identical for each record. In this case a linear form would yield

$$\Phi_2(\mathbf{y}) = \sum_{j>1:y_j \neq \perp} w_{y_1} = w_{y_1} \times (\# \text{ of disclosed attributes})$$

where w_{y_1} is the weight associated with the release of each attribute value of the record \mathbf{y} . Incorporating different weights for different attribute values and different records yields the linear form

$$\Phi_3(\mathbf{y}) = \sum_{j>1:y_j \neq \perp} w_{j,y_1}$$

where w_{j,y_1} represents the sensitivity of attribute j in record \mathbf{y} . These weights may be represented by a two dimensional table of numbers. A possible special case is to assume the decomposition of attribute-record weights $w_{j,y_1} = w_j w'_{y_1}$ leading to a representation of the weight table as two vectors of weights (w_1, \dots, w_m) , (w'_1, \dots, w'_n) .

An extension of the linear representation of Φ is accomplished through k -order interactions. In k -order interaction we use additional weights to capture interactions of k attributes that are not accounted for in the linear forms above. k -order interactions take into account cases in which the simultaneous disclosure of multiple attribute values needs to be weighted differently when compared to the independent disclosure of each single value. An example of $k = 2$ -order interaction yields the form

$$\Phi_4(\mathbf{y}) = \sum_{j>1:y_j \neq \perp} w_{j,y_1} + \sum_{j>1:y_j \neq \perp} \sum_{h>j:y_h \neq \perp} w_{j,h,y_1}.$$

As k increases in magnitude, the class of functions represented by Φ becomes richer. Reaching $k = m$ would provide arbitrary flexibility with respect to the

functional form of Φ (however, as mentioned above, the representation and computation are impractical for large m). If the simple linear form is not sufficient to capture the user-specified privacy values, it is likely that increasing k to 2 or 3 will bring the functional form of Φ quite close to the user-specified values.

In some cases, a multiplicative rather than linear form is preferred. In this case, a convenient form is

$$\Phi_5(\mathbf{y}) = \exp \left(\sum_{j>1:y_j \neq \perp} w_{j,y_1} \right) = \prod_{j>1:y_j \neq \perp} e^{w_{j,y_1}}$$

or its k -order extensions analogous to the linear forms above. The multiplicative form Φ_5 has the advantage that if we increase one privacy weight w_{ij} while fixing the other weights, Φ increases exponentially rather than linearly. Since the disclosure of extremely private information should not be considered even if the remaining attributes are non-private, the multiplicative form Φ_5 is the most appropriate in many settings.

Experiments

The goals of our experiments are 3-fold: (i) to validate the risk associated with different dictionaries, (ii) to assess the impact of different parameters on the privacy risk, and (iii) to use the proposed framework to assess the relationship between the estimated risk and the true risk.

We conducted our experiments on a real Wal-Mart database: An **item description** table of more than 400,000 records each with more than 70 attributes is used in the experiments. Part of the table is used to represent the disclosed data whereas the whole table is used to generate different dictionary. Throughout all our experiments, the risk components are computed as follows. First, the identification risk is computed with the aid of the Jaro distance function[7] that is used to identify dictionary items consistent with a released record to a certain extent (we used 80% similarity threshold to imply consistency.) Second, the sensitivity of the disclosed data is assessed by means of random weights that are generated using a uniform random number generator.

The impacts of the number of disclosed attributes per record, k , and the dictionary size on the privacy risk are reported in Figure 1 (left). As k increases (i.e. extra data is being disclosed) and by fixing the dictionary size, the possibility of identifying the entity, to which the data pertain, increases, thus increasing the privacy risks. We increase k from 25% to 100% of the total number of attributes. On the other hand, by fixing the number of data attributes that are disclosed, the relation between the risk and dictionary size is inversely related. The larger the size of the dictionary the attacker uses, the lower the probability that the entity be identified. Different dictionaries are generated from the original table with sizes varying from 10% to 100% of the size of the whole table. Moreover, the experimental data show that the multiplicative model for sensitivity is always superior in terms of the modeled risk to the additive model.

The relationship between the true risk $R(\delta, \theta^{\text{true}})$ and the estimated risk $R(\delta, \hat{\theta})$ is reported in the scatter plot in Figure 1 (right). As we proved before,

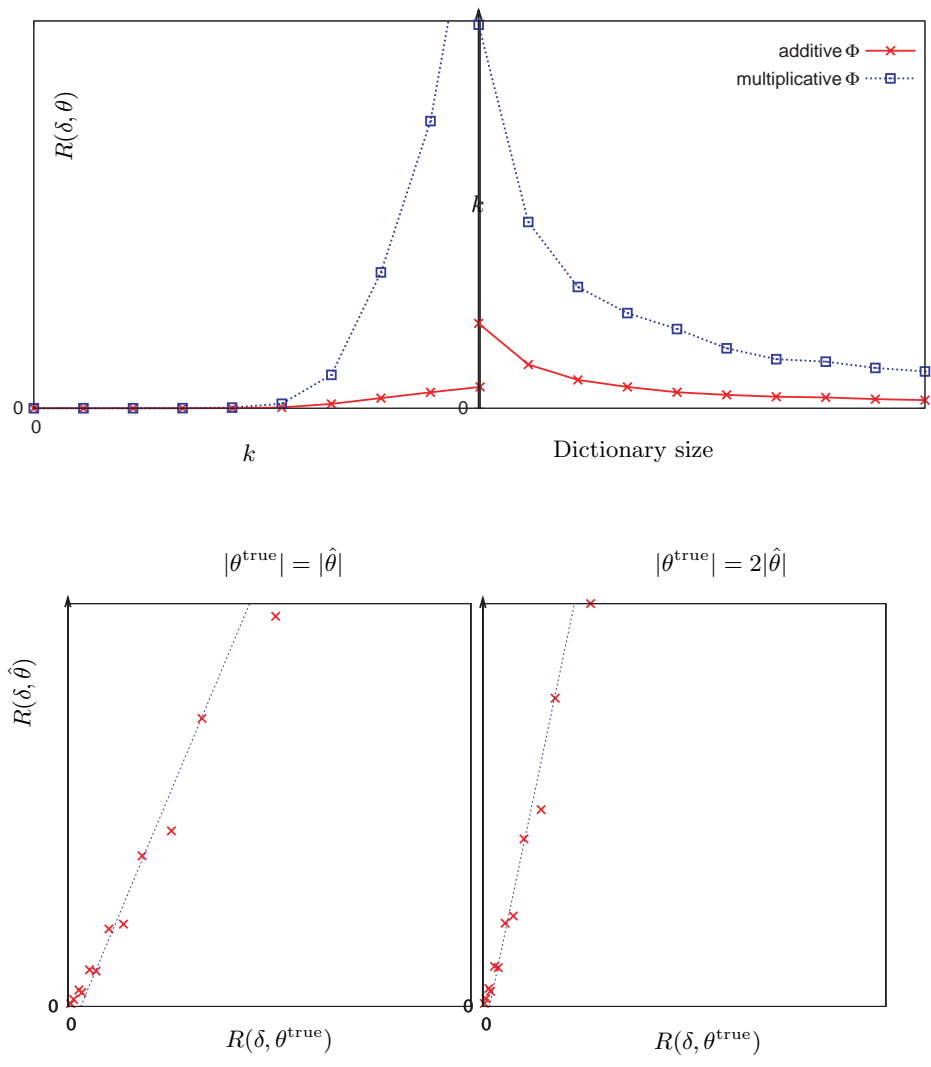


Fig. 1. The risk associated with different dictionaries and k values (left) and the relationship between the true risk and the estimated risk (right).

$R(\delta, \hat{\theta})$ is always an upper bound of $R(\delta, \theta^{\text{true}})$ (all the points occur above the line $y = x$). Note that, as the size of the true dictionary becomes significantly larger than the size of the estimated dictionary, the points seem to trace a steeper line which means that the estimated risk becomes a looser upper bound for the true risk.

References

1. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, P. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *Proc. of ICDT 2005*.
2. Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The sulq framework. In *Proc. of PODS 2005*.
3. I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proc. of PODS 2003*.
4. G.T. Duncan, S.A. Keller-McNulty, and L.S. Stokes. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences (NISS), December 2001.
5. A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proc. of PODS 2003*.
6. Guidelines for Providing Appropriate Access to Geospatial Data in Response to Security Concerns. Federal geographic data committee, 2005. http://fgdc.er.usgs.gov/fgdc/homeland/access_guidelines.pdf.
7. M.A. Jaro. UNIMATCH: A record linkage system, user's manual. In *Washington DC: U.S. Bureau of the Census*, 1978.
8. L.V.S. Lakshmanan, R.T. Ng, and G. Ramesh. To do or not to do: the dilemma of disclosing anonymized data. In *Proc. of SIGMOD 2005*.
9. P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proc. of PODS 1998*.
10. A. Wald. *Statistical Decision Functions*. Wiley, 1950.
11. S. Zhong, Z. Yang, and R.N. Wright. Privacy-enhancing k-anonymization of customer data. In *Proc. of PODS 2005*.