

2009

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

Guy Lebanon

Monica Scannapieco

Mohamed Fouad

Elisa Bertino

Purdue University, bertino@cs.purdue.edu

Follow this and additional works at: <http://docs.lib.purdue.edu/ccpubs>

 Part of the [Engineering Commons](#), [Life Sciences Commons](#), [Medicine and Health Sciences Commons](#), and the [Physical Sciences and Mathematics Commons](#)

Lebanon, Guy; Scannapieco, Monica; Fouad, Mohamed; and Bertino, Elisa, "Beyond k -Anonymity: A Decision Theoretic Framework for Assessing Privacy Risk" (2009). *Cyber Center Publications*. Paper 236.

<http://docs.lib.purdue.edu/ccpubs/236>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

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

Guy Lebanon*, Monica Scannapieco**, Mohamed R. Fouad***, Elisa Bertino***

*College of Computing, Georgia Institute of Technology, Atlanta, USA.

E-mail: lebanon@cc.gatech.edu

**Department of Systems and Computer Sciences, Rome University, Italy.

E-mail: monscan@dis.uniroma1.it

***Department of Computer Science, Purdue University, West Lafayette, USA.

E-mail: {mrf, bertino}@cs.purdue.edu

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 using additional information thus resulting in privacy violations. To date, however, a systematic approach to quantify such risks is not available. In this paper we develop a framework, based on statistical decision theory, that assesses the relationship between the disclosed data and the resulting privacy risk. We model the problem of deciding which data to disclose, in terms of deciding which disclosure rule to apply to a database. We assess the privacy risk by taking into account both the entity identification and the sensitivity of the disclosed information. Furthermore, we prove that, under some conditions, the estimated privacy risk is an upper bound on the true privacy risk. Finally, we relate our framework with the k -anonymity disclosure method. The proposed framework makes the assumptions behind k -anonymity explicit, quantifies them, and extends them in several natural directions.

Keywords. Privacy, Security, Risk Management, Data Sharing, Decision Theory, Anonymity

1 Introduction

Data sharing has 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 of individuals' privacy, can result in financial

*The work reported here has been supported by the NSF grants IPS-0712846 "Security Services for Healthcare Applications" and IPS-0712856 "Decision Theoretic Approaches to Measuring and Minimizing Customized Privacy Risk".

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 geospatial data.

Preserving the privacy of such data is a complex task driven by two important privacy goals: (i) preventing the identification of the entity relating to the data, and (ii) preventing the disclosure of sensitive information. Entity identification occurs when the released information makes it possible to identify the entity either directly (e.g., by publishing identifiers like SSNs), or indirectly (e.g., by linkage with other sources). Sensitive information includes information that must be protected by law such as medical data, or is deemed sensitive by the entity to whom the data pertains. In the latter case, data sensitivity is a subjective measure whose nature may differ across entities.

In many cases, a careful evaluation needs to be carried out in order to assess whether the privacy risk associated with the dissemination of certain data outweighs the benefits of such dissemination. As pointed out in the recent guidelines issued by the [3], "Some organizations have curtailed access without assessing the risk to security, the significance of consequences associated with improper use of the data, or the public benefits for which the data were originally made available. Contradictory decisions and actions by different organizations easily negate each organization's actions." [21] introduces a way for providing privacy protection while constructing algorithms that learn information from disparate data and introduces the notion of privacy-enhanced linking. [7] shows that it is impossible to achieve privacy with respect to worst-case external knowledge.

To date, 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, the main focus of such approaches has been the tradeoff between data privacy and data quality e.g., [18, 8]. Similar approaches based on output perturbation have been proposed by [2] and [4].

An important practical requirement for any privacy solution is the ability to quantify the privacy risk that can be incurred by the release of certain 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, or by record linkage techniques [26]. A possible adversarial scenario is depicted in Figure 1: an attacker exploits data released by an organization by linking it with previously obtained data concerning the same entity to gain an enhanced insight about this organization. Indeed, this insight would help the attacker narrow down possible mismatches when it is compared against a public dictionary, and consequently raising the identification risk. The goal of the work presented in this paper is to develop, for the first time, a comprehensive framework for quantifying such privacy risk and supporting informed disclosure policies.

The framework we propose 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 taking 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 concerning the knowledge or dictionary that the adversary may exploit. The privacy risk function naturally 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, for either the linkage or the no-linkage scenario. We prove several interesting results within our framework including that under reasonable hypotheses, the estimated privacy risk is an upper bound on the true privacy risk. Finally, we gain insight by showing that the

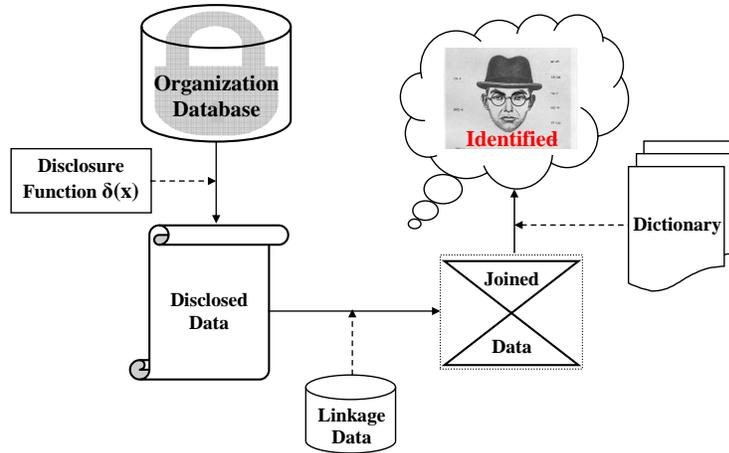


Figure 1: Adversarial framework for identity discovery

privacy risk is a quantitative framework for exploring the assumptions and consequences of k -anonymity. The work presented in this paper was initially presented in the conference paper [13].

2 Basics of Statistical Decision Theory

Statistical decision theory [24] offers a natural framework for measuring the quantitative effect of information disclosure. As the necessary modifications of decision theory are relatively minor, we are able to adapt a considerable array of tools and results from over 50 years of impressive research. We describe below only the principal concepts of this theory in its traditional abstract setting, and then proceed to apply it to the information disclosure problem.

Statistical decision theory deals with the abstract problem of making decisions in an uncertain situation. Decisions, their properties and the resulting effect are specified formally, enabling their quantitative and rigorous study. 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 $p_\theta(x)$ that generates the observed data (x_1, \dots, x_n) . Given the state of nature θ , the loss incurred by taking an action $\delta(x_1, \dots, x_n) \in \mathcal{A}$ is determined by a non-negative loss function

$$\ell : \mathcal{A} \times \Theta \rightarrow [0, +\infty] \quad \text{or} \quad \ell(\delta(x_1, \dots, x_n), \theta) \geq 0.$$

We sometime denote $\ell(\delta(x_1, \dots, x_n), \theta)$ as $\delta_\theta(x_1, \dots, x_n)$ when we wish to emphasize it as a function, parameterized by θ , of the observed data.

Rather than measuring the loss incurred by a specific decision rule and a specific set of observations, it makes sense to consider the expected loss, or risk, where the expectation

is taken over observations being generated from the distribution generating the data p_θ . Denoting expectations in general as

$$E_{p(x)}(h(x)) = \begin{cases} \int_{\mathcal{X}} p(x)h(x) dx & \text{continuous } x \\ \sum_{x \in \mathcal{X}} p(x)h(x) & \text{discrete } x \end{cases}$$

the expected loss or risk associated with the decision rule δ and $\theta \in \Theta$ is

$$R(\delta, \theta) = E_{p_\theta(x_1, \dots, x_n)}(\ell(\delta(x_1, \dots, x_n), \theta)) = E_{\prod p_\theta(x_i)}(\ell(\delta(x_1, \dots, x_n), \theta))$$

where the last equality assumes independence of x_1, \dots, x_n .

The two main statistical paradigms, classical statistics and Bayesian statistics, carry over to decision theory. In the classical setting of decision theory, the risk $R(\delta, \theta)$ is the main quantity of interest and its properties and relations to different decision rules δ and states θ are studied. The Bayesian approach to decision theory assumes that another piece of information is available: our prior beliefs concerning the possibility of various states of nature $\theta \in \Theta$. This prior belief is represented by a prior probability $q(\theta)$ over possible states leading to the Bayes risk

$$R(\delta) = E_{q(\theta)}(R(\delta, \theta)) = E_{q(\theta)}\{E_{p_\theta(x_1, \dots, x_n)}(\ell(\delta(x_1, \dots, x_n), \theta))\}. \quad (1)$$

Much has been said in the statistics literature over the controversy between the classical and the Bayesian points of view. Without going into this discussion, we simply point out that an advantage of the Bayes risk is that we can compare different policies δ_1, δ_2 based on a single number - their associated risks $R(\delta_1), R(\delta_2)$ leading to a partial order on all possible policies. An advantage of the classical framework is that there is no need for a prior distribution q - which is often hard or impossible to specify. In both cases, we need to have a precise specification of the probabilistic model $p_\theta(x)$, a set of possible states of nature Θ and a loss function ℓ . While p_θ and Θ depend on modeling assumptions or estimation from data, the loss function ℓ is typically elicited from a user and its subjective quality reflects the personalized nature of risk-based analysis and decision theory.

3 Privacy Risk Framework

As private information in databases is being disclosed, undesired effects occur such as privacy violations, financial loss due to identity theft, and national security breaches. 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 two arguments (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 patients' gender and whether they have a certain medical condition or not. Due to the presence of medical information, such data is clearly sensitive. However, the data sensitivity does not provide any information about the chance of tying the disclosed data to specific individuals and as a result the patients maintain their anonymity and no harmful effect is produced. The clear distinction between data sensitivity and identification, and their combination via a probabilistic framework, is a central part of our framework. The quantification of the identification probability depends on (i) the disclosed data, (ii) available side information such as national archives or a phone-book, and (iii) an attacker model.

In contrast to the identification probability, the second argument of the loss function concerning the data sensitivity depends on the entity associated with the data. Data such as annual income, medical history, and Internet purchases relating to specific users may be very sensitive to some but only marginally sensitive to others. Such personalized or customized sensitivity measures are important to be taken into consideration when measuring harmful effects and deciding on a disclosure policy. Clearly, ignoring it may lead to offering insufficient protection to a subset of people while applying excessive protection to the privacy of another subset. It is worth pointing out that we do not draw a distinction between sensitive attributes and quasi-identifiers [15, 27, 14]. Rather, our framework provides more flexibility by enabling the owners of the data to supply the sensitivity of their attributes at their discretion.

We assume that the data resides in a relational database with the relational scheme (A_1, \dots, A_m) , where each attribute A_i takes values in a domain Dom_i which includes a possible missing value symbol \perp . The space

$$\mathcal{X} = \text{Dom}_1 \times \text{Dom}_2 \times \dots \times \text{Dom}_m$$

represents the set of all possible records, both original records residing in a database and disclosed records. We make the following assumptions for the sake of notational simplicity, none of which are crucial to the presented framework. First, we assume that one of the attributes A_1 uniquely identifies the entity associated with the record. This attribute will typically not be disclosed, but is important for notational convenience. Second, we assume that the symbol $\perp \in \text{Dom}_i$ for all i , corresponds to both a missing value in the database and to attribute values that are suppressed during the disclosure procedure. Suppression of the (non-missing) j -attribute in a record $\mathbf{y} \in \mathcal{X}$ may thus be represented by a function $\delta : \mathcal{X} \rightarrow \mathcal{X}$ for which $[\delta(\mathbf{y})]_j = \perp$. Finally, we assume that the space \mathcal{X} is sufficiently rich to denote attribute generalizations, for example

$$\text{North America} \in \text{Dom}_{\text{country}}$$

represents a generalization of the country attribute to a more vague concept.

We will usually refer to an arbitrary record as \mathbf{x} , \mathbf{y} or \mathbf{z} and to a specific record in a particular database using a subscript \mathbf{x}_i (note the bold-italic typesetting representing vector notation). The attribute values of records are represented using the notation $[\mathbf{x}]_j$, $[\mathbf{x}_i]_j$ or just x_j or x_{ij} , respectively (note the non-bold typesetting). A collection of n records, for example a database containing n records, is represented by $(\mathbf{x}_1, \dots, \mathbf{x}_n) \subseteq \mathcal{X}^n$.

3.1 Disclosure Rules and Privacy Risk

Adapting the decision theory framework described in the previous section to privacy requires relatively a few changes. Instead of decision policies $\delta : \mathcal{X}^n \rightarrow \mathcal{A}$ we have disclosure policies $\delta : \mathcal{X} \rightarrow \mathcal{X}$ representing disclosing the data as is (e.g., $\delta(\mathbf{z}) = \mathbf{z}$), attribute suppression (e.g., $[\delta(\mathbf{z})]_j = \perp$), or attribute generalization (e.g., $[\delta(\mathbf{z})]_j = \text{North America}$).

The state of nature θ that influences the incurred loss $\ell_\theta = \ell(\cdot, \theta)$ is the side information used by the attackers in their identification attempt. Such side information θ is often a public data resource composed of identities and their attributes, for example a phone-book. The record distribution p is the distribution that generates the disclosed data where we omit the dependence on θ since in our case p is independent of the attacker's side information θ . In the case of disclosing a specific set of records $\mathbf{x}_1, \dots, \mathbf{x}_n$ that are known in advance, a convenient choice for p is the empirical distribution \tilde{p} over these records, defined below.

	Statistical decision theory	Privacy risk framework
\mathcal{X}	Space of abstract data	Space of disclosed or stored records
x_1, \dots, x_n	Available observations sampled from $p_{\theta_{\text{true}}}$	Records to be (partially) disclosed; determine $\tilde{p}(\mathbf{x})$
$\theta \in \Theta$	Determines the data distribution $p_{\theta}(x)$	Side information; unrelated to data
δ	Determines abstract action based on x_1, \dots, x_n	Determines what to disclose from a single record x_i
ℓ	Abstract loss; based on x_1, \dots, x_n and the model θ	Privacy loss incurred from disclosing $\delta(x_i)$ in the presence of the side information θ
$R(\delta, \theta)$	Abstract risk associated with decision rule δ and the model θ	Privacy risk associated with disclosure rule and side information θ
$R(\delta)$	Bayes risk associated with decision rule δ	Bayes risk associated with disclosure rule δ

Figure 2: Similarities and differences between statistical decision theory and the privacy framework

Definition 1. The empirical distribution \tilde{p} on \mathcal{X} associated with a set of records x_1, \dots, x_n is

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

where $1_{\{z=x_i\}}$ is 1 if $z = x_i$ and 0 otherwise.

Note that the expectations under \tilde{p} reduce to empirical means $E_{\tilde{p}}(f(\mathbf{x}, \theta)) = \frac{1}{n} \sum_{i=1}^n f(x_i, \theta)$ and the expected loss reduces to the average incurred loss with respect to disclosing x_1, \dots, x_n : $E_{\tilde{p}}(\ell_{\theta}(\delta(\mathbf{x}))) = \frac{1}{n} \sum_{i=1}^n \ell_{\theta}(x_i)$. Taking expectation with respect to distributions other than \tilde{p} can lead to a weighted average of losses, representing a situation in which some records are more important than others (although this effect can be more naturally incorporated into ℓ as described in the Section 3.3). More generally, in case of a streaming or sequential disclosure of records generated from a particular distribution p , we should compute the expected loss over that distribution in order to obtain a privacy risk relevant to the situation at hand.

The following definitions complete the adaptation of statistical decision theory to the privacy risk setting. The similarities and differences between these definitions and their counterparts of the previous section are summarized in Figure 2.

Definition 2. The loss function $\ell : \mathcal{X} \times \Theta \rightarrow [0, +\infty]$ measures the loss incurred by disclosing the data $\delta(z) \in \mathcal{X}$ due to possible identification based on the side information $\theta \in \Theta$.

Definition 3. The risk of the disclosure rule δ in the presence of side information θ is the expected loss $R(\delta, \theta) = E_{p(z)}(\ell(\delta(z), \theta))$.

Definition 4. The Bayes risk of the disclosure rule δ is $R(\delta) = E_{q(\theta)}(R(\delta, \theta))$ where $q(\theta)$ is a prior probability distribution on Θ .

3.2 Identification Probabilities, Data Sensitivity, and Loss Functions

We turn at this point to consider in detail the process of identifying the entity represented by the disclosed record, the data sensitivity, and their relation to the loss function. The identification attempt is normally carried out by the attacker who uses the disclosed record $\mathbf{y} = \delta(\mathbf{x})$ and additional side information (or dictionary θ) whose role is to tie the disclosed data to a list of possible candidate identities.

The specification of the loss function ℓ is typically entity and problem dependent. We can, however, make significant progress by decomposing the loss into two parts: (i) the attacker's probability of identifying the data owner based on the disclosed data $\delta(\mathbf{x})$ and side information θ , and (ii) the user-specified data sensitivity. While the data sensitivity is a subjective measure specified by users, the attacker's probability of identifying the data owner should be computed based on the side information θ and a probabilistic attacker model. We proceed below with describing a reasonable derivation of the attacker's identification probability and then proceed with a description of the user-specified data sensitivity function.

Given a disclosed record $\delta(\mathbf{x})$, and available side information or dictionary θ , the attacker can narrow down the list of possible identities to the subset of entity entries in θ that are consistent with the disclosed attributes $\delta(\mathbf{x})$. For example, consider \mathbf{x} being `(first-name, surname, phone-number)` and the dictionary θ being a phone-book. The attacker needs only to consider dictionary entities that are consistent with the disclosed record $\delta(\mathbf{x})$. If there are no missing values and the entire record is disclosed, i.e. $\delta(\mathbf{x}) = \mathbf{x}$, it is likely that only one entity exists in the dictionary that is consistent with the disclosed information. On the other hand, if the attribute value for `phone-number` is suppressed, the phone-book θ may yield more than a single consistent entity, depending on the popularity of the combination `(first-name, surname)`.

Formalizing the above idea we define the binary random variable Z which equals 1 if the attacker successfully identified the data owner and 0 otherwise. The identification probability $p(Z = 1)$ depends on the attacker, but in the absence of additional information we may assume that the identification attempt is a uniform selection from the set of entities in θ consistent with the disclosed $\delta(\mathbf{x})$, denoted by $\rho(\delta(\mathbf{x}), \theta)$,

$$\begin{aligned} p(Z = 1) &= \begin{cases} |\rho(\delta(\mathbf{x}), \theta)|^{-1} & \text{if } \rho(\delta(\mathbf{x}), \theta) \neq \emptyset \\ 0 & \text{if } \rho(\delta(\mathbf{x}), \theta) = \emptyset \end{cases} \quad \text{and} \\ p(Z = 0) &= 1 - p(Z = 1). \end{aligned}$$

The data sensitivity is determined by two user specified functions $\Phi, \Psi : \mathcal{X} \rightarrow [0, +\infty]$. Φ measures the harmful effect of disclosing the data assuming that the attacker's identification is successful, i.e., $\Phi(\mathbf{x})$ is $\ell(\delta(\mathbf{x}), \theta)$ provided that $\{Z = 1\}$. Similarly, Ψ measures the harmful effect of disclosing the data assuming that the attacker's identification was unsuccessful, i.e., $\Psi(\mathbf{x})$ is $\ell(\delta(\mathbf{x}), \theta)$ provided that $\{Z = 0\}$.

Putting the identification probability and sensitivity function together, we have that the harmful effect is a random variable with two possible outcomes: $\Phi(\delta(\mathbf{x}))$ with probability $p(Z = 1)$ and $\Psi(\delta(\mathbf{x}))$ with probability $p(Z = 0)$. Accounting for the uncertainty resulting from possible identification we define the loss $\ell(\mathbf{y}, \theta)$ as the expectation

$$\ell(\delta(\mathbf{x}), \theta) = p(Z = 1) \cdot \Phi(\delta(\mathbf{x})) + p(Z = 0) \Psi(\delta(\mathbf{x})).$$

Allowing Φ, Ψ to take on the value $+\infty$ enables us to model situations where the data sensitivity is so high that its disclosure is categorically prohibited (if $\Psi(\delta(\mathbf{x})) = +\infty$) or is prohibited under any positive identification chance (if $\Phi(\delta(\mathbf{x})) = +\infty$).

It is often the case that no harmful effect is caused if the attacker's identification attempt fails leading to $\Psi \equiv 0$. For simplicity, we assume this is the case in the remainder of the paper, leading to $\ell(\delta(\mathbf{x}), \theta) = p(Z = 1) \Phi(\delta(\mathbf{x}))$. The risk $R(\delta, \theta)$ with respect to the empirical

distribution \tilde{p} over the disclosed records is

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

and the Bayes risk under the prior $q(\theta)$ is

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

or its discrete equivalent if Θ is a discrete space. Similar expressions can be computed if the assumption $\Psi \equiv 0$ is relaxed.

As mentioned above, the concepts defined in this section are somewhat different from the traditional use of decision theory. The parameter θ describes the attacker's knowledge rather than a parameter governing the data generation process. We believe this is appropriate as it represents an unknown state of nature that we wish to protect from, e.g., minimax risk with respect to θ corresponds to a worst case scenario. We also typically use the empirical distribution when computing expectations, rather than a full probabilistic model. However, if a full probabilistic model exists, it may replace the empirical distribution. This interpretation of decision theory in the context of privacy is different from [22], [23] (see related work section) which assume a more traditional setting including a prior and posterior belief functions, both of which are focused on inferring the state of nature θ . We believe, however, that our approach is more appropriate to our problem, which is measuring personalized utility and risk associated with disclosing private information.

Returning to our example of data being phone-book entries, we have that $p(Z = 1)$ is the probability that a disclosed data (potentially suppressed or generalized) may be mapped to a specific individual, assuming identification is done by random selection. The loss $\ell(\delta(\mathbf{x}), \theta)$ is the expected sensitivity of the disclosed information times the identification probability. The risk measures the expected sensitivity of the disclosed information in the long run, as repeated trials and identification attempts are made.

3.3 Parametric Families of Sensitivity Functions

We now present several possible families of expressions for the data sensitivity function Φ . Since Φ is defined on the set \mathcal{X} of all possible records, defining it by a lookup table is impractical for a large number of attributes. We therefore consider several options leading to compact and efficient representations. Given a disclosed record $\mathbf{y} = \delta(\mathbf{x})$, perhaps the simplest meaningful form for Φ is a linear combination of non-negative weights $w_j \geq 0$ over the disclosed attributes

$$\Phi_1(\mathbf{y}) = \sum_{j: y_j \neq \perp} w_j \quad (2)$$

where w_j represents the sensitivity associated with the corresponding attribute A_j . A weight of $+\infty$ represents critically sensitive information that may only be disclosed if there is zero chance of it leading to identification.

In some cases, the data sensitivity significantly depends on the entity associated with the record. In other words attributes A_i may be highly sensitive for some records and less so

for other records. Recalling the assumption that one of the attributes, say y_1 , represents a unique identifier, we can construct the following personalized linear sensitivity function

$$\Phi_2(\mathbf{y}) = \sum_{j: y_j \neq \perp} w_{j,y_1}. \quad (3)$$

The weights $\{w_{j,r} : j = 1, \dots, n\}$ should be elicited from the different entities corresponding to the records or otherwise assigned by the database according to the group or cluster they belong to. Normalization constraints such as

$$\forall r \quad \sum_j w_{j,r} = c \quad \text{or} \quad \forall r \quad \sum_j w_{j,r} \leq c$$

can be enforced to provide all entities with similar privacy protection, or to make sure that no single entity dominates the privacy risk.

There are a number of ways to increase the flexibility of sensitivity functions beyond linear forms. One way to do so is by forming linear expressions containing k -order interaction terms, e.g., for $k = 2$

$$\Phi_3(\mathbf{y}) = \sum_{j>1: y_j \neq \perp} w_{j,y_1} + \sum_{j>1: y_j \neq \perp} \sum_{h>j: y_h \neq \perp} w_{j,h,y_1}. \quad (4)$$

Expressions containing k -order interaction terms use additional weights to capture interactions of at most k attributes that are not accounted for in the expressions (2) and (3). As k increases in magnitude, the class of functions represented by Φ becomes richer and, in the case of $k = m$, provides arbitrary flexibility. However, increasing k beyond a certain limit is impractical as both the number of weights specified by the users as well as the computational complexity associated with Φ_4 grow exponentially with k .

A possible alternative to the linear sensitivity function is a multiplicative function

$$\Phi_4(\mathbf{y}) = \exp \left(\sum_{j: y_j \neq \perp} w_{j,y_1} \right) = \prod_{j: y_j \neq \perp} e^{w_{j,y_1}} \quad (5)$$

in which case increasing one weight $w_{i,j}$ while fixing the others causes the sensitivity to increase exponentially in contrast to (2)-(4). The precise choice of the sensitivity function Φ (or Ψ) ultimately depends on the database policy and entities relating to the data. A simple expression such as (3) or (5) has the advantage of being easier to elicit and interpret.

In some cases the elicitation of the sensitivity function may not be easy. It is similar to eliciting prior from an expert in Bayesian statistics. However, it is necessary in order to provide some degree of customization which is important in privacy applications.

We also note that the sensitivity may be defined with respect to the the user target rather than the entity disclosing the data. The two entities may have different sensitivities and choosing one over the other depends on the context of the problem and whose risk we are really trying to minimize. The framework presented here applies with no modification regardless of whose sensitivity we are actually measuring or whose risk we are minimizing.

3.4 Data Suppression and Generalization and the Privacy Risk

A common practice in privacy preservation is to replace data records with suppressed or more general values [18, 19] in order to ensure anonymity and prevent the disclosure of

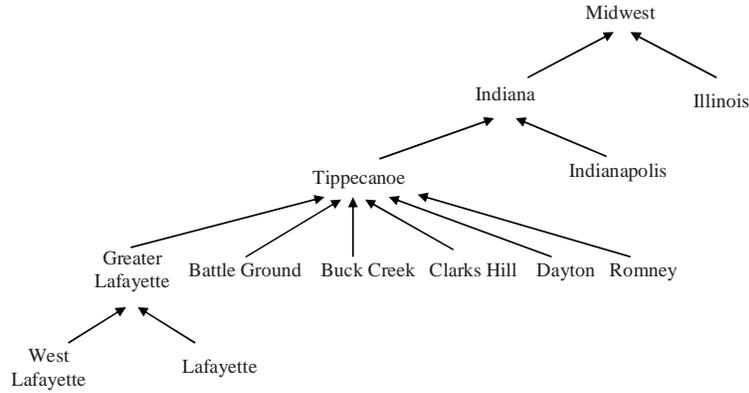


Figure 3: A partial value generalization hierarchy (VGH) for the address field

sensitive data. A disclosure policy $\delta : \mathcal{X} \rightarrow \mathcal{X}$ can suppress an attribute by assigning a \perp symbol to the appropriate attribute i.e. $[\delta(\mathbf{x})]_j = \perp$.

Assuming that the space \mathcal{X} is rich enough to contain the necessary generalizations, attribute value generalization may be accomplished by assigning a disclosed value that is more general than the original attribute value $x_j \prec [\delta(\mathbf{x})]_j$. Formally, we assume that Dom_i is a partially ordered set (S_i, \prec) whose smallest elements correspond to non-generalized attribute values and whose single maximum element is the ultimate generalized value, which we identify with the suppressed or missing value introduced earlier, \perp .

The partially ordered set Dom_i may be illustrated using its Hasse diagram in which every node correspond to a member of Dom_i and the edges correspond to the covering relation: x covers y if $y \prec x$ and $\nexists z : y \prec z \prec x$ [20]. Furthermore when drawing the Hasse diagram we draw more general nodes vertically higher than less general nodes. As an example, consider the attribute value representing a location or address and several levels of generalized values. A partial Hasse diagram representing the partial value generalization hierarchy for this attribute is illustrated in Figure 3. In this particular case, the Hasse diagram is relatively simple and is graphically described using a tree structure. More general examples and properties of partially ordered set may be found in [20].

Replacing an attribute value x_j by a more general value \hat{x}_j , i.e. $x_j \prec \hat{x}_j$, increases the set of entities consistent with that value in the attacker's dictionary θ , i.e.

$$x_j \preceq \hat{x}_j \implies \rho((x_1, \dots, x_j, \dots, x_m), \theta) \subseteq \rho((x_1, \dots, \hat{x}_j, \dots, x_m), \theta) \quad (6)$$

where $\rho(\mathbf{x}, \theta)$ is the set of entities in θ consistent with \mathbf{x} . Equation (6) indicates that as expected, generalizing an attribute value (which includes suppression as a special case) reduces the identification probability $p(Z = 1)$.

Equation (6) together with the assumption that the data sensitivity function Φ assigns smaller values to more general data ensures that the loss $\ell(\delta(\mathbf{x}), \theta)$ decreases with the amount of data generalization. The precise constraint on Φ depends on its parametric form such as expressions (2)-(5). For example, in the case of a personalized linear sensitivity (3), the appropriate constraints on the weights are

1. $w_a \geq 0$
2. $w_{a,r} \leq w_{b,r} \quad \forall a \preceq b \quad \forall r$.

3. $w_{\perp} = 0$.

The last constraint above is not crucial, but it ensures that fully suppressed data have zero sensitivity $\Phi(\perp, \dots, \perp) = 0$.

In summary, as we generalize or suppress data both the identification probability $p(Z = 1)$ and data sensitivity $\Phi(\delta(\mathbf{x}))$ decrease leading to lower loss $\ell(\delta(\mathbf{x}), \theta)$ and lower risk $R(\delta, \theta)$. Considering the disclosure risk $R(\delta, \theta)$ by itself leads to the conclusion that in order to minimize the risk the data needs to be completely suppressed or generalized. However, such a conclusion misses the point since it ignores the benefit obtained from the data disclosure. In order to appropriately appreciate the trade-off between the risk and benefit associated with private data disclosure we extend our discussion in the next section to include a quantification of the benefit associated with data disclosure.

4 The Optimal Disclosure Policies

Apart from incurring a privacy risk, disclosing private data $\delta(\mathbf{x})$ has some benefit, or else data would never be disclosed. We represent this benefit by a utility function $u : \mathcal{X} \rightarrow \mathbb{R}_+$ whose expectation

$$U(\delta) = E_{p(\mathbf{x})}(u(\delta(\mathbf{x})))$$

plays a similar but opposing role to the risk $R(\delta, \theta)$. While the loss $\ell(\delta(\mathbf{x}), \theta)$ may change from user to user, the utility is typically specified by the disclosing organization or the data recipient and is not user dependent.

The relationship between the risk and expected utility is schematically depicted in Figure 4 which displays disclosure policies δ on a 2-D plane using the corresponding risk R and expected utility U as coordinates (R, U) . The shaded region in the figure corresponds to the set of achievable disclosure policies, i.e., every coordinate (R, U) in that region corresponds to one or more policies δ realizing it. The unshaded region corresponds to un-achievable policies, i.e., there does not exist any δ with the corresponding risk and expected utility. The vertical line in the figure corresponds to all rules whose risk is fixed at a certain level. Similarly, the horizontal line corresponds to all rules whose expected utility is fixed at a certain level. Since the disclosure goal is to obtain both low risk and high expected utility, we are naturally most interested in disclosure policies occupying the boundary or frontier of the shaded region. Policies in the interior of the shaded region can be improved upon by projecting them to the boundary.

The vertical and horizontal lines suggest the following two ways of resolving the risk-utility tradeoff. Assuming that we cannot afford incurring risk higher than some acceptable level, we can define the optimal policy as

$$\delta^* = \arg \max_{\delta} U(\delta) \quad \text{subject to} \quad R(\delta, \theta) \leq c. \quad (7)$$

Alternatively, insisting on having expected utility no less than a certain acceptable level, we can define the optimal policy as

$$\delta^* = \arg \min_{\delta} R(\delta, \theta) \quad \text{subject to} \quad U(\delta) \geq c. \quad (8)$$

A more symmetric definition of optimality is given by

$$\delta^* = \arg \min_{\delta} R(\delta, \theta) - \lambda U(\delta) \quad (9)$$

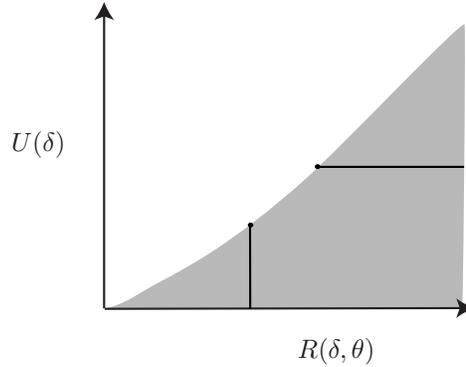


Figure 4: Space of disclosure rules and their risk and expected utility. The shaded region correspond to all achievable disclosure policies δ .

where $\lambda \in \mathbb{R}_+$ is a parameter controlling the relative importance of minimizing risk and maximizing utility.

The formation and interpretation of optimality depend on the situation at hand and are ultimately up to policy makers. We focus below on the case (8), but to simplify the notation we denote $\Delta = \{\delta : U(\delta) \geq c\}$ so that (8) becomes $\delta^* = \arg \min_{\delta \in \Delta} R(\delta, \theta)$. Solving (8) may often be computationally challenging as it is not easy to get a closed form definition of the constraint set $\Delta = \{\delta : U(\delta) \geq c\}$. More efficient computational search can usually be obtained by considering instead $\delta^* = \arg \min_{\delta \in \hat{\Delta}} R(\delta)$ where $\hat{\Delta} = \{\delta : \forall i, u(\delta(\mathbf{x}_i)) \geq c\} \subseteq \Delta$.

Solving the optimization problems (7)-(9) requires knowledge of the attacker's side information θ . Indeed, in some cases the attacker's side information is known - for example when θ constitutes national archives or some other publicly available dataset. In cases where the attacker's side information θ is unknown we can proceed instead using one of the following approaches.

Bayes Risk Replacing $R(\delta, \theta)$ in (7)-(9) with the Bayes risk $R(\delta) = E_{q(\theta)}(R(\delta, \theta))$ provides Bayesian-optimal policies that are independent of θ .

Estimating θ In some cases we can obtain an estimate of the attacker's side information $\hat{\theta}$. In these cases we can use expressions (7)-(9) with $R(\delta, \theta)$ replaced by $R(\delta, \hat{\theta})$. Mathematical analysis can be used to study the quality of the approximation $R(\delta, \hat{\theta}) \approx R(\delta, \theta)$ in terms of the approximation $\hat{\theta} \approx \theta$.

Worst Case Scenario In the absence of any information concerning θ we can use (7)-(9) with the worst case risk $\max_{\theta \in \Theta} R(\delta, \theta)$ instead of $R(\delta, \theta)$. The resulting policies, for example the minimax risk $\delta^* = \arg \min_{\delta \in \Delta} \max_{\theta \in \Theta} R(\delta, \theta)$, have the best worst case scenario.

Bounding the Risk This approach is described in the next section.

4.1 Bounding the True Risk by the Estimated Risk

In some of the cases where the attacker's side information θ^{true} is unknown, we can use a more specific side information $\hat{\theta}$, for example the relational database itself, to upper-bound

the true risk $R(\delta, \theta^{\text{true}})$ by the risk $R(\delta, \hat{\theta})$. The relation between the attacker's side information θ^{true} and the more generic version $\hat{\theta}$ is formally defined in Definition 5. Intuitively the relation indicates that all entries in $\hat{\theta}$ may be found in θ^{true} where they may appear with less, or more general, attribute values.

Definition 5. We define a partial order relation \ll between dictionaries $\theta = (\theta_1, \dots, \theta_{l_1})$ and $\eta = (\eta_1, \dots, \eta_{l_2})$ by saying that $\theta \ll \eta$ if for every $\theta_i, \exists \eta_j$ such that $\eta_{j1} = \theta_{i1}$ and $\forall_k \theta_{ik} \preceq \eta_{jk}$.

Above, we considered dictionaries $\theta = (\theta_1, \dots, \theta_{l_1})$ 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.

It is often the case that the situation described above holds with $\hat{\theta}$ representing proprietary or organizational database unknown to the attacker and the attacker's θ^{true} corresponding to a public resource such as a phone-book or some other generic listing. Since θ^{true} is a more general purpose listing than $\hat{\theta}$ it has records corresponding to all records in $\hat{\theta}$. Furthermore, since $\hat{\theta}$ is more specific than the general purpose listing θ^{true} the entries in $\hat{\theta}$ will be more specific and will have less missing values than θ^{true} . The motivation for considering a more specific (but known) $\hat{\theta}$ instead of the unknown θ^{true} is that it leads to the following upper bound of the privacy risk.

Proposition 1. If $\hat{\theta}$ contains records that correspond to x_1, \dots, x_n and $\hat{\theta} \ll \theta^{\text{true}}$, then

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

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

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

The probability of identifying $\delta(x_i)$ by the attacker is thus smaller than the identification probability based on $\hat{\theta}$ and it follows that

$$\forall_i \quad \ell(\delta(x_i), \theta^{\text{true}}) \leq \ell(\delta(x_i), \hat{\theta}) \quad \Rightarrow \quad R(\delta, \theta^{\text{true}}) \leq R(\delta, \hat{\theta}).$$

□

4.2 Independence and Integrity Constraints

Computing and minimizing the risk may be computationally demanding in the general case. In this section we discuss how the independence assumption or its relaxation by introducing integrity constraints affects such computational efficiency considerations. Section 8 describes using branch and bound, a discrete approximations method, in order to further accelerate the computation and minimization of the privacy risk.

The assumption that different attributes are statistically independent is somewhat questionable but still often used in high dimensions due to its practicality. For instance, returning to the simple phone-book example, the independence assumption may imply that the

popularity of first names does not depend on the popularity of last names, e.g.,

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

Clearly, some attributes are strongly correlated while others are generally believed to be independent. The introduction of integrity constraints to model correlated attributes (e.g., $\{\text{profession} = A\} \Rightarrow \{\text{salary} \in C\}$) while assuming independence between uncorrelated attributes is an effective relaxation of the complete independence assumption. We first discuss the implication of complete independence to risk computation and then proceed to consider the presence of integrity constraints.

Under the assumption of statistical independence on the attribute values the identification distribution factors as a product

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

for some appropriate functions α_j . As a result the loss function (assuming a parametric multiplicative form) decomposes to

$$\begin{aligned} \ell(\mathbf{y}, \theta) &= \frac{\prod_{j \in C_2(\mathbf{y})} e^{w_j, y_1}}{|\rho(\mathbf{y}, \theta)|} = \frac{\prod_{j \in C_2(\mathbf{y})} e^{w_j, y_1}}{\prod_{k > 1} \alpha_k(y_k, \theta)} = \prod_{j \in C_2(\mathbf{y})} \frac{e^{w_j, y_1}}{\alpha_j(y_j, \theta)} \cdot \prod_{l \in C_1(\mathbf{y})} \frac{1}{\alpha_l(\perp, \theta)} \\ &= \prod_{j \in C_2(\mathbf{y})} e^{w_j, y_1} \frac{\alpha_j(\perp, \theta)}{\alpha_j(y_j, \theta)} \cdot \prod_{l=2}^m \frac{1}{\alpha_l(\perp, \theta)}. \end{aligned}$$

where $C_1(\mathbf{y}) = \{j : j > 1, y_j = \perp\}$ and $C_2(\mathbf{y}) = \{j : j > 1, y_j \neq \perp\}$.

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 elements of $\{e^{w_j, y_1} \frac{\alpha_j(\perp, \theta)}{\alpha_j(y_j, \theta)}\}_{j=2}^m$ which may be efficiently computed in time $O(nNm)$ where n, N, m are the number of disclosed records, dictionary size, and number of attributes, respectively.

Extensions of the above decomposition are straightforward when the attributes can be divided to several clusters satisfying statistical dependence for attributes within the same cluster and statistical independence for attributes belonging to different clusters. An alternative decomposition for more general integrity constraints or statistical dependencies may be obtained through the product factorization of graphical models in statistics, e.g., [25].

5 Privacy Risk and k -Anonymity

k -Anonymity [18] has recently received considerable attention by the research community [29, 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.

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 refer to a privacy loss. However, a closed examination reveals that k -anonymity implicitly makes strong assumptions whose presence may undermine its original motivation. Following the formal presentation of k -anonymity in the privacy risk context, we analyze these assumptions and their possible relaxations.

Since the k -anonymity requirement is enforced on the relation T (the database containing the original records) the anonymization algorithm considers the attacker's side information θ^{true} as equal to the database T . Representing the k -anonymity rule by δ_k^* we have that the following expression of k -anonymity constraints

$$\forall i \quad |\rho(\delta_k^*(\mathbf{x}_i), T)| \geq k. \quad (10)$$

Since k -anonymity is concerned only with satisfying the constraints (10) and ignores the role of the data sensitivity we consider its sensitivity function to be constant $\Phi \equiv c$.

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 an optimal risk minimizer policy.

Proposition 2. 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 proposition implies, a 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. Viewed as a privacy risk minimizer, we can examine the now explicit assumptions behind k -anonymity:

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

The first assumption may be taken as an indication that k -anonymity simply assumes that the database relation T is available as side information to the attacker. This assumption can be expanded as described earlier by assuming an estimated $\hat{\theta}$, using a Bayesian averaging, worst case risk $\max_{\theta \in \Theta} R(\delta, \theta)$ or that θ^{true} is a publicly available resource. Such adaptation of k -anonymity are likely to more faithfully protect privacy and yet should not require a major conceptual change to the k -anonymity framework.

The second assumption of the sensitivity function $\Phi \equiv c$ being constant is a result of k -anonymity's singular attention to protection from identification. In other words, disclosing data incurs the same loss regardless of the data itself and the entity to whom the data pertains, as long as there exists a certain protection from identification. This is a problematic assumption since under imperfect identification protection, the notion of privacy

preservation is not synonymous with identification. In cases where a positive probability of identification exists, the nature of the disclosed data and in particular its sensitivity with respect to the the entity it relates to should play a crucial role.

As a simple example, consider facing two possible disclosure options: the disclosure of data containing a substantial medical diagnosis (e.g., HIV positive) and the disclosure of data containing a recent grocery shopping transaction. Intuitively, disclosing the first data would lead to a greater privacy violation than the second data under non-zero identification probability. However k -anonymity, assuming a constant sensitivity function, considers the disclosure of both data equally harmful if they provide similar identification protection. Furthermore, it may favor the disclosure of very sensitive data over non-sensitive data as long as it provides a slightly better identification protection.

Section 8 presents a case study illustrating this point further in the context of a commercial organization's customer transaction database. For a diverse commercial organization, transactions should be classified according to varying sensitivity levels. k -Anonymity protection would exert undesired privacy protection in some areas while lacking in other areas. The privacy risk framework presented in this paper provides a natural extension to k -anonymity by making Φ non-constant. The resulting privacy loss combines data sensitivity and identification protection in quantitative probabilistic manner.

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.

6 Privacy Risk and Record Linkage

We have thus far discussed the usage of a dictionary to identify the entity associated with a disclosed record. The re-identification process can be whether an exact linkage or an approximate linkage. The latter type of linkage is better known as record linkage [9]. Though we have not explicitly mentioned record linkage, our framework does include the possibility of performing linkage in both ways.

However, in this section we provide some further considerations that are intended to extend our framework to also consider additional information that the attacker may have as a consequence of the linkage. Indeed, the linkage, if successful, enlarges the available information thereby influencing both the data sensitivity and subsequent identification probability. The probabilistic framework of the privacy risk can be naturally extended to account for such cases. Figure 1 illustrates the linkage process in the context of the privacy risk framework.

We say that the linkage of the disclosed data $\delta(\mathbf{x})$ and public record z is successful if $\delta(\mathbf{x})$ and z are records that relate to the same entity. The linkage of $\delta(\mathbf{x}_i)$ and z creates an enlarged set of attributes $\delta(\mathbf{x}_i) \vee z$ combining information from both sources, which if successful, improves identification based on a dictionary.

The disclosed record $\mathbf{y} = \delta(\mathbf{x})$ and the linked record z are random variables with a joint distribution $p(\mathbf{x}, z) = p(\mathbf{x})p(z|\mathbf{y})$ where $p(\mathbf{x})$ may be the empirical $\tilde{p}(\mathbf{x})$ described earlier and the conditional $p(z|\mathbf{y})$ is the probability of linking record z with record \mathbf{y} . In this case,

it is important to estimate the linking probability based on what a sensible attacker might do. In the case of linking $\delta(\mathbf{x}_i) \vee \mathbf{z}$, the risk is

$$R_{\text{link}}(\delta, \theta) = E_{p(\mathbf{x})p(\mathbf{z}|\mathbf{x})}(\ell(\delta(\mathbf{x}) \vee \mathbf{z}, \theta))$$

where the loss function $\ell(\delta(\mathbf{x}_i) \vee \mathbf{z})$ can be structured in a similar way to our previous discussion. The expectation in the definition of R_{link} is with respect to a joint distribution over x, y defined by $p(x, z) = p(x)p(z|x)$ as described above. Introducing a binary random variable W representing successful linking we have that the loss incurred under successful linking and identification is equal to the sensitivity of the enlarged data $\Phi(\delta(\mathbf{x}) \vee \mathbf{z}, \theta)$ is $\ell(\delta(\mathbf{x}) \vee \mathbf{z}, \theta)$ provided that $\{W = 1, Z = 1\}$. Continuing as before, we can define the loss $\ell(\delta(\mathbf{x}) \vee \mathbf{z}, \theta)$ as the expectation of the sensitivity taking into consideration probabilities of successful linking and identification.

7 Applications and Experiments

In this section, we define two operators that implement disclosure rules on relations (Section 7.1), and then proceed to illustrate some experiments further validating our framework (Section 7.2). The following section contains experiments outlining a particular case study.

7.1 Implementation of Disclosure Rules

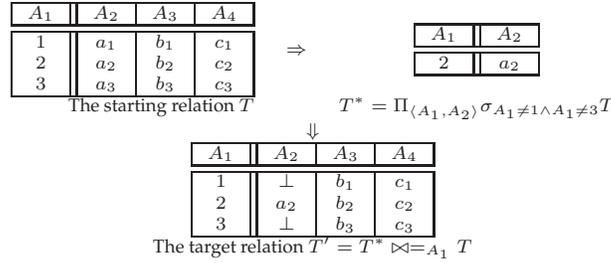
As described in Section 3.4, disclosure rules can lead to data generalization or even to data suppression. In this section we describe two operators that implement disclosure rules in a relational setting. In particular, we show that we can define such operators by relying on relational algebra, thus enabling the usage of relational technology. The definition in relational algebra allows us to prove some interesting properties as well as to obtain consistent advantages in terms of standardization and efficiency of the disclosure rules implementation. More specifically, the algebraic specification of the operators paves the way for the realization of optimization strategies within relational technology.

As previously mentioned, we consider the relation $T(A_1, \dots, A_m)$ and the set $\mathcal{X} = \text{Dom}_1 \times \dots \times \text{Dom}_m$ with the attribute A_1 representing a record identifier. We further assume that F is a formula involving: (i) a set of operands that are either variables or constants, (ii) the set of arithmetic operators $\{=, \neq, <, \leq, >, \geq\}$, and (iii) the set of logical operators $\{\vee, \wedge, \neg\}$. Such formulas are used to specify disclosure conditions, i.e. to identify the tuples of a relation which are subject to privacy constraints and on which disclosure rules must be enforced. We first define the operator $hide_T$ that enables attribute-level suppression on the table T .

Definition 6. Given an attribute A_j of the relation T , and a formula F , the operator $hide_T$ is defined as

$$hide_T(A_j, F) = (\Pi_{\langle A_1, A_j \rangle} \sigma_F T) \bowtie_{=A_1} T$$

First, the operator selects the tuples satisfying the condition F in the relation T . The selection σ_F specifies which tuples do not have privacy requirements on the attribute A_j . Second, the projection $\Pi_{\langle A_1, A_j \rangle}$ builds a partial result used to recompose the original relation, with \perp values replacing the values of A_j to be kept private. For this latter step, the right outer join operator $\bowtie_{=}$ is applied over the record identifier A_1 , and is used to introduce the \perp values wherever specified by the disclosure conditions. Note that the outer join operator

Figure 5: The $hide_T(A_2, F' \wedge F'')$ operator

is used in cases when it is required that the resulting relation contains all tuples from both relations, even if they do not participate in the join. In such cases they are padded with \perp values [17].

The following proposition formally states the commutative and associative properties of the operator, that are directly derived from properties of relational algebra.

Proposition 3. The operator $hide_T$ is commutative and associative with respect to disclosure conditions. Given two disclosure conditions F' and F'' with respect to the T -attribute A_j , and a disclosure condition F''' with respect to the T -attribute A_h :

- $hide_{hide_T(A_j, F')}(A_h, F''') = hide_{hide_T(A_h, F''')}(A_j, F')$;
- $hide_{hide_T(A_j, F')}(A_j, F'') = hide_T(A_j, F' \wedge F'')$.

The commutative property is particularly relevant as it means that the order according to which disclosure rules are enforced on the original relation is not significant.

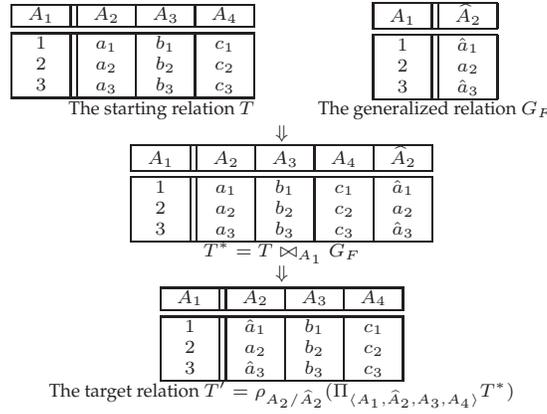
In the following, we provide an example of how the $hide_T$ operator is applied in order to enforce suppression disclosure rules. Suppose that we want to enforce the rules: (i) “For tuple 1, the attribute value a_1 of A_2 is private” and (ii) “For tuple 3, the attribute value a_3 of A_2 is private” on the relation T shown in Figure 5. Two disclosure conditions F' and F'' are formulated as $(A_1 \neq 1)$ and $(A_1 \neq 3)$, respectively. Notice that the disclosure conditions are formulated such that they exclude the tuples requiring privacy enforcement on a specific attribute from the partial result. Due to the associativity property of the disclosure conditions, we can apply the operator $hide_T(A_2, F' \wedge F'')$ according to the steps shown in Figure 5, thus obtaining the resulting relation T' .

We next define the operator gen_T that implements the generalization disclosure rules on a table T .

Definition 7. Given an attribute A_j of the relation T , a formula F , and a table $G_F(A_1, \widehat{A}_j)$ that contains the generalization values for A_j . Such values are supposed to have been chosen from the domain generalization hierarchy of A_j corresponding to tuples selected by F , i.e., tuples affected by privacy constraints. The operator gen_T is defined as follows:

$$gen_T(A_j, F) = \rho_{A_j/\widehat{A}_j}(\Pi_{\langle A_1, \dots, A_j, \dots, \widehat{A}_j \rangle - \langle A_j \rangle}(T \bowtie_{A_1} G_F))$$

Notice that the relational algebra operator ρ_{A_j/\widehat{A}_j} is used for *renaming* \widehat{A}_j as A_j . In Figure 6, we provide an example of how the gen_T operator is applied in order to enforce generalization disclosure rules. We modify the example shown in Figure 5 such that the

Figure 6: The $gen_T(A_2, F' \wedge F'')$ operator

rules are: (i) “For tuple 1, the attribute value a_1 of A_2 is to be released as \hat{a}_1 ” and (ii) “For tuple 3, the attribute value a_3 of A_2 is to be released as \hat{a}_3 ”. The two disclosure conditions F' and F'' do not change and are $(A_1 \neq 1)$ and $(A_1 \neq 3)$, respectively. We apply the operator $gen_T(A_2, F' \wedge F'')$ according to the steps shown in Figure 6, thus obtaining the resulting relation T' . The generalized relation G_F is supposed to contain generalized values for A_2 in correspondence to tuples selected by F' and F'' . The (i) join of T and G_F , (ii) projection on all attribute except A_2 , and (iii) renaming of \hat{A}_2 as A_2 are performed in this order to obtain the disclosed table T' .

Finally, we notice the following properties of the gen_T operator:

- It is easy to verify that the commutative and associative properties proved for the $hide_T$ operator are also valid for the gen_T operator.
- Suppression as a special case of generalization is coherent with the semantics specified for the gen_T operator. Indeed, if the table G_F includes \perp symbols, then such symbols will be released.

7.2 Experiments

The goals of our experiments are three-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 database obtained from Wal-Mart. The database represents an `item description` table of more than 400,000 records each with more than 70 attributes. Part of the table is used to represent the disclosed data whereas the whole table is used to generate a 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 [10] 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. These random weights could be,

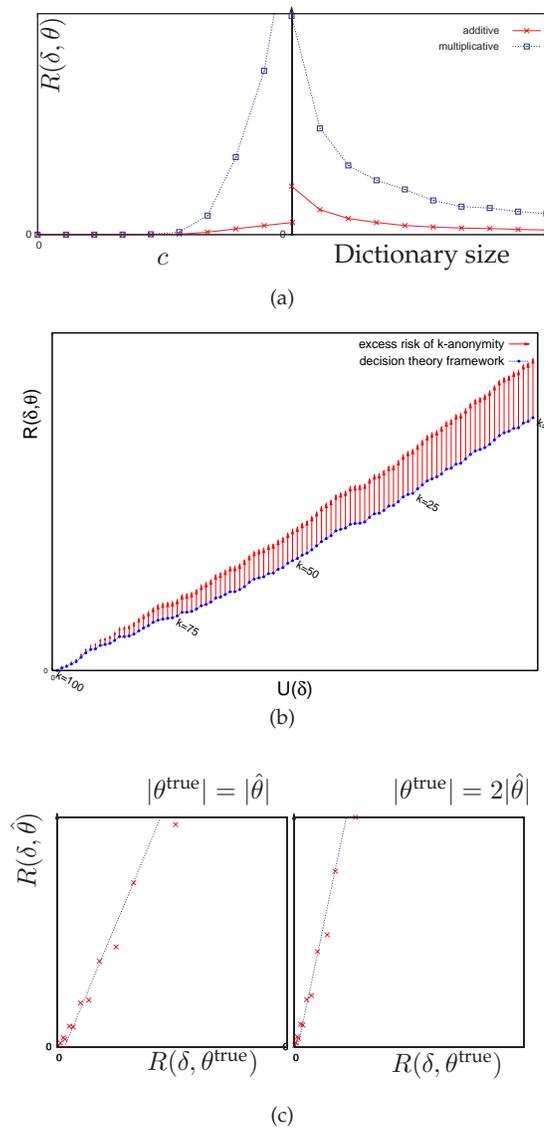


Figure 7: (a) The risk associated with different dictionaries and c values, (b) a comparison between our decision theory framework and k -anonymity, and (c) the relationship between the true risk and the estimated risk.

without loss of generality, replaced with more realistic values if the organization to which data belongs provides such an information.

We use a simplified utility function $u(\mathbf{y})$ to capture the information benefit of releasing a record \mathbf{y} : Data usefulness is intuitively measured by how far it is from the root. The farther the node is from the root, the more specific it is, and the more information it provides. Roots carry no information whereas leaves carry the most useful information. For simplicity, the distance of each node from the root is equivalently computed by subtracting the height of the DGH minus the distance from the leaf. Therefore,

$$u(\mathbf{y}) = \sum_{i=1}^m \text{Dist}(\text{Root}_{DGH_i}, y_i)$$

where $\text{Dist}(\text{Root}_{DGH_i}, y_i)$ represents the sum of all heights in the domain generalization hierarchies or DGH (see Figure 8) minus the number of generalization steps that were performed to obtain the record \mathbf{y} . For each record x_i , the minimum risk is obtained subject to the constraint set $\hat{\Delta} = \{\delta : \forall_i u(\delta(x_i)) \geq c\}$. The impacts of the parameter c and the dictionary size on the privacy risk are reported in Figure 7(a). As c increases from 0 to 10 (i.e., more specific data is being disclosed) and by fixing the dictionary size the identification probability, as well as the privacy risk increase.

On the other hand, fixing c at a certain value $c = 8$, the relation between the risk and dictionary size is inversely related. As the attacker's dictionary size increases, more consistent records are found reducing the identification probability. The different dictionaries were generated from the original table with sizes varying from 10% to 100% of the size of the whole table. Interestingly, the experimental data shows that the multiplicative sensitivity model is always superior to the additive model. The superiority of multiplicative model over the additive model comes from the observation that the former always yields a higher modeled risk. A conservative user would always be more alarmed if this model is used.

We compare the risk and utility associated with a disclosed table based on our decision theory framework and arbitrary k -anonymity rules for k from 0 to 100. In Figure 7(b) we compare the utility and risk of optimally selected disclosure policies and standard k -anonymity rules (averaged over a random selection of 10 k -anonymity rules). The optimal disclosure policies consistently outperform standard k -anonymity rules. The arrows in the figure represent the difference in risk between both approaches which increases as k increases.

The relationship between the true risk $R(\delta, \theta^{\text{true}})$ and the estimated risk $R(\delta, \hat{\theta})$ is illustrated in the scatter plot in Figure 7(c). All the points occur above the line $y = x$ which agrees with our result of upper bounding the $R(\delta, \theta^{\text{true}})$ by $R(\delta, \hat{\theta})$. 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 on the true risk.

8 Case Study: An Organization Releasing Customers' Data

We demonstrate the practical applicability of the privacy risk framework in this section by describing its application to customer databases of commercial organizations. In such cases, the organizations often benefit from disclosing customer records, for example due to outsourcing its datamining efforts or otherwise sharing customer records with partnering organizations. The initial customer suspicion towards sharing their records, is often relaxed

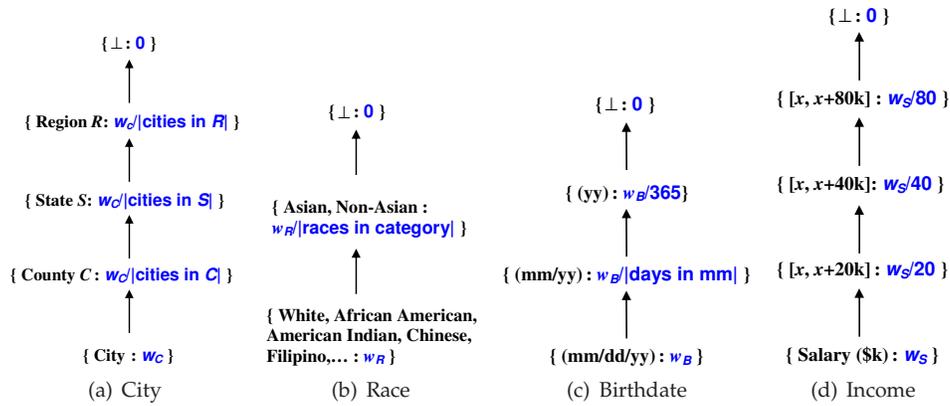


Figure 8: Domain Generalization Hierarchies (DGHs) with the Associated Sensitivity Weights

by offering them benefits such as loyalty cards or some other discount plans in return for their participation.

Willing participants may rate the privacy of various parts of their data as non-private, semi-private and very-private. For example, consider a customer database such as Amazon.com where some of the customer transactions are non-private, some are semi-private, and some are very private (for example a purchase disclosing a health condition or an embarrassing area of interest). The organization will treat the transactions according to the user specified sensitivity which will in turn constitute a user specified loss function. To prevent customers from registering all transactions as very-private the organization may enforce constraints on the supplied loss, which must be obeyed in order to participate in the discount plan. It is the organizations' responsibility to determine at which level of generalization hierarchy each attribute is to be disclosed such that: (i) minimizing the inherited risk associated with violating customers' privacy (e.g., the potential law suit resulting from releasing very-private information) and (ii) maximizing (or at least establishing a floor for) the benefit of the released information. Note that in this example, the utility of disclosing data may be quantified by the monetary amount the organization can expect to obtain, for example by selling the data or by projected increase in efficiency due to data mining activity.

An illustrative scenario is described as follows. Suppose that Wal-mart needs to assess the risk associated with releasing its members' data while maximizing its benefit. To carry out our experiments, a projected Wal-mart members table of 5,667,004 records is used. The projection contains 4 attributes: City, Race, Birthdate, and Household Income. Figure 8 depicts the used domain generalization hierarchies for these 4 attributes with the associated sensitivity weights computed as follows.

First of all, we assume, without loss of generality, that all leaf nodes belonging to the same DGH are equally sensitive. We use a modified harmonic mean¹ to compute the sensitivity of a parent node w_p with l immediate children given the sensitivities of these children

¹Different means may lead to alternative ways to define the sensitivity function. For the sake of experimentation, we elect to use one of these means.

$w_i, 1 \leq i \leq l$:

$$w_p = \frac{1}{\sum_{1 \leq i \leq l} \frac{1}{w_i}}$$

with the exception that the root node (corresponding to suppressed data) has a sensitivity weight of 0. Clearly, the modified harmonic mean satisfies the following properties: (i) the sensitivity of any node is greater than or equal 0 provided that the sensitivity of all leaves are greater than or equal 0, (ii) the sensitivity of a parent node is always less than or equal (in case of 1 child) the sensitivity of any of its descendent nodes, and (iii) the higher the number of children a node has the lower the sensitivity of this node. For example, given a constant city weight w_c , the weight of the County node j in the DGH for the City is $\frac{1}{\sum_{1 \leq i \leq l_j} \frac{1}{w_c}} = \frac{w_c}{l_j}$, where l_j is the number of cities in the county j . Moreover, the sensitivity of the State node in the same DGH is $\frac{1}{\sum_{1 \leq j \leq m} \frac{1}{w_c/l_j}} = \frac{w_c}{\sum_{1 \leq j \leq m} l_j} = \frac{w_c}{n}$, where m is the number of counties in the state and $n = \sum_{1 \leq j \leq m} l_j$ is the number of cities in the state.

The multiplicative form $\Phi_4(\mathbf{y})$ of the sensitivity function is used to compute the overall sensitivity of a released record. The weights w_c, w_r, w_b , and w_s are set at the values 0.3, 0.4, 0.5, and 0.75, respectively. Therefore, the sensitivity associated with the record (West Lafayette, White, 05/10/1975, \$52k), for example, is $e^{0.3+0.4+0.5+0.75} = 7.03$, whereas the sensitivity associated with the record (West Lafayette, \perp , May 1975, [\$20k, \$99k]) is $e^{0.3+0+\frac{0.5}{31}+\frac{0.75}{80}} = 1.38$.

As a dictionary θ we use the Adult database² which is comprised of 9,857,623 records extracted from US Census data. The database contains 5 attributes: Age, Gender, Zipcode, Race, and Education. Each record \mathbf{y} (and its generalizations) from Wal-mart members table is matched with this dictionary to identify the number of dictionary records consistent with it $\rho(\delta(\mathbf{y}))$. The matching process is performed on the corresponding attributes representing age, race, and address in both tables. For example, the record (West Lafayette, White, 05/10/1975, \$52k) has 7 dictionary records consistent with it, whereas the record (West Lafayette, \perp , May 1975, [\$20k, \$99k]) has 198 dictionary records consistent with it.

The loss function associated with releasing a record \mathbf{y} is $\ell(\mathbf{y}, \theta) = \frac{\Phi(\mathbf{y})}{|\rho(\mathbf{y}, \theta)|}$. For example, from the above results, the loss associated with releasing the record (West Lafayette, White, 05/10/1975, \$52k) is $7.03/7 = 1.004$, whereas the loss associated with releasing the record (West Lafayette, \perp , May 1975, [\$20k, \$99k]) is $1.38/198 = 0.007$. The overall risk associated with releasing a whole table is computed as the average loss associated with releasing its individual records.

We use the same utility function explained in Section 7.2. For example, the utility function corresponding to the record (West Lafayette, White, 05/10/1975, \$52k) is $4 + 2 + 3 + 4 = 13$ or equivalently $(4 + 2 + 3 + 4) - (0) = 13$, whereas the utility function corresponding to the record (West Lafayette, \perp , May 1975, [\$20k, \$99k]) is $4 + 0 + 2 + 1 = 7$ or equivalently $(4 + 2 + 3 + 4) - (0 + 2 + 1 + 3) = 7$.

By following the procedure explained above, the organization goal to determine the disclosure rule that yields the minimal risk while maintaining the utility above a certain threshold is achievable. For each potentially disclosed table T , our model can be applied to assess both the risk and utility associated with releasing this table. As the case with the risk, the utility of a given table is the average utility of all individual records constituting this table rounded to the nearest integer. The table that poses the minimal risk with an acceptable utility is released.

²Downloaded from <http://www.ipums.org>

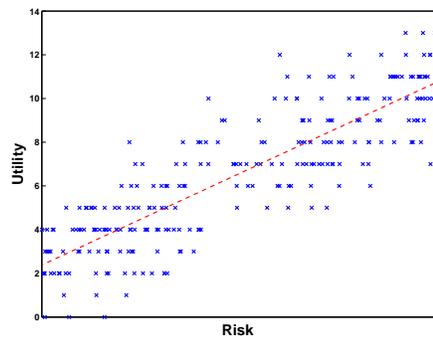
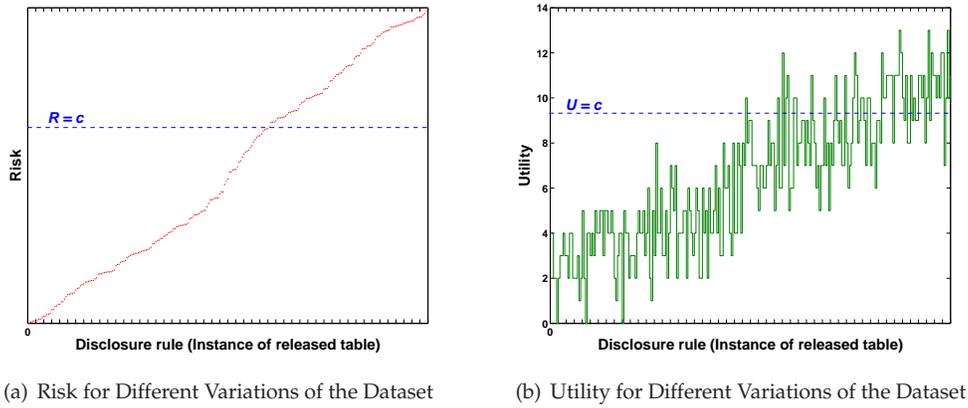


Figure 9: Risks and Utilities for Different Disclosure Rules

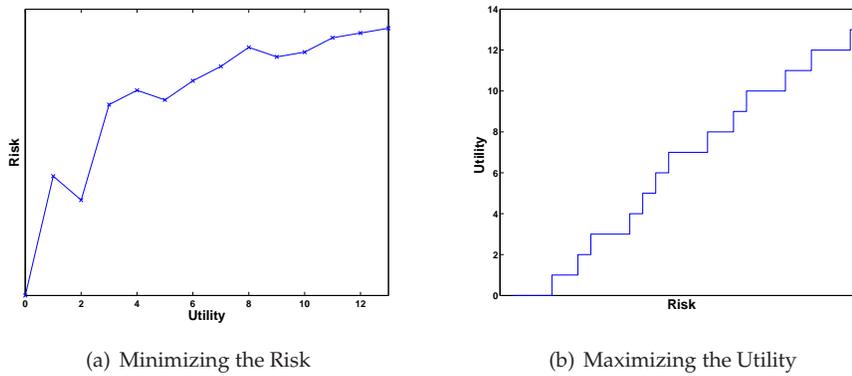


Figure 10: The Optimal Disclosure Rule

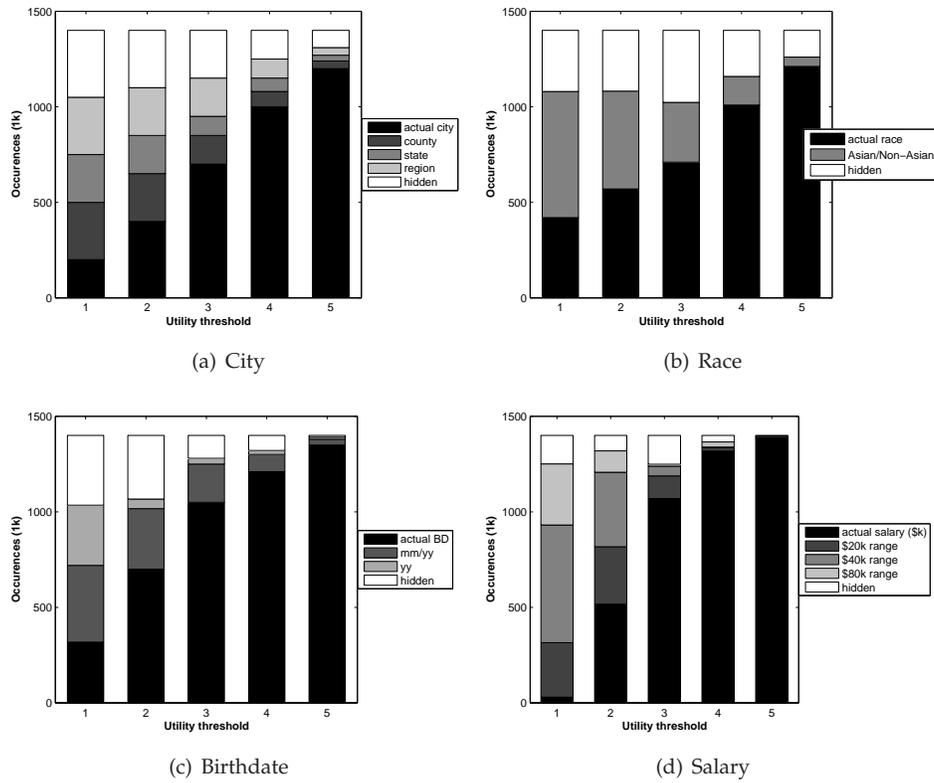
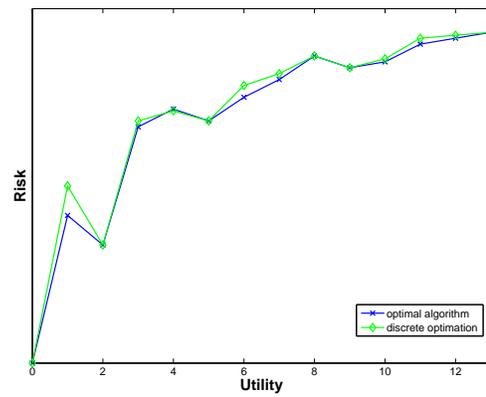
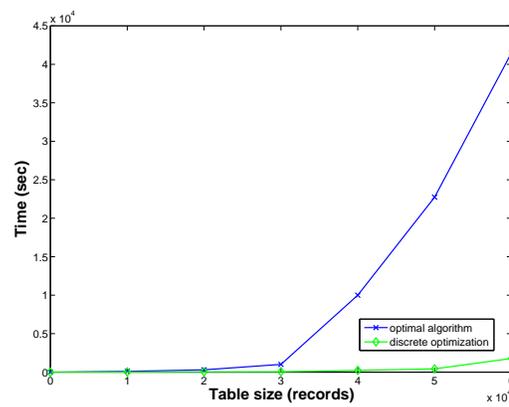


Figure 11: Effect of Utility Threshold on the Level of Attribute Disclosure



(a) Risk



(b) Time

Figure 12: The Discrete Optimization Algorithm

Figure 9 shows some plots of the risks and associated utilities for various disclosure rules. Recall that a disclosure rule $\delta_i(T)$ (or simply δ_i) is a combination of transformations (suppression, generalization, and disclosure of actual data) performed on the attributes of the original table T which results in the disclosed table T'^3 . Figures 9(a)(b) plot the computed risks (in increasing order) and the corresponding utilities for random instances of the released table T' , $\mathcal{R} = \{(\delta_i, R(\delta_i, \theta)), i = 1, 2, \dots\}$ and $\mathcal{U} = \{(\delta_i, U(\delta_i)), i = 1, 2, \dots\}$, respectively. As expected the utility increases as the risk increases. However, sometimes this is not the case due to the settings of the sensitivity weights and the topologies of different DGHs. The scatter plot in Figure 9(c) depicts the high positive correlations between risk and utility with a computed correlation coefficient of 0.858.

Had the organization goal been focusing only on one factor (i.e., minimize the risk or maximize the utility), these 2 curves would have been sufficient to identify the optimal disclosure rule. However, it is often the case that the goal is to optimize one of the factors while maintaining an acceptable level of the other factor. Figure 10 shows how the optimal disclosure rule is determined for the example at hand. By using different values for the constant c and obtaining the minimum risk, Figure 10(a) can be plotted. It shows the optimal risk (and accordingly the optimal disclosure rule) that yields utility $U \geq c$. Specifically, it helps determine $\delta^* = \arg \min_{\delta} R(\delta)$ subject to $U(\delta) \geq c$. Likewise, by fixing the risk at different values for the constant c and obtaining the maximum utility, we obtain Figure 10(b) which shows that the optimal utility (and accordingly the optimal disclosure rule) poses a risk $R \leq c$. Specifically, it helps determine $\delta^* = \arg \max_{\delta} U(\delta)$ subject to $R(\delta) \leq c$.

We implemented a heuristic discrete optimization algorithm, Branch and Bound [12], to obtain the heuristic optimum disclosure rule. Figure 12 shows that the discrete optimization algorithm is superior in terms of execution time compared to the brute-force algorithm with no significant risk increase.

Figure 11 shows some statistics about the frequencies of generalization steps carried out on each attribute at different utility levels to obtain the optimal risk. For instance, when setting the utility $U \geq 5$, Figure 11(d) indicates that the actual salaries of almost all members are released. Clearly, the tendency towards releasing the actual data increases as the utility level increases. Moreover, depending on attribute settings, the level of aggressiveness with which the tendency to release the actual data occurs varies. The statistics shows, for instance, that most of the time the actual birthdate is released when the required utility $U \geq 3$. An organization that is willing to apply the disclosure rule which has been applied the most may elect to release the actual birthdate for a newly added record when the utility is sought to be no less than 3.

9 Related Work

The task of avoiding privacy violations has been investigated by different communities. Nevertheless, to the best of our knowledge, our approach is the first in providing a comprehensive theoretical framework for assessing privacy risks. Our framework, published initially in [13], is based on statistical decision theory and is a highly flexible tool for modeling the trade-off between disclosure benefits and risks. It incorporates into a probabilistic framework the notion of personalized data sensitivity, identification model and attacker's side information. The connection between personalized privacy preservation and decision theory may lead to further theoretical insights and practical techniques.

³An example of T' is $\{(West\ Lafayette, White, 05/10/1975, \$52k), (Indiana, Asian, 1948, \perp), (\perp, Chinese, August\ 1965, [\$20k, \$40k]), \dots\}$.

Section 5 relates the privacy risk framework to k -anonymity and includes a discussion of related work in that context. We describe here additional related work in the statistical databases and data mining literatures. In statistical databases, queries result in some statistical information, for example the average of a set of values. The techniques for preserving privacy can be divided into two categories: (i) query restriction and (ii) input-output perturbation. Query restriction methods pose limitations on query parameters while input-output perturbations alter the data by introducing noise to either the data or the query results. Unlike statistical databases that are concerned with disclosing statistical data summaries, our framework focuses on disclosing elementary data and thus incorporates a broader class of queries. Moreover, while recent proposals in statistical database [4, 2] focus on the tradeoff between meaningfulness of information and privacy loss, we are interested in the fundamentally different tradeoff between disclosure benefit and privacy loss. In the data mining area, several approaches have been developed for privacy preserving data mining. Unlike our framework, such approaches (e.g., [8]) are based on perturbing the original data and at the same time achieving correct data mining results.

Duncan et al. [6] describes a framework, called Risk-Utility (R-U) confidentiality map, which addresses the tradeoff between data utility and disclosure. [11] propose an approach to risk analysis for disclosed anonymized data based on modeling a database as a series of transactions and the attacker's knowledge as a belief function. Our model is fundamentally different since we deal exactly with relational instances rather than data frequencies, we do not consider simply anonymized data and we incorporate the concept of data sensitivity into our framework. [16] provide a measure of the privacy risk in the context of the query-view security problem but such measure does not result in a complete framework for privacy risk assessment.

Another decision theory based approach to privacy measurement was proposed by Trotini and Fienberg [22], [23]. In that approach the risk is based on the deviation of the posterior from the prior. Despite the fact that both [22], [23] and our approach are based on decision theory, there are some fundamental differences. In our approach the risk is based on a user defined sensitivity function which enables personalization without the need for specifying a prior distribution and computing a posterior distribution. We also introduce both a utility and loss function which take complementary roles (but neither is determined by the other). The state of nature θ now denotes the attacker's knowledge as opposed to a parameter governing the data generation process.

These differences reflect the somewhat non-classical approach to decision theory taken in this paper. The role of θ , the use of the empirical distribution in definition of the risk, and the complementary roles of the utility and loss function are somewhat different than traditional decision theory settings. Nevertheless, we believe that our setting is more appropriate to our problem, which is measuring personalized utility and risk associated with disclosing private information.

10 Discussion

In this paper we describe a novel framework for quantitatively assessing and optimizing privacy risk in a variety of situations. Our framework combines both the identification and data sensitivity components with the latter being a personalized subjective measure. This framework substantially differs from the recent alternatives proposed in the literature which usually neglect the role of data sensitivity. Moreover, only little of the recent work are based on a probabilistic identification model [5, 28].

The quantitative framework allows defining and computing optimal disclosure policies, where optimality is defined in a number of alternative ways. The computation and optimization of the risk are described in the context of exact knowledge, partial knowledge, and no knowledge whatsoever with respect to the attacker's side information. While computing the optimal policy may require massive computation for large datasets, we demonstrate a discrete approximation method based on branch and bound that closely approximates the optimal policy while being computationally efficient.

We demonstrate the generality of our framework by showing that k -anonymity is a special case of it and we have highlighted the decision theory based assumptions underlying k -anonymity. Viewed in this perspective, the various often implicit assumptions underlying k -anonymity become explicit allowing their more rigorous examination and possible generalization.

Our framework makes the following basic assumptions.

- Subjective sensitivity information is available at the attribute level. In the case that personalized or subjective sensitivity information is not available, default values provided by the database designer may be used instead.
- Loss is defined by measuring separately the sensitivity of the disclosed data under successful and unsuccessful identification.
- Risk is defined as expected or average loss. Different loss aggregations such as worst-case may be incorporated into the framework as necessary.
- Preventing privacy violations is achieved by data suppression and generalization with the particular choice of suppression or generalization possibly revealing information concerning the original data. This 'second order' effect is ignored in the current presentation for simplicity purposes.
- The adversarial external knowledge is given in terms of a side information referred to as a dictionary. We provide a variety of ways to compute or bound the risk in situations where the dictionary is completely known, partially known, or unknown.

In practice, some situations may require slightly altering the assumptions above. However, we believe our framework to be rather general and easily modified to account for many practical cases. While it may seem that our framework makes substantially more assumptions than alternative models, we point out that we simply make these assumptions explicitly rather than implicitly, as is the case for example with k -anonymity. In order to make any meaningful claim concerning privacy preservation, some assumptions have to be made concerning the resources that are available to the attacker and the sensitivity of the data. Making these assumptions explicitly rather than implicitly, allows an open discussion of their strengths and weaknesses and suggests different adaptations depending on the situation at hand.

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] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The SuLQ framework. In *Proc. of PODS*, 2005.

- [3] U.S. Federal Geographic Data Committee. Guidelines for providing appropriate access to geospatial data in response to security concerns, 2005. http://fgdc.er.usgs.gov/fgdc/homeland/access_guidelines.pdf.
- [4] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proc. of PODS*, 2003.
- [5] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure methods for microdata. In P. Doyle, J.I. Lane, J.J.M. Theeuwes, L.V. Zayatz (Eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 111–133, 2001.
- [6] G.T. Duncan, S.A. Keller-McNulty, and L.S. Stokes. Disclosure risk vs. data utility: The r-u confidentiality map. In *Technical Report, National Institute of Statistical Sciences*, 121, 2001.
- [7] C. Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.
- [8] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proc. of PODS*, 2003.
- [9] I.P. Fellegi and A.B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64, 1969.
- [10] M.A. Jaro. UNIMATCH: A record linkage system, user’s manual. In *U.S. Bureau of the Census*, 1978.
- [11] L.V.S. Lakshmanan, R.T. Ng, and G. Ramesh. Toq do or not to do: the dilemma of disclosing anonymized data. In *Proc. of SIGMOD*, 2005.
- [12] E.L. Lawler and D.E. Wood. Branch-and-bound methods: A survey. *Operations Research*, 14(4), 1966.
- [13] G. Lebanon, M. Scannapieco, M.R. Fouad, and E. Bertino. Beyond k-anonymity: A decision theoretic framework for assessing privacy risk. In *Privacy in statistical databases, Springer Lecture Notes in Computer Science*, volume 4302, 2006.
- [14] T. Li and N. Li. T-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. of ICDE*, 2007.
- [15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.
- [16] G. Miklau and D. Suci. A formal analysis of information disclosure in data exchange. In *Proc. of SIGMOD*, 2004.
- [17] P. Mishra and M.H. Eich. Join processing in relational databases. *ACM Computing Surveys*, 24(1), 1992.
- [18] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proc. of PODS*, 1998.
- [19] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Technical Report, SRI International*, 1998.
- [20] R.P. Stanley. Enumerative combinatorics. *Wadsworth & Brooks/Cole Mathematics Series*, 1, 1986.
- [21] L. Sweeney. Privacy-enhanced linking. *ACM SIGKDD Explorations*, 7(2), 2005.
- [22] M. Trottni. A decision-theoretic approach to data disclosure problems. In *Research In Official Statistics*, volume 4, pages 2–17, 2001.
- [23] M. Trottni and S.E. Fienberg. Modelling user uncertainty for disclosure risk and data utility. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, volume 10, pages 511–528, 2002.
- [24] A. Wald. Statistical decision functions. *Wiley*, 1950.
- [25] J. Whittaker. Graphical models in applied multivariate analysis. *John Wiley and sons*, 1990.
- [26] W. Winkler. Methods for evaluating and creating data quality. *Information Systems*, 29(7), 2004.
- [27] X. Xiao and Y. Tao. Personalized privacy preservation. In *Proc. of SIGMOD*, 2006.

-
- [28] W.E. Yancey, W.E. Winkler, and R.H. Creecy. Disclosure risk assessment in perturbative micro-data protection. In *J. Domingo-Ferrer (ed.) Inference Control in Statistical Databases, Lecture Notes in Computer Science, Vol. 2316*, volume 2316, pages 135–152, 2002.
- [29] S. Zhong, Z. Yang, and R.N. Wright. Privacy-enhancing k-anonymization of customer data. In *Proc. of PODS*, 2005.