An Information Theoretic Privacy and Utility Measure for Data Sanitization Mechanisms

Mina Askari, Reihaneh Safavi-Naini, Ken Barker
Department of Computer Science
University of Calgary, Calgary, Canada
{maskari, rei, kbarker}@ucalgary.ca

ABSTRACT

Data collection agencies publish sensitive data for legitimate purposes, such as research, marketing and *etc.*. Data publishing has attracted much interest in research community due to the important concerns over the protection of **individuals privacy**. As a result several sanitization mechanisms with different notions of privacy have been proposed. To be able to measure, set and compare the level of privacy protection, there is a need to translate these different mechanisms to a unified system. In this paper, we propose a novel information theoretic framework for representing a formal model of a mechanism as a noisy channel and evaluating its privacy and utility. We show that deterministic **publishing property** that is used in most of these mechanisms reduces the privacy guarantees and causes information to leak. The great effect of adversary's background knowledge on this metric is concluded. We also show that using this framework we can compute the sanitization mechanism's preserved utility from the point of view of a data user. By using the specifications of a popular sanitization mechanism, k-anonymity, we analytically provide a representation of this mechanism to be used for its evaluation.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms

Security, Measurement

Keywords

Privacy, Utility, Sanitization Mechanism, Information Theory

1. INTRODUCTION

The explosive growth of collecting, storing and publishing of private information about individuals or organizations by government, statistical and business agencies, health networks and social networking systems together with public desire of receiving and analyzing these data, raises a significant concern for the privacy of individuals in the published datasets. The importance of this concern has expedited the emergence of several sanitization mechanisms to be used by data collectors who have collected personal data and want to publish them for new data users. A data sanitization system aims to transform original sensitive data so that the data are useful while the privacy is preserved. Usually a notion of privacy is attached to a sanitization mechanism based on the adversary's ability in identifying an individual in a published dataset (identity disclosure) [29] or disclosure of the users' sensitive information (attribute disclosure)[18][20] with or without adversary's access to some external data. Achieving privacy via sanitization comes at the cost of information loss and/or utility.

Although these metrics are useful within a sanitization mechanism, most of them are individual and data related that makes it hard to evaluate sanitization mechanisms and compare them in general. In addition, the adversarial models in the evaluations are often not appropriately formalized. A common approach in privacy literature to evaluate the effectiveness of any emerging or popular privacy preserving mechanism is to use some data examples and show that privacy issues such as inferring one individual's sensitive attribute, or change in the prior belief of an adversary about the sensitive attribute after observing the published data, still is remained. Theses ad hoc and scenario-based approaches cannot be used to evaluate any arbitrary sanitization mechanism or compare the results of the evaluation. To be able to measure, set and compare the level of privacy protection and utility of sanitization mechanisms, it is necessary to quantify the privacy and utility of such mechanisms within a unified and generic formal framework. In this paper, we propose an information theoretic framework for representing a formal model of a mechanism as a noisy channel and use it for evaluating privacy and utility of such mechanisms.

In our framework, a sanitization mechanism can be analogized to a noisy channel in which some information is sent by a sender over a communication channel to the receiver who tries to reconstruct the original information. The quality of this communication depends upon how accurately the receiver is able to recover the information (even in the presence of errors) from the transferred data. We consider a sanitization mechanism as an information theoretic channel and the inference of private information is regarded as a hypothesis-

CODASPY'12, February 7–9, 2012, San Antonio, Texas, USA.

testing problem, whereas the adversary is analogous to the receiver who wants to reconstruct the original data after their transformation through the sanitization mechanism. Hence, the quality of the privacy provided by a sanitization mechanism depends on how unsuccessful the adversary is in reconstructing the input dataset. What makes a collected dataset valuable for analysis is the knowledge that exists in that dataset. We quantify the application-independent utility of a sanitization mechanism first by defining an information theoretic utility definition for any given dataset using joint statistical properties of that dataset. We then use this utility definition to measures the amount of utility degradation between any input and output dataset. The expected value of the utility degradation is used as a measure of utility for the mechanism. The associated expected degradation provides a measure of utility of the sanitized data, in the intuitive sense that low degradation approximately preserves the values of the original data, and their joint statistical properties. This is the first time that the utility of a mechanism is defined over all possible usage of a mechanism.

We include two different types of background knowledge for our evaluation. The first type is the amount of information an adversary has about original dataset that will affect the privacy measure. The ability of a given adversary in accurately guessing the original dataset depends on adversary's knowledge. We model this knowledge as a probability distribution over datasets. This type of background knowledge has been recognized in literature and we adapted it for our specific model. We also include the amount of information the data user already has about the dataset that will affect the utility of the released dataset for that specific user. To the best of our knowledge this is the first time that this type of background knowledge is considered for a utility metric. After representing the sanitization mechanism by a channel's matrix, based on this representation and background knowledge of the **adversary**, the privacy of the mechanism and based on this representation and the background knowledge of the **data user**, the utility of the mechanism are computed.

By using the proposed framework, we provide a formal representation of k-anonymity as a popular generalization-based sanitization mechanism to be used to evaluate its privacy and utility. In k-anonymity a set of *quasi-identifier* attributes are recognized [29]. Quasi-identifiers are attributes which can be combined and linked to external data to identify individuals. Then generalization, and/or suppression operations are applied on quasi-identifier values. We show that deterministic **publishing property** that is used in most of these mechanisms reduces the privacy guarantees and causes information to leak. We also show the utility that these mechanisms preserve in the published dataset from the point of view of a neutral data user is high for the mechanisms that output less distorted datasets with higher probability.

Contributions

To the best of our knowledge this is the first general analytical framework for measuring both privacy and utility of sanitization mechanisms using information theory, which includes representation of the datasets and sanitization mechanisms and modeling the adversary and user's background knowledge. The proposed framework is general enough to formally express different sanitization mechanisms in the literature and specific enough to include the details of each mechanism to compute the privacy and utility of them.

Our second main contribution is the representation of k-anonymity, a popular generalization method, as a channel matrix. Representation of k-anonymity as the sanitization method has the challenge of providing a realistic setting for this mechanism. We use a global-recoding full-domain approach to cover a wide range of algorithms. We chose global recoding algorithm since its low time complexity and high quality results make it more likely to be used in practice.

Paper organization

The rest of this paper is organized as follows. In next section we will describe how this work relates to existing literature and how it is distinct from previous work (Section 2). The description of our information theoretic model is provided in Section 3 where we define the representation of the datasets, modeling the sanitization mechanisms as noisy channels and representation of the adversary and data user's background knowledge. In Section 4, we then show in detail how to compute the privacy and utility of such models. In Section 5, a formal representation for k-anonymity as a popular sanitization mechanism with realistic settings is shown. Finally, Section 6 provides the conclusion and suggests multiple future directions and improvements to this work.

2. RELATED WORK

Ensuring privacy in data publishing, is a challenging problem, and has been studied extensively in the past [8][29][18]. A notion of privacy is attached to most of these sanitization mechanisms to show the effectiveness of these methods. Lambert [14] provides a formal description of the risk and harm of possible disclosures, and discusses how to evaluate a dataset in terms of these risks and harms. Dalenius [4] poses the problem of re-identification in sanitized census records and introduces the notion of *quasi-identifiers*. Sweeny [29] introduced a popular method in data publishing known as *k-anonymity*. Based on this definition, a table satisfies *k-anonymity* if every record in the table is indistinguishable from at least *k - 1* other records with respect to every set of quasi-identifier attributes. The *k-anonymity* property aims at protecting against *identity disclosure* (that the adversary could uniquely identify a **victim**'s record from the quasi-group), and does not guarantee protection against *attribute disclosure* (the adversary may not precisely identify the record of a victim, but could infer its sensitive values from the published data). This problem has been recognized by several authors [19][27] and the notion of ℓ-*diversity* [20] was developed to address this problem. Observing that when the overall distribution of a sensitive attribute is skewed diversity does not prevent attribute linkage attacks led Li *et al.* [18] to propose t-closeness in which the distribution of a sensitive attribute in any quasi-group should be close to the distribution of the attribute in the overall table and the distance between these distributions should be no more than a threshold t. Dwork [9][10] proposes a new model called ϵ-*differential* privacy that requires that the disclosure of any individual's privacy should not substantially (bounded by ϵ) increase as a result of participating in a statistical database. In this model, privacy disclosure is compared with and without the record owner's data in the dataset, while in previous works the comparison

was between the prior probability and the posterior probability before and after accessing the dataset.

Achieving privacy via sanitization comes at the cost of losing information. For example, for the mentioned t-closeness, enforcing t-closeness would greatly degrade the data utility because it requires the distribution of sensitive values to be the same in all quasi-groups and this would significantly damage the correlation between quasi identifier and sensitive attributes. Information loss and utility metrics try to measure/keep the data quality of the sanitized data during or after the sanitization process. When the purpose of publishing data is unspecified, similarity between the original data and the anonymized data is considered as information loss and can be used during the sanitization process. For example, for k-anonymity, average size of equivalence classes [17] and discernibility [13] are two generic metrics which take equivalence class size into account to measure utility of a sanitized dataset.

To define more reliable utility measures in the context of data applications such as data mining and queries, other metrics such as information-gain-privacy-loss ratio [12] considering a specific mining task have been proposed. Since in most cases, the data publisher does not know how the published data will be analyzed by the recipient, using a utility metric that only targets some specific workloads, will result in a poor dataset for other purposes. Sramka *et al.*[28] developed a general data mining framework that considers the tradeoff between the privacy and utility measure not in the process of anonymization but after the sanitization process. The defined utility and privacy measures depend on dataset and sanitization mechanism and are independent of the data mining task that will be performed on the data. The proposed privacy and utility measures in our information theoretic framework are not tied to any specific sanitization mechanism or dataset. It is not a method to be used during the sanitization process, rather to model and evaluate the sanitization mechanisms based on their specifications.

Information-theoretic measures play a crucial role in sanitization mechanisms [6][7][22][18][11][1]. Reviewing the literature, however, we realize that most of the current information theoretic metrics are used during the sanitization process and are related to random perturbation techniques. In [6][7], for example, the privacy risk is measured as the mutual information between perturbed key attributes and sensitive attributes. The conditional entropy has also been used in this context. Chatzikokolakis *et al.*[2] considers randomized protocols for hiding private information as noisy channels in anonymous protocols. They use this model to define the degree of anonymity for an anonymity protocol. Sankar *et al.*[24] model the databases as a sequence of values. In this abstract universal database model, sanitization is a problem of mapping a set of database entries to a different set subject to specific utility and privacy requirements. Privacy requirements are specified using entropy, while utility requirements are expressed using rate-distortion theory. This method cannot be used to model the existing sanitization mechanisms in order to evaluate them, rather to show regions of privacy-utility tradeoff for databases.

3. PROBLEM MODELING

Given a **dataset** as an input, the objective of a sanitization mechanism is to release a modified (sanitized or anonymized) version of this dataset such that the released version does not allow an adversary to confidently derive the sensitive information of any individual (who is present in that table), and yet, the released dataset can be used to analyze the statistical patterns or other useful information in original dataset. To achieve this goal a sequence of specific operations (such as suppressing identifiers, noise addition or generalization) are done on data prior to publishing them. To be able to provide a general framework to evaluate the effectiveness of these mechanisms, in terms of privacy and utility, we aim to quantify the privacy and utility of such mechanisms within a unified and generic formal framework.

To develop our framework, we model sanitization mechanisms as noisy channels. In a noisy communication channel, there is a set of possible inputs $a \in A$ that are sent through the channel from the source (Figure 1). In these channels, for an input a that is sent to channel, several outputs $o \in O$ may be observed in destination (Figure 2). Suppose that $\mathbb{P}(a)$ is the probability distribution of inputs to a noisy communication channel. The conditional probability $\mathbb{P}(o|a)$ or **Channel's Matrix** is assigned to such noisy channels such that it gives the conditional probability of observing output o when a is the input (Figure 2).

At destination, observing an output o, the receiver should decide what input was sent to the destination. The quality of this communication depends upon how accurately the receiver is able to recover the information (even in the presence of errors) from the transferred data. Assume that function f is a guess function or decision function $f : O \rightarrow A$ that observing an output o, decides that input $f(o)$ was sent at the source. In a noisy channel, the goal is to compute the error probability of possible decision functions and compute an upper bound for such an error probability. The best strategy for the receiver is to apply the MAP (Maximum *A posteriori* Probability) criterion, which says that one should choose the input with the maximum conditional probability given the observation ($f(o) = a_1$ iff $\max_a \mathbb{P}(a|o) = \mathbb{P}(a_1|o)$). Best decision function means that it results in the smallest probability of guessing the wrong hypothesis. The probability of error, in this case, is also called *Bayes risk*. When the distribution of input is known, error probability for a given noisy channel (with a channel's matrix assigned to it) is computable. The challenge in this area is to compute the maximum error probability for the MAP method, over all possible input distributions.

We consider a sanitization mechanism as an information theoretic channel and the inference of private information is regarded as a hypothesis-testing problem, whereas the adversary is analogous to the receiver who wants to reconstruct the original data after their transformation through the sanitization mechanism. We assume that **input is the dataset** (in the form of a table) to be kept hidden, the **output is the sanitized table** and the **matrix represents the conditional probability of having a sanitized table for different inputs**. If we consider a dataset (represented as a table T) as an input to such a model, we may observe one or several possible released tables (output) which are the anonymized versions of the dataset based on the characteristics of the sanitization mechanism. It is also possible that one or several input tables be sanitized to the same output table (source of privacy for a sanitization mechanism).

Having modeled a privacy protocol as an information channel (computing the channel's matrix), the privacy of the protocol is then measured as the expected value of guessing the

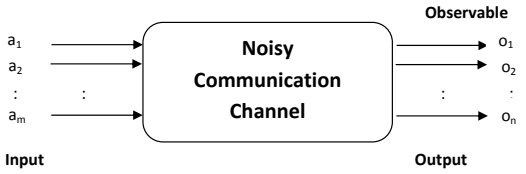

Figure 1: Noisy Channel Model

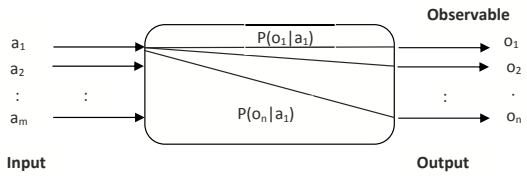

Figure 2: Channel's Matrix for Different Outputs

wrong input by the adversary using the MAP method. In subsection 4.1, we show how to compute probability of error.

3.1 Channel's Matrix

To compute the channel's matrix for a sanitization mechanism, we will first specify the input set T that contains all possible tables T_1 with r records and the set of outputs T' that contains all possible released tables T_1' of size r. We assume that there is a conditional probability (channel's matrix) $\mathbb{P}(T'|T)$, that is derived based on the characteristics of the sanitization mechanism and represents the possibility of releasing table T' by sanitization mechanism if the input table is T. If the sanitization mechanism is known, then the channel's matrix can be derived. It is not unrealistic to assume that adversary is aware of sanitization mechanism as it has been shown before [31][30]. In Section 5, we will show that how we use the characteristics of k-anonymity to compute such channel's matrix and how we use the assumptions of the algorithm to compute the probability distribution $\mathbb{P}(T'|T)$.

3.2 Conditions of Attack

In previous works, the adversary is given a sanitized table T' generated from a **single** data table T, and the quasi identifier for some target individuals that their information are in the released table. The goal of the adversary is to either identify the targeted records in the table or know the sensitive attributes of the targets, or improve its knowledge regarding the sensitive attributes of the targets. The privacy metrics are usually defined over these two definitions and privacy violation is due to implicit dependencies between sensitive and non-sensitive attributes of T and T'. This attack model assumes a weak adversary that uses some auxiliary information (also referred to as background knowledge of the adversary) to do linking or inference attacks in order to get information about some individuals. In our attack model, however, the adversary's goal is to unsanitized the published table, given the characteristics of the sanitized table and the privacy of the sanitization mechanism is defined against such a strong adversary. The effects of an adversary knowledge about the sanitization mechanism on the privacy definition has been studied before [31], but this is the first time that the privacy definition is directly defined considering this attack model. It is important to mention that using our framework it is possible to model the weak adversaries, when we only include one column of the channel's matrix and by changing the probabilities on input datasets to remove the unrelated inputs.

3.2.1 Modeling Adversarial's Knowledge

Measuring the privacy in the presence of additional adversarial knowledge allows the data holders to give higher protection to the individuals in the dataset when they are releasing the data to the public.

Recent studies consider many cases in which the adversary may possess different information about the data; known as adversarial's knowledge. Martin *et al.*[21] modeled background knowledge in the form of conjunctions of k basic implications. Each basic implication is a rule specifying the implication relationship between two predicate about a person and his sensitive values. This work was extended by Chen *et al.*[3] in a scheme called privacy skyline that uses a triple (ℓ, k, m) to quantify the three types of knowledge:(1) ℓ sensitive values that target individual t does not have, (2) the sensitive values of k other individuals, and (3) m people who tend to have the same sensitive values. The type of background knowledge in these models are the knowledge about the sensitive attributes of specific individuals in the population and/or the table and/or partial knowledge about the distribution of some sensitive and non sensitive attributes in the population. The adversary can acquire this knowledge from either public datasets that considered to be safe for access (they only contain identifier attributes and no sensitive information) or from previously released anonymized data from the same organizations or other organizations. It can be demonstrated that giving the adversary more background knowledge will result in more disclosure.

In our framework we model the additional background knowledge of the adversary by a probability distribution over the inputs of the channel, *i.e.*, the probability distribution over the input set T. We postulate that adversary may possess different types of knowledge ranging from access to external datasets to some piece of information about the individuals' attributes in the original dataset and *etc.*, but the results of all can be modeled in a probability distribution over the possible original datasets. In subsection 4.1, to evaluate the privacy of the mechanism or the error probability of the adversary we first assume that the background knowledge of the adversary is zero. In other words, we assume that the probability distribution over all possible datasets is uniform. It means that the adversary has maximum uncertainty about the possible data inputs before observing the output. We will then extend our model to include the prior knowledge of the adversary in the form of a probability distribution on inputs. We will show how the privacy metric will change using this prior knowledge.

The other types of knowledge that usually assist the adversary to jeopardize the individual privacy are the knowledge about the mechanism used for data publication and the optimality goal of the mechanism [31][30]. Our model is in fact taking into account the initial background knowledge of adversary about sanitization mechanism that is modeled as the probability distribution $\mathbb{P}(T'|T)$ or channel's matrix.

3.2.2 Modeling User's Knowledge

In the proposed information theoretic utility for a dataset, we include the amount of information a data user already has about the dataset. The fact that the prior knowledge of adversary about a dataset can affect the level of privacy that a mechanism offers, has long been considered in different analysis. Similarly, we show that the utility that a mechanism offers, not only depends on the application of the released dataset, but also to the knowledge level of potential users of the released dataset. For example, the utility of a released dataset for two different users p in population Ω is different if one of the users has more information about the released dataset, *i.e.*, has the previous release of the dataset in form of joint probability distribution of some attributes. Measuring the utility in the presence of additional user knowledge allows the data holders to give a better utility to the data users when they are releasing the data to the public.

We model the background knowledge of the user as prior probability distribution on different attributes of the dataset and joint probability distribution of any subsets of attributes. Since we will use probability distribution of different attributes as well as the joint probability distribution of any subsets of them to define the utility of a dataset, we model the background knowledge of the data user by her prior knowledge about these distributions. In subsection 4.2, to evaluate the utility of a mechanism we first assume that the background knowledge of user is zero. This means that user's prior knowledge about these distributions are in form of uniform distribution. In subsection 4.2.4, we will extend our model to include the background knowledge of the data users in the utility.

Most of the time, the purpose of data release is not known at the time of publication. Although this model of user's background knowledge does not depend on any specific application for dataset, we can customize it for specific applications of dataset. For example, assume that the goal of releasing dataset is rule association data mining. For this specific application, the joint probability distribution between the targeted attributes in dataset is a good indication of user's background knowledge. If the data is published for modeling the classification of a target attribute in the table, then probability distribution of the attributes that are essential for discriminating the class labels in the target attribute form the background knowledge of the user for that specific purpose.

If the user's prior knowledge about these distributions are in the form of uniform distribution, then the user will get the maximum utility from the dataset. On the other hand, if the background knowledge of the user is close to the knowledge that exists in the dataset, the utility of the dataset for that user is negligible.

4. PRIVACY AND UTILITY

In this section we show how to use the channel's matrix model of the sanitization mechanism to compute its privacy and utility.

4.1 Measuring Privacy

We define the privacy of a sanitization mechanism as the error probability of a guess by an adversary. The attack model in this work is different from the model in which the goal of the adversary is to infer the sensitive information of some targeted individuals. Here, the adversary's goal is to find the whole table T_1 of size r that was released by observing the released table T_1'. For now, the success of adversary is defined as the probability of guessing the exact input. Later, we will extend this definition to decrease the error probability. That is, another table T_1'' is guessed by the adversary that is not the exact Table T_1, but is close to it by some definition.

To compute our privacy metric, we follow MAP (maximum *a posteriori* probability) method. Let T be a random variable that gets its value from the input set to sanitization mechanism containing all n possible tables of size r. The probability distribution $\mathbb{P}(T)$, is *a priori* probability for this random variable. Let T' be another random variable that gets its value from possible outputs of the sanitization mechanism containing m possible released tables of size r.

We assume that there is a conditional probability (channel's matrix) $\mathbb{P}(T'|T)$, that is derived based on the characteristics of the sanitization and represents the possibility of releasing a table $T_1' \in T'$ by sanitization if the input table is $T_1 \in T$. Having $\mathbb{P}(T)$ and $\mathbb{P}(T'|T)$, we derive a posteriori distribution $\mathbb{P}(T|T')$ over all possible values for random variables T and T'. We then apply the MAP to compute the probability of guessing the wrong hypothesis. That is:

$$\mathbb{P}(T'|T) = \begin{bmatrix} p(T_1'|T_1) & \dots & p(T_{m-1}'|T_1) & p(T_m'|T_1) \\ p(T_1'|T_2) & \dots & p(T_{m-1}'|T_2) & p(T_m'|T_2) \\ \vdots & \vdots & \vdots & \vdots \\ p(T_1'|T_{n-1}) & \dots & p(T_{m-1}'|T_{n-1}) & p(T_m'|T_{n-1}) \\ p(T_1'|T_n) & \dots & p(T_{m-1}'|T_n) & p(T_m'|T_n) \end{bmatrix},$$

Suppose that function $f : T' \rightarrow T$ is a guess function or decision function that by observing an output $T_1' \in T'$, gives the input $f(T_1') \in T$ to the adversary. Suppose that for each $T_1 \in T$, $E(T_1)$ is the set of outputs that observing them do not give us T_1 as the input by function f. That is,

$$E(T_1) = T' - f^{-1}(T_1). \tag{1}$$

Probability of error for function f when we have uniform distribution $\mathbb{P}(T) = \frac{1}{n}$ is defined:

$$\begin{aligned} Error_f &= \sum_T \frac{1}{n} \sum_{E(T)} \mathbb{P}(T'|T) \\ &= \sum_T \frac{1}{n} (1 - \sum_{f^{-1}(T)} \mathbb{P}(T'|T)) \\ &= \sum_T \frac{1}{n} - \sum_T \sum_{f^{-1}(T)} \frac{1}{n} \mathbb{P}(T'|T) \\ &= 1 - \sum_T \sum_{f^{-1}(T)} \mathbb{P}(T') \mathbb{P}(T|T') \end{aligned} \tag{2}$$

Based on MAP, we assume that $f(T') = T_1$ iff $\max_T \mathbb{P}(T|T') = P(T_1|T')$. We then obtain that:

$$Error_f = 1 - \frac{1}{n} \sum_{T'} \max_T \mathbb{P}(T'|T) \tag{3}$$

Having computed the channel's matrix for the sanitization mechanism, it is easy to compute 3. The higher this error probability, the less successful is the adversary and the higher the privacy of the protocol. The advantage of this

metric is that it gives a value between 0 and 1 that is comparable for all different sanitization mechanisms, no matter what operations have been performed on data to achieve that privacy.

The maximum value for $\sum_{T'} \max_{T} \mathbb{P}(T'|T)$ is m and it happens when there is a deterministic sanitization mechanism that any output T' is the sanitization of at least one input set T with probability of 1 and when there are $m \leq n$ possible outputs for this mechanism. The minimum value for $\sum_{T'} \max_{T} \mathbb{P}(T'|T)$ is 1 and it happens when the maximum probability of $\mathbb{P}(T'|T)$ is $1/m$ in all columns or all the rows are the same:

$1 - \frac{m}{n} \leq Error_f = 1 - \frac{1}{n} \sum_{T'} \max_{T} \mathbb{P}(T'|T) \leq 1 - 1/n$

When the number of inputs (n) increases or the number of possible outputs decreases, this error probability gets closer to 1. It is a correct conclusion as privacy increases in these cases.

4.1.1 Effect of Adversarial's Knowledge on the Privacy Metric

When we model the channel's matrix we consider the adversary's knowledge about the sanitization mechanism in use. We take into account the adversary's additional background knowledge using a probability distribution over the set of all possible database instances $\mathbb{P}(T)$. To include this probability distribution, we rewrite our privacy metric as follows.

$$\begin{aligned} Error_f &= \sum_{T} \mathbb{P}(T) \sum_{E(T)} \mathbb{P}(T'|T) \\ &= \sum_{T} \mathbb{P}(T)(1 - \sum_{f^{-1}(T)} \mathbb{P}(T'|T)) \\ &= \sum_{T} \mathbb{P}(T) - \sum_{T} \sum_{f^{-1}(T)} \mathbb{P}(T)\mathbb{P}(T'|T) \\ &= 1 - \sum_{T} \sum_{f^{-1}(T)} \mathbb{P}(T')\mathbb{P}(T|T') \end{aligned} \tag{4}$$

Based on MAP, we assume that $f(T') = T_1$ iff $\max_{T} \mathbb{P}(T|T') = P(T_1|T')$. We then obtain that:

$$\begin{aligned} Error_f &= 1 - \sum_{T} \sum_{f^{-1}(T)} \mathbb{P}(T')\mathbb{P}(T|T') \\ &= 1 - \sum_{T'} \mathbb{P}(T') \max_{T} \mathbb{P}(T|T') \\ &= 1 - \sum_{T'} \max_{T} \mathbb{P}(T'|T)\mathbb{P}(T) \end{aligned} \tag{5}$$

The defined $Error_f$ is related to $H(T|T')$ by the Santhi-Vardy bound [25] as follows:

$$Error_f(\mathbb{P}(T)) \leq 1 - 2^{-H(T|T')} \tag{6}$$

where the conditional entropy $H(T|T')$ is defined as

$$H(T|T') = \sum_{T'} \mathbb{P}(T') \sum_{T} \mathbb{P}(T|T') \log \mathbb{P}(T|T') \tag{7}$$

Since the defined privacy depends on probability distribution of the input set (adversarial's knowledge), the minimum error (minimum privacy) that a mechanism offers over all possible distributions of input set happens when the adversary does the maximum knowledge. It is easy to show that for all sanitization mechanisms (no matter what the $\mathbb{P}(T'|T)$ is), when the $H(T)$ is zero the error probability of adversary and hence the privacy of the mechanism will become zero.

In summary, when adversary does not know the probability distribution of the inputs $\mathbb{P}(T)$ but knows the mechanism of release, then the error probability of the adversary will be computed as stated in equation 3. If the adversary has not have access to channel matrix ($\mathbb{P}(T'|T)$) and does not have the probability distribution over the input, then if there are n possible input tables, adversary will randomly pick one of these tables with probability $1/n$. In this case the error probability of the adversary or the success rate of picking the right table is $1 - 1/n$. Now assume that adversary has the probability distribution of the inputs $\mathbb{P}(T)$ but does not know the mechanism of release ($\mathbb{P}(T'|T)$ (uniformly distributed). In this case if adversary chooses the input with the highest probability, the error probability is $1 - \max_{T} \mathbb{P}(T)$. If we add the mechanism of release to the adversary's knowledge, such that it enables him to derive the conditional probability $\mathbb{P}(T'|T)$, then the error probability of the adversary will be computed as stated in equation 5.

4.2 Measuring Utility

The cost of performing the privacy operations on the original collected data in order to achieve privacy is the loss of some information that could have been useful for a third party.

In our work, we argue that data utility is a specific measure since it depends on the specific application and on a dataset. To show this, we will first define an application-independent utility metric for a dataset and then will explain how this utility measure can be customized to be specific to a dataset regarding an application. We also argue that data utility is a relative measure when we consider how much utility is preserved in the released data after sanitization with respect to the original data. To cover this relativity, we propose a distance metric to measure the utility remained in the dataset after sanitization in general and with respect to an application. We also argue that the utility of a sanitization mechanism not only depends on the utility difference between a pair of original and released dataset, but also on the average utility that will be kept for different possible datasets. We also include the user background knowledge in these definitions.

4.2.1 Utility of a dataset

To understand the utility of a dataset, it is important to highlight that the utility of a dataset is the direct result of the correlation between the attributes of the dataset. As the result, data distribution of a dataset is a good indicator of its usefulness. For example suppose that the probability distribution (*i.e.*, frequency) of an attribute's values in a dataset is uniformly distributed. Then, the utility of that dataset is less than another dataset with non-uniform distribution for the same attribute, if the application of the dataset only considers this attribute. Higher entropy of the dataset indicates that the distribution contains more information, and lower entropy indicates that it has less [26]. However, when the application concerns the correlations between several attributes together, the probability distributions of all these

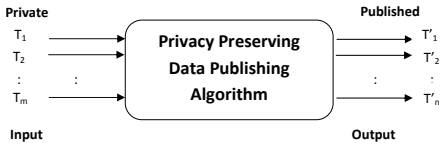

Figure 3: The Sanitization Mechanism as a Noisy Channel

attributes as well as their joint probability is important. The correlation between a subset of attributes may be independent of another subset of attributes and hence does not effect their utilities. For example, if the data is published for modeling the classification of a target attribute in the table, generalizing the values whose distributions are essential for discriminating the class labels in the target attribute will destroy the utility of it but generalizing other attributes may still keep the utility.

To define the utility of our dataset, we start with the utility of a single attribute $U(A_1)$ without considering the other attributes in the dataset and then extend it to the utility of multiple attributes:

$$U_1(A_1) = H_{max}(A_1) - H(A_1) \tag{8}$$

$H_{max}(A_1)$ or the maximum entropy of an attribute A_1 is $\log k$, where k is the number of values for attribute A_1 and it happens when A_1 has a uniform distribution. When there is a correlation between attribute A_1 with another attribute A_2, attribute A_2 can be use to reduce the uncertainty about it and to better predict the value of attribute A_1 . We use conditional entropy $H(A_1|A_2)$ (that can be computed using the joint entropy) to compute the utility of attribute A_1 as follows:

$$U_2(A_1) = H_{max}(A_1) - H(A_1|A_2) \tag{9}$$

Using this metric, we can justify why generalizing an attribute is considered a source of information loss. We can also relate the utility of a sensitive attribute to the metrics that in literature have been defined as privacy metric.

The increase of utility for attribute A_1 by using attribute A_2 or utility gain by using attribute A_2 is equal to

$$\begin{aligned} Utility_{gain} &= U_2(A_1) - U_1(A_1) \\ &= H(A_1) - H(A_1|A_2) = I(A_1; A_2) \end{aligned} \tag{10}$$

If we add another attribute A_3 in this utility measurement, the conditional entropy $H(A_1|A_2, A_3)$ can be used to compute the utility of attribute A_1:

$$U_3(A_1) = H_{max}(A_1) - H(A_1|A_2, A_3) \tag{11}$$

The increase of utility for attribute A_1 by using attributes A_2 and A_3 or utility gain by using attribute A_3 is equal to

$$\begin{aligned} Utility_{gain} &= U_3(A_1) - U_2(A_1) \\ &= H(A_1|A_2) - H(A_1|A_2, A_3) = I(A_1; A_3|A_2) \end{aligned} \tag{12}$$

We can similarly, compute the utility of a dataset, when the goal is to get knowledge about attribute A_1, considering the correlated attributes $A_2, A_3, \ldots, A_n$ as follows

$$U_{max}(A_1) = H_{max}(A_1) - H(A_1|A_2, A_3, \ldots, A_n) \tag{13}$$

Since the defined utility of an attribute is non-decreasing while considering all other attributes in the dataset, we call this the maximum utility of an attribute when all other attributes are used.

To compute the average utility of a dataset considering all possible usage of the dataset, we first compute the maximum utility for all attributes in the dataset, $U_{max}(A_1), U_{max}(A_2), \ldots, U_{max}(A_n)$. We then assume that the data publisher has a priori distribution $\mathbb{P}(A)$ on the possibility of selection of attributes for an application and computes the average utility of a dataset over all possible applications. That is:

$$Utility_T = \Sigma_i \mathbb{P}(A_i) U_{max}(A_i) \tag{14}$$

4.2.2 Comparing the Utilities

In the previous section, we computed the utility of a dataset regardless of the fact that it is an original dataset or a sanitized one. Thus, the utility of a sanitized dataset can also be computed using the defined utility. For a dataset T and its sanitized version T', we define utility degradation as the reduction in quality of T when it is sanitized to T':

$$Utility_{Deg}(T, T') = Utility_T - Utility_{T'} \tag{15}$$

4.2.3 Utility Evaluation of a Sanitization Mechanism

Having the same data utility metric, we still have different utilities for different released tables by the same sanitization mechanism. To evaluate the utility performance of a mechanism regardless of its dataset inputs, we use the channel model as described before to compute the utility of a mechanism on average. This will allow us to compare the utility of different sanitization mechanisms. In the previous section, we quantified the average utility of a dataset and the average utility degradation of a dataset caused by applying a sanitization mechanism. We compute the utility degradation between all input and output datasets $Utility_{Deg}(T, T')$ and then obtain the expected value of the degradation to be used as a measure for utility comparison for the mechanism. The utility degradation is defined as:

$$Utility_G = \sum_T \mathbb{P}(T) \sum_{T'} Utility_{Deg}(T, T') \mathbb{P}(T'|T) \tag{16}$$

$Utility_{Deg}(T, T')$ is the utility degradation of an output table T', when the input table is T. $\mathbb{P}(T)$ is the probability distribution on the input table T and $\mathbb{P}(T'|T)$ is the probability of releasing T', when the input table is T.

The associated expected degradation provides a measure of utility loss of the perturbed datasets, in the intuitive sense that low degradation of a mechanism approximately preserves the values of the original data, and their joint statistical properties with respect to any other mechanism. This is the first time that the utility of a mechanism is defined over all possible usage of a mechanism.

We extend this metric in section 4.2.4 to include the previous knowledge of the user.

4.2.4 *Effect of User Background Knowledge on the Utility Metric*

To consider the user background knowledge in the utility metric, we assume that there is a population Ω of potential users of the released dataset. We model the background knowledge of the user as prior probability distribution on the different attributes of the dataset and joint probability distribution of any subsets of attributes. We then use the background knowledge of a user to define the background utility of that user for a dataset T with attributes $A_1, A_2, \ldots, A_n$ similar to what we did in equation 13. First the background utility of the user for an attribute A_1 is:

$$U^p_{Bg}(A_1) = H_{max}(A_1) - H'(A_1|A_2, A_3, \ldots, A_n) \quad (17)$$

The average background utility for a dataset considering background utility for all attributes in the dataset, $Umax_{A_1}$, $Umax_{A_2}$, ..., $Umax_{A_n}$ is computed as:

$$Utility^p_{Bg}(T) = \Sigma_i \mathbb{P}(A_i) U^p_{Bg}(A_i) \quad (18)$$

Where we assume that the same apriori distribution $\mathbb{P}(A)$ on the possibility of selection of attributes for an application.

When we consider the background utility, then the utility of a dataset T for a user p with background utility is given by:

$$Utility^p(T) = \Sigma_i \mathbb{P}(A_i)(U_{max}(A_i) - U^p_{Bg}(A_i)) \quad (19)$$

So far we have studied the influence of user's background knowledge on the utility of a dataset. Similarly, the effect of users background knowledge on the utility of a mechanism can be captured. Let $\mathbb{P}(p)$ be the probability of giving the released dataset to the user p in the population Ω, then

$$\begin{aligned} Utiulity^{\Omega}_G &= \sum_{p\in\Omega} \mathbb{P}(p) Utility^p_G \\ &= \sum_{p\in\Omega} \mathbb{P}(p) \sum_T \mathbb{P}(T) \sum_{T'} Utility^p_{Deg}(T, T') \mathbb{P}(T'|T) \end{aligned} \quad (20)$$

is the expected utility of a mechanism for population Ω. $Utility^p_{Deg}(T, T')$ is the utility degradation of an output table T', when the input table is T considering the background knowledge of user p for both datasets.

5. EVALUATION OF K-ANONYMITY

In this section we use our framework to represent a popular sanitization mechanism, *i.e.*, k-anonymity [5][23] [29] as channel's matrix. This representation can be used in our defined framework to evaluate privacy and utility of this mechanism. To represent k-anonymity as a channel's matrix, we will first give the details of this mechanism that will lead us to define a representation for a dataset, a method to specify the input set and output set of the mechanism and a way to compute the conditional probabilities. After computing these parameters, we are able to compute the privacy of the mechanism as described before.

5.1 Preliminaries and Assumptions

To enforce a specified privacy requirement on a dataset before being published based on a sanitization mechanism, a sequence of operations are applied on the dataset. Based on these operations the sanitization mechanism can be grouped in three general groups: (1) **generalization-based** methods that use generalization and suppression, (2) **anatomization** methods that use anatomization and permutation operations, and (3) **randomization** methods that use data perturbation operations such as adding noise, data swapping, and synthetic data generation. k-anonymity is an instance of a generalization method. To be able to represent a mechanism we need to know the details of it. Here we briefly review the basic definitions of a generalization-based sanitization mechanism in general and k-anonymity as a specific instance of these mechanisms.

A dataset to be released contains some **sensitive** attributes, **identifying** attributes, and ***quasi-identifying*** attributes. The values of quasi-identifying attributes can be used to uniquely identify at least a single individual in the dataset via linking attacks. In generalization-based mechanisms, the dataset is assumed to be in the form of a data table T and there is a domain associated with attributes of this table. Generalization and suppression are two operations that are used in these mechanisms that replace values of some attributes (usually are applied only to quasi identifiers, with sensitive attributes left intact), with less specific values. For example, ordered attributes such as "age" are partitioned into intervals, and categorical attributes are partitioned according to some domain or value generalization hierarchies (for example, cities are generalized to counties, counties to states, and states to regions). After theses operations, some subsets of tuples in dataset share the same values for quasi-identifiers. Every subset of tuples in dataset that share the same values for quasi-identifiers (and are indistinguishable from each other) is often referred to as an ***equivalence class***. A released dataset is said to satisfy k-anonymity, if for each existing combination of quasi-identifier attribute values in the dataset, there are at least $k-1$ other records in the database that contain such a combination.

The advantage of the generalization approach is that the released data is semantically consistent with the original data which means that it preserves the truthfulness of data. k-anonymity and most of its derivations follow this method for providing privacy. Because of its conceptual simplicity, k-anonymity has been widely discussed as a viable definition of privacy in data publishing. Also, due to algorithmic advances in creating k-anonymous versions of a dataset; k-anonymity has grown in popularity. For these reasons we also use it for evaluation using our framework.

There are different schemes in literature for generalizing one or several attributes such as **global recoding** in which if a value of an attribute is generalized, all its instances in the dataset are generalized to the same value or **local recoding** in which some instances of a value may remain ungeneralized while other instances may be generalized to different levels. For global recoding we have **full-domain** generalization[15][29] that generalizes all values in an attribute to

the values of the same domain in domain generalization hierarchy and **subtree** generalization [16][13] generalizes all child values to the same value in a value generalization hierarchy. There are several methods to achieve k-anonymity. Among all possible ways of generalization, we will use the algorithm that uses global recoding, full-domain generalization in which a quasi-identifier is mapped to a generalized value at the same level of the hierarchy structure. We assume that there is no cell suppression and consider all possible generalizations by predefined generalization hierarchies.

5.2 Dataset Representation

We assume that the data table T follows the relational model. The relational model represents the dataset as a collection of relations, each one representing a collection of related data values. We consider a collected dataset as a single relational table T, containing non aggregate personal data. Based on this, we model the dataset of the data owner as a collection of n tuples $t_1, t_2, ..., t_n$ and m attributes $A = \{A_1, ..., A_m\}$. In a tuple $t = (a_1, \ldots, a_m) : D_1 \times \ldots \times D_m$, a_i represents the attribute value for A_i and D_i's denote the domain for attribute A_i.

If C is a subset of attributes: $C = \{C_1, C_2, \cdots, C_p\} \subseteq A$, for a tuple t, we use the notation $t[C]$ to denote the tuple $(t[C_1], ..., t[C_p])$, which is the projection of t onto the attributes in C.

5.3 Enumerating the Input Set (T)

To specify the input set T, we first assume that the number of tuples in each data table $T_1 \in T$ is r. For this, we assume that r also represents the number of tuples in the released table T' and can be specified by observing the output table T'. This is a valid assumption if we assume that the anonymization algorithm performs suppression by generalizing all the QI attributes to the last level.

To compute m, the size of input set T, assume that we have a table $T_1 \in T$ with r tuples and each tuple $t = (a_1, \ldots, a_p) \in T$ contains p attributes and a_i in tuple t presents the value for attribute A_i. Suppose that $|A_i|$ shows the size of the domain that the values of A_i are taken from.

Each row in this table can get $\prod_{i=1}^{p} |A_i|$ possible values. For r rows, there are

$$m = \binom{\prod_{i=1}^{p} |A_i| + r - 1}{r} = \frac{(\prod_{i=1}^{p} |A_i| + r - 1)!}{r!(\prod_{i=1}^{p} |A_i| - 1)!} \qquad (21)$$

of ways of assigning values to the attributes. This number represents the size of input set T. It is easy to see that following this procedure, we can build the input set.

Example- Suppose that we have the released table which has 5 rows. From this we realize that $r = 5$. There are three attributes in the released table: Age, gender and disease. The disease attribute is a sensitive attribute and has not been generalized. It gets its value from domain $D = \{Cold, Heartattack, Diabetes\}$. Age and gender are QI attributes and have been generalized. The value and domain generalization hierarchy for these attributes are depicted in Fig 4 and 5:

For this example, where $p = 3$, so we will have $\prod_{i=1}^{3} |A_i| = 4 \times 2 \times 3 = 24$ possible values for each row. For 5 rows,

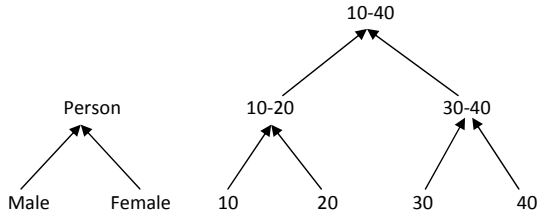

Figure 4: Value Generalization Hierarchy for Age and Gender Attributes

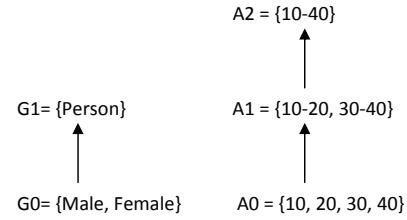

Figure 5: Domain Generalization Hierarchy for Age and Gender Attributes

there are $m = \binom{24+5-1}{5} = \frac{28!}{5!23!} = 98280$ of ways of assigning values to the attributes. This number represent $|T|$.

5.4 Enumerating the Output Set (T')

In generalization-based methods, the values in table T_1 are substituted with their generalized values according to a generalization hierarchy. The number of distinct values associated with each attribute is non-increasing in each level, so the substitution tends to map several values to the same result, thereby decreasing the number of distinct tuples in T. The number of possible outputs, *i.e.*, the size of T', depends on the generalization hierarchy that the anonymization algorithm is using and the generalization method. For example we consider full-domain generalization in our study. We assume that we have a table $T_1' \in T'$ with r tuples and each tuple $t = (a_1, \ldots, a_p) \in T'$ contains p attributes and a_i in tuple t presents the value for attribute A_i. Suppose that $|A_i|$ shows the size of the domain that the values of A_i are taken from before generalization and $|DH_i|$ shows the height of generalization hierarchy for each attribute A_i.

The number of different generalizations of a table T, when generalization is enforced at the attribute level (full-domain generalization and single-domain), is equal to the number of different combinations of domains that the attributes in the table can assume. Given domain generalization hierarchies DH_i for attributes $A_i, i = 1, \ldots, p$; the number of generalizations, enforced at the attribute level, for table $T(A_1, \ldots, A_n)$ is computed as

$$N = \prod_{i=1}^{p} (|DH_i| + 1). \qquad (22)$$

Given any of these generalizations such as $G_{l_1,l_2,\ldots,l_p}$ $(l_1 = 1 \ldots |DH_1| + 1, \ldots l_p = 1 \ldots |DH_p| + 1)$, we will now com-

pute the possible ways of having a table with r rows and p attributes such that each attribute A_i gets its value from the level of l_i in the domain generalization hierarchy DH_i (suppose that $|l_i|$ shows the size of domain that the values of A_i are taken from in the level of l_i in the generalization).

Each row in this table can get $\prod_{i=1}^{p}|l_i|$ possible values. For r rows, there are

$$n'_{l_1,l_2,\ldots,l_p} = \binom{\prod_{i=1}^{p}|l_i|+r-1}{r} = \frac{(\prod_{i=1}^{p}|l_i|+r-1)!}{r!(\prod_{i=1}^{p}|l_i|-1)!} \quad (23)$$

of ways of assigning values to the attributes. Since we had $\prod_{i=1}^{p}(|DH_i|+1)$ ways of generalization,

$$n = \sum_{1,\ldots,1}^{|DH_1|+1,\ldots,|DH_p|+1} n'_{l_1,l_2,\ldots,l_p} \quad (24)$$

will be the size of T'. It is easy to see that following this procedure, we can build the output set.

It is important to notice that not all of these computed output tables are k-anonymized. In fact we can reduce to the output tables that are k-anonymized:

First for any given generalization such as $G_{l_1,l_2,\ldots,l_p}$ ($l_1 = 1\ldots|DH_1|+1, \ldots l_p = 1\ldots|DH_p|+1$), we will compute all possible ways of having a table with r rows that are k-anonymized. Again we assume that each attribute A_i gets its value from the level of l_i in the domain generalization hierarchy DH_i and $|l_i|$ shows the size of domain that the values of A_i are taken from in the level of l_i in the generalization.

If we have r records, if the output satisfies k-anonymity, for any $k \leq r$, there would be $\lfloor \frac{r}{k} \rfloor \ldots 1$ quasi groups in the table. When we have $\lfloor \frac{r}{k} \rfloor$ quasi groups, then there will be $\binom{\prod_{i=1}^{p}|l_i|}{\lfloor \frac{r}{k} \rfloor}$ possible ways of having k-anonymized tables. Now assume that we have $\lfloor \frac{r}{k} \rfloor - 1$ quasi groups. In this case we will have $\lfloor \frac{r}{k} \rfloor - 1$ quasi groups with k records and we need to distribute additional k records in theses quasi groups, so we will have in total $\binom{\prod_{i=1}^{p}|l_i|}{\lfloor \frac{r}{k} \rfloor - 1}\binom{(\lfloor \frac{r}{k} \rfloor - 1)+k-1}{k}$ possible ways of having k-anonymized tables. Continuing this approach we will have $\binom{\prod_{i=1}^{p}|l_i|}{\lfloor \frac{r}{k} \rfloor - i}\binom{(\lfloor \frac{r}{k} \rfloor - i)+ik-1}{ik}$ k-anonymized tables with $\lfloor \frac{r}{k} \rfloor - i$ quasi groups.

The total k-anonymized tables that get their values from the generalization level $G_{l_1,l_2,\ldots,l_p}$ ($l_1 = 1\ldots|DH_1|+1, \ldots l_p = 1\ldots|DH_p|+1$) is then computed as:

$$n'_{l_1,l_2,\ldots,l_p} = \sum_{i=0}^{(\lfloor \frac{r}{k} \rfloor - 1)} \binom{\prod_{i=1}^{p}|l_i|}{\lfloor \frac{r}{k} \rfloor - i}\binom{(\lfloor \frac{r}{k} \rfloor - i)+ik-1}{ik} \quad (25)$$

Since we had $\prod_{i=1}^{p}(|DH_i|+1)$ ways of generalization, the size of T' will be $n = \sum_{1,\ldots,1}^{|DH_1|+1,\ldots,|DH_p|+1} n'_{l_1,l_2,\ldots,l_p}$. It is easy to see that following this procedure, we can build the output set.

5.5 Computing the channel's matrix

After specifying the sets of input and output (T and T'), to compute the channel's matrix we first need to find the mapping between two sets and then compute the probability of this mapping.

To find the mapping between the input and output sets, we compute all possible inputs T that can be generalized to an output table T', for each generalized output T'. After determining the link between the input and output sets, for each (T, T') we will compute the probability of $\mathbb{P}(T'|T)$ in the following way:

We assume that for each input table T there are more than one possible outputs T', such that $\mathbb{P}(T'|T) > 0$. Suppose that for one row of the matrix, there are x possible outputs $T' \in DB'$, then $\mathbb{P}(T'|T) = \frac{1}{x}$. To see that the output table T' is the possible release of how many other input table T, any generalized value g_i of an attribute can be the generalization of ni values (all the leaf children of g_i in generalization hierarchy). For all the generalized values g_i in the table T', there are $\sum_{i=1}^{rp} n_i$ possible tables that will be the possible input.

6. CONCLUSIONS AND FUTURE WORK

We proposed the first general analytical framework for measuring both privacy and utility of sanitization mechanisms using information theory, which includes representation of the datasets and sanitization mechanisms and modeling the adversary and user's background knowledge. The proposed framework is general enough to formally express different sanitization mechanisms in the literature and specific enough to include the details of each mechanism to compute the privacy and utility of them. To use the proposed framework, we provided a formal representation of k-anonymity as a popular generalization-based sanitization mechanism as a channel's matrix.

We included two different types of background knowledge for our evaluation. The first type is the amount of information an adversary has about original dataset and about the privacy mechanism that will affect the privacy measure. We also included the amount of information the data user already has about the dataset that will affect the utility of the released dataset for that specific user.

The scope of our work is on non-interactive data publishing mechanisms,*i.e.*, the dataset is sanitized before publication and the whole sanitized dataset is published. For interactive data privacy mechanisms, either the dataset is sanitized and the queries are run on it, or the queries are run on the original database but query restriction-based methods are used to provide privacy. Our goal is to extend our proposed framework to be used for interactive sanitization mechanisms as well. In the future, we also plan to design an information theoretic-based sanitization mechanism that takes as the input a data user's current background knowledge about the data and its expected utility, the individuals expected privacy and the maximum level of adversary's current background knowledge about the data and generates a channel's matrix to be used to publish an input dataset.

7. REFERENCES

[1] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *PODS*, 2001.

[2] K. Chatzikokolakis, C. Palamidessi, and P. Panangaden. On the bayes risk in information-hiding protocols. *J. Comput. Secur.*, 16(5):531–571, 2008.

[3] B.-C. Chen, K. LeFevre, and R. Ramakrishnan. Privacy skyline: privacy with multidimensional adversarial knowledge. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 770–781. VLDB Endowment, 2007.

[4] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444, 1977.

[5] S. De Capitani di Vimercati and P. Samarati. k-Anonymity for protecting privacy. October 2006.

[6] J. Domingo-Ferrer, A. Oganian, and V. Torra. Information-theoretic disclosure risk measures in statistical disclosure control of tabular data. In *SSDBM*, pages 227–231, 2002.

[7] J. Domingo-Ferrer, F. Sebé, and J. Castellà-Roca. On the security of noise addition for privacy in statistical databases. In *Privacy in Statistical Databases*, pages 149–161, 2004.

[8] G. Duncan and D. Lambert. The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2):207–17, April 1989.

[9] C. Dwork. Differential privacy. In *ICALP (2)*, pages 1–12, 2006.

[10] C. Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19, 2008.

[11] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222, 2003.

[12] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, pages 205–216, 2005.

[13] R. J. B. Jr. and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, pages 217–228, 2005.

[14] D. Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9:313–331, 1993.

[15] K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, and D. J. DeWitt. Limiting disclosure in hippocratic databases. In *VLDB*, pages 108–119, 2004.

[16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD Conference*, pages 49–60, 2005.

[17] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, page 25, 2006.

[18] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115, April 2007.

[19] A. Machanavajjhala and J. Gehrke. On the efficiency of checking perfect privacy. In *PODS '06: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 163–172, New York, NY, USA, 2006. ACM.

[20] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l -diversity: Privacy beyond k -anonymity. *TKDD*, 1(1), 2007.

[21] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE*, pages 126–135, 2007.

[22] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer. From t-closeness to pram and noise addition via information theory. In *Privacy in Statistical Databases*, pages 100–112, 2008.

[23] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *PODS*, page 188, 1998.

[24] L. Sankar, S. R. Rajagopalan, and H. V. Poor. A theory of privacy and utility in databases. *CoRR*, abs/1102.3751, 2011.

[25] N. Santhi and A. Vardy. On an improvement over rényi's equivocation bound. *CoRR*, abs/cs/0608087, 2006.

[26] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27, 1948.

[27] A. Solanas, F. Sebé, and J. Domingo-Ferrer. Micro-aggregation-based heuristics for p-sensitive k-anonymity: one step beyond. In *PAIS*, pages 61–69, 2008.

[28] M. Sramka, R. Safavi-Naini, J. Denzinger, and M. Askari. A practice-oriented framework for measuring privacy and utility in data sanitization systems. In *EDBT/ICDT Workshops*, 2010.

[29] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10:2002, 2002.

[30] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, pages 543–554, 2007.

[31] L. Zhang, S. Jajodia, and A. Brodsky. Information disclosure under realistic assumptions: privacy versus optimality. In *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*, pages 573–583, New York, NY, USA, 2007. ACM.

APPENDIX

A. EXAMPLE

Assume that we have computed the following channel's matrix for a given sanitization mechanism:

$p(T'\|T)$	T'_1	T'_2	T'_3	T'_4
T_1	5/17	2/17	0	10/17
T_2	0	20/ 35	5/35	10/35
T_3	15/28	5/28	3/28	5/28
T_4	5/20	8/20	0	7/20

Assume that the adversary's background knowledge is zero, *i.e.*, $P(T_1) = P(T_2) = P(T_3) = P(T_4) = 1/4$, then the privacy of this sanitization mechanism is computed as follows:

$$\begin{aligned} Error_f &= 1 - \sum_{T'} \underset{T}{max} \mathbb{P}(T'|T)\mathbb{P}(T) \\ &= 1 - 1/4(15/28 + 20/35 + 5/35 + 10/17) \\ &= 1 - 0.46 = 0.54 \end{aligned} \tag{26}$$

However, if adversary has prior background knowledge of $P(T_1) = 1$ and $P(T_2) = P(T_3) = P(T_4) = 0$ about the input datasets the probability of error will become zero for the same sanitization mechanism. This example show the importance of adversary's background knowledge in privacy of a mechanism.

If the adversary did not know about the mechanism of release, he would randomly pick one of input tables with probability of 0.25 and error probability of 0.75. In fact this is similar to the case when the knowledge of adversary about the above mechanism can be modeled as:

$p(T'\|T)$	T'_1	T'_2	T'_3	T'_4
T_1	1/4	1/4	1/4	1/4
T_2	1/4	1/4	1/4	1/4
T_3	1/4	1/4	1/4	1/4
T_4	1/4	1/4	1/4	1/4

And the error probability is computed as:

$$\begin{aligned} Error_f &= 1 - \sum_{T'} \underset{T}{max} \mathbb{P}(T'|T)\mathbb{P}(T) \\ &= 1 - 1/4(1/4 + 1/4 + 1/4 + 1/4) \\ &= 1 - 1/4 = 0.75 \end{aligned} \tag{27}$$

This means that for a non deterministic algorithm with the uniformly distributed channel's matrix the maximum privacy is achieved if the background knowledge of adversary is zero. For the same sanitization mechanism, if adversary has prior background knowledge of $P(T_1) = 1$ and $P(T_2) = P(T_3) = P(T_4) = 0$ about the input datasets, again the probability of error will become zero.