1-1-2006

Comparisons of k-anonymization and randomization schemes under linking attacks

Zhouxuan Teng
Syracuse University, Department of Electrical Engineering and Computer Science, zhteng@syr.edu

Wenliang Du
Syracuse University, Department of Electrical Engineering and Computer Science, wedu@ecs.syr.edu

Comparisons of K-Anonymization and Randomization Schemes Under Linking Attacks*

Zhouxuan Teng and Wenliang Du
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, USA
Email: zhteng@syr.edu,wedu@ecs.syr.edu

Abstract

Recently K-anonymity has gained popularity as a privacy quantification against linking attacks, in which attackers try to identify a record with values of some identifying attributes. If attacks succeed, the identity of the record will be revealed and potential confidential information contained in other attributes of the record will be disclosed. K-anonymity counters this attack by requiring that each record must be indistinguishable from at least $K-1$ other records with respect to the identifying attributes.

Randomization can also be used for protection against linking attacks. In this paper, we compare the performance of K-anonymization and randomization schemes under linking attacks. We present a new privacy definition that can be applied to both k-anonymization and randomization. We compare these two schemes in terms of both utility and risks of privacy disclosure, and we promote to use R-U confidentiality map for such comparisons. We also compare various randomization schemes.

1 Introduction

With the advance of information technologies, the amount of information collected by different entities is increasing exponentially. These data, if available to the general public, can significantly benefit the society. However, many databases contain people's confidential information, which, as required by law, cannot be disclosed. Some necessary transformations have to be performed on a database before its publication, but such transformation is non-trivial. One of the greatest threats faced by database publication is linking attacks. In this type of attack, adversaries who have already known partial information (through other means) about a person try to identify the record that belong to this person; if adversaries can successfully identify the record, all the information (including confidential information) of that record will be disclosed.

K-anonymity is a popular way of specifying privacy requirements over a to-be-published database which might be subject to external linking attacks. This privacy requirement is conceptually simple - it only imposes one restriction: a database is *k-anonymized* if, for each existing combination of identifying attributes, there are at least K records that contain such a combination. Attackers are thus forced to accept this coarse granularity of K records.

The most straightforward way to achieve K-anonymity is to group some records with similar identifying attributes' values together and merge these values to one "new" value which is a set containing all original values merged. This technique is called the generalization approach due to the fact that some attributes' values of certain records are generalized to a coarser, or a more general value.

Another way to achieve data privacy in a published database is to introduce extra noise to a database by either adding random noises or mixing different values together. An advantage of this randomization approach is that this process is partially reversible, i.e., these added noises can be partially removed so that some aggregate information can be recovered from the disguised database while the precise reconstruction of individual record is still impossible.

The possibility of reconstruction of original data distribution from disguised database makes randomization more attractive than pure generalization where multiple values are mapped to a same fake label and this label can only provide a uniform guess of original value. Thus we propose to apply randomization technique to prevent linking attacks, which might yield a disguised database with a higher data utility than generalization.

Contributions We propose a new privacy definition that can be applied to many data disguising techniques, including K-anonymization and randomization. We also promote to use the R-U confidentiality map to compare K-

*This work was supported by Grant ISS-0219560, ISS-0312366 and CNS-0430252 from the United States National Science Foundation.

anonymization and randomization. Extensive simulations are also carried out to reach a conclusion.

2 Background

2.1 K-Anonymity

K-anonymity was formulated by Samarati and Sweeney [7] to protect a public database from linking attacks. It is a popular way of specifying privacy requirements over a to-be-published database which might be subject to external linking attacks. This privacy requirement is conceptually simple - it only imposes one restriction: a database is *k-anonymized* if for each existing combination of identifying attributes, there are at least K records that contain such a combination. Attackers are thus forced to accept this coarse granularity of K records.

All identifying attributes values that are generalized into the same macro-value form a set, and this set is characterized by the following concept of equivalence class.

Definition 2.1 (Equivalence Class) *An equivalence class is a set E_c of combinations of identifying attributes values that satisfy the following two conditions:*

1. *$\forall x \in E_c, \forall y \in E_c$, the probability of transforming x to s is not zero.*
2. *$\forall y \notin E_c$, the set $\{y\} \cup E_c$ does not satisfy the above condition.*

Some algorithms have been proposed to K-anonymity problem [5]. In our paper, we will apply the Incognito algorithm by LeFevre et al [5]. In Incognito algorithm, a complete search is done to find the least generalization schemes given a generalization hierarchy.

2.2 Randomization

The Randomized Response (RR) technique was proposed by Warner [8] in the statistics community in 1965 as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute *A*. Warner developed a randomized response technique to prevent the interviewer from learning the actual answers from the clients, while allowing the interviewer to compute the aggregate result, i.e., the percentage of clients that has attribute *A*. Warner's methods focus on binary case and can be extended to categorical attributes.

The randomization approach for privacy preserving data mining was first reintroduced by Agrawal and Srikant proposed in [1]. To randomize categorical data, Rizvi and Haritsa presented a scheme to mine associations with secrecy constraints in [6]; Evfimievski et al. proposed an approach to conduct privacy preserving association rule mining [3]; Du and Zhan proposed an approach to conduct privacy preserving decision tree building [2]. Zhu and Liu proposed a general framework for randomization based on mixture models [9].

3 Quantifying Privacy for Linking Attacks

Given a published database, the goal of *linking attacks* is to find out the value t, the index of a record, given values of identifier attributes, so that they can link the identity ID_t with the t-th record. Generally speaking, the effectiveness of linking attacks depends on the probability of discovering the correct index t of the target record. The damage of linking attacks is alleviated to some degree when this probability is bounded by some number less than 1.

We use an oracle model in the quantification of privacy, i.e., we assume that the original database is totally known in calculation, but not to attackers. This corresponds to the case when the database owner evaluates the privacy breach of the published database.

Some notations need to be clarified before we proceed. Let N be the number of records in the database. Let R_t be the values of identifiers of the t-th record. Assume that attackers have known the values R_t of the t-th record. Also, we only focus on identifying attributes.

Our quantification will be presented in a series of definitions and theorems, and many proofs will be omitted due to page limitations. Please see our website for the complete version with proofs and intuitive explanations.

Definition 3.1 *The success of linking attacks for the t-th record with identifiers values R_t is defined to be the following probability*

$$\Pr(t \mid R_t) = \Pr(\textit{Find } t \textit{ successfully given } R_t).$$

Definition 3.2 *Let $\Pr(A \rightarrow B)$ denote the probability of reconstructing values of identifier attributes as B from the disguised data, given that the original values are A.*

$\Pr(A \rightarrow B)$ implies how much information is preserved after the disguising transformation of the original database and it is similar to the probability defined in binary attribute case [6].

In the reconstructing process, only the records whose identifiers' values are in the equivalence class which A belongs to should be considered. Assume the size of this equivalence class is m, and denote its elements as $C_1, C_2, \cdots, C_m$. Also, denote the original and disguised values of identifiers as X and Y respectively.

Lemma 3.1 *The probability $\Pr(A \rightarrow B)$ can be computed*

as follows:

$$\Pr(A \to B) = \sum_{j=1}^{m} P(Y = C_j \mid X = A) \qquad P(X = B \mid Y = C_j).$$

Lemma 3.2 *The probability that a disguised identifiers value is C is*

$$P(Y = C) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} P(Y = C \mid X = R_i).$$

where R_i is the identifiers value of the i-th record.

Definition 3.3 *The probability of the success of linking attacks can be computed as follows:*

$$\Pr(t \mid R_t) \quad = \quad \frac{\Pr(R_t \to R_t)}{NP(X = R_t)}.$$

Linking attacks involve two steps. First, attackers need to make sure the values of identifier attributes of the re-identified record are correct; second, if there are multiple records in the original database with these values, attackers need again to choose one from this many records. The denominator and the nominator of the above equation reflect these two steps.

Theorem 3.1 *The probability of successful identification of the record t is*

$$\Pr(R_t \to R_t) = \sum_{j=1}^{m} \frac{P^2(Y = C_j \mid X = R_t) P(X = R_t)}{\sum_{i=1}^{N} P(Y = C_j \mid X = R_i) \cdot \frac{1}{N}},$$

where $\{C_1, C_2, \cdots, C_m\}$ is the set of equivalence class that identifier values R_t belongs to.

Theorem 3.2

$$\Pr(t \mid R_t) = \sum_{j=1}^{m} \frac{P^2(Y = C_j \mid X = R_t)}{\sum_{i=1}^{N} P(Y = C_j \mid X = R_i)}.$$

Proof. By definition 3.3 and theorem 3.1. ■

Armed with this probability of identifying original records, we are ready to define privacy against linking attacks. In particular, we are interested in the following two definitions.

Definition 3.4 *The Bound Privacy against linking attacks is defined to be*

$$\textit{Bound Privacy} = 1 - \max_{t=1}^{N} \Pr(t \mid R_t).$$

Definition 3.5 *The Top q-percentile Privacy against linking attacks is defined to be*

$$\textit{Top q-percentile Privacy} = 1 - \frac{1}{N'} \sum_{i=1}^{N'} \Pr(i \mid T_i).$$

where in the second term we compute the average of the q-percent largest values of $Pr(i|T_i)$, and $N' = Nq/100$. $\{T_1, T_2, \cdots, T_N\}$ is the set of sorted records in decreasing order of $Pr(i|T_i)$.

Definition 3.6 (Risk) *The risk against linking attacks is defined to be*

$$\textit{Risk} = 1 - Privacy.$$

Accordingly, we have Bound Risks and top q-percentile risks. These two risk definitions characterize different aspects of linking attacks. The Bound Risk is concerned with the worst case where the identifying probability of a certain record is the largest among all records. On the other hand, the top q-percentile privacy is concerned with the average case of the q-percent largest values.

The above defined privacy and risk for linking attacks is very general, and thus can be applied to any data disguising technique which transforms the original database to the disguised database.

We can easily show that the bound risk in K-Anonymized dataset is $1/K$, which means that the privacy definition of the K-Anonymity fits into our framework of privacy analysis, and it is a special case of our Bound Privacy. Furthermore, we have one more privacy - top q-percentile privacy, which is not considered in K-Anonymity, to capture the average case of the privacy breach.

Randomization schemes also fit into this privacy analysis naturally due to the data disguising method used. In a randomization scheme, each categorical value is retained with certain probability and is transformed to one of the other categorical values in the same equivalence class with some probabilities. These probabilities are utilized in the computation of the Bound Privacy and the Top q-percentile Privacy.

4 R-U Confidentiality Map

There are good and bad sides of the published database, for example, utility for data mining, and disclosure risk due to possible linking attacks. The ultimate goal of data disguising is to maximize utility and minimize risk of privacy disclosure at the same time.

Obviously, this ultimate goal is not feasible in reality, thus the trade-off between these two conflicting ends need to be made. Likewise, these two aspects need to be considered in any comparison between any two data disguising schemes as countermeasures against linking attacks.

While it is easy to compare either utility or risk of two different data disguising schemes, any such comparisons are incomplete if done separately. Instead, we need to consider these two aspects at the same time. Specifically, we should only compare the privacy of different schemes when both schemes have exactly the same utility, and we should only compare the utilities of different schemes when both schemes have exactly the same privacy.

However, it is very difficult for two different schemes to yield exactly the same utility or privacy in comparisons, because precise control of either utility or privacy in a data disguising scheme is almost impossible. This requires the generation of a curve describing relationships between risk and utility for each data disguising scheme, and we should compare such curves of different schemes.

A theoretical relationship between risk and utility is highly desired but very difficult to obtain due to the complexity of whole data disguising and utilization process. Instead, we can quantify this relationships in an empirical curve by simulations. This curve is named empirical R-U confidentiality map (abbreviated as R-U map), proposed by Duncan et al [4].

Each such empirical curve depicts the characteristics of the data disguising technique in terms of relations between the privacy disclosure risk under linking attacks and the utility in various data mining processes. It not only makes the trade-off between these two conflicting aspects prominent, but also enables us to compare performance of two data disguising schemes in an easy and complete way.

Therefore, we will mainly use this R-U Confidentiality map as a metric in comparing different data disguising schemes, and we will draw conclusions from these curves.

Utility

In most, if not all, data mining processes, the most important information is the probability distribution of the data, which motivates us to focus on the data distribution when evaluating the utility of a database. Therefore, the Kullback-Leibler Distance (KL-Distance), or relative entropy, which is commonly used as a measure of difference between two probability distributions in information theory, is used as the metric of the utility of the disguised database in this paper.

Specifically, for a disguised database, we compute the KL-distance between it and the original. The larger the KL-distance value, the bigger the difference between two databases, and the worse the utility of the disguised database.

5 Methodology

Our strategy is to compare different data disguising techniques using R-U Maps. Specifically, we fill first compare randomization and generalization schemes, and then compare different randomization schemes.

5.1 Comparisons of Randomization and K-Anonymity Schemes

We may compare randomization and K-Anonymity schemes with any equivalence class definitions. To make comparisons more fair to K-Anonymity, we will run a K-Anonymity algorithm, and the resulting solutions, which are the best cases for K-Anonymity requirement, are used to generate R-U maps. We will use a complete search algorithm that does not use any approximations, i.e., Incognito algorithm proposed by LeFevre et al [5].

Our intuition is that, unlike in the database disguised via generalization methods, in the disguised database via randomization, we can partially recover original data distribution, which improves the utility of the disguised database. Therefore, we believe that R-U maps will show that randomization methods will generally outperform generalization method.

5.2 Comparisons of Different Randomization Schemes

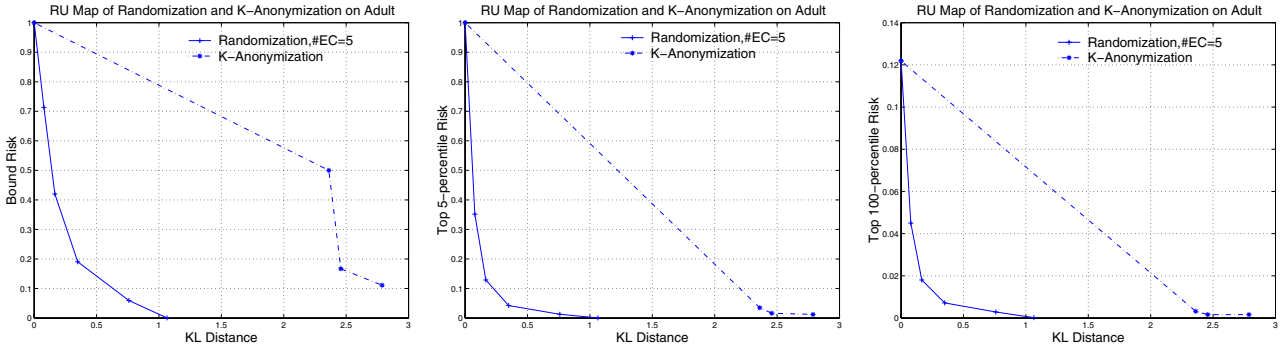
Many different randomization schemes exist that randomize the original database in different ways, calling for the method to choose the best one that is most suitable for a specific application context.

The equivalence class setup reflects the way the original domains of all attributes are reorganized and randomized. Consequently, the equivalence class definition is inherent in a randomization process and thus we use it as an essential parameter when comparing different randomization schemes.

To make comparisons more fair, we enforce another rule - the equivalence classes in different randomization schemes should form a total order relationship, with the order being defined as follows.

Definition 5.1 *We say the equivalence classes in two randomization schemes, denoted as R_1, R_2, form a **less than order** relationship, $R_1 \subseteq R_2$ if and only if we can obtain R_1 by dividing some equivalence classes R_2 into more equivalence classes.*

Note that we have defined privacy in two different ways, and they are bound privacy and top q-percentile privacy. For q-percentile privacy, we will choose q values to be 5 and 100. By letting q be 5 we focus on the privacy of a few records with the worst privacy breaches, which may help us better understand and explain bound privacy results. With q being 100 we are computing the average of all records.



(a) Bound Risk (b) Top 5-percentile Risk (c) Top 100-percentile Risk

Figure 1. Comparison between Randomization and K-Anonymization

Our intuition for the comparison is that under the constraint of the total order relationship, the randomization scheme with a smaller number of equivalence classes will better preserve privacy when utilities are the same.

6 Evaluation

We use the adult database from UCI data mining repository in our evaluations. There are 14 attributes in adult database, and we used 5 of them as identifying attributes.

6.1 Comparisons between K-Anonymization and Randomization

To generate R-U maps for K-Anonymity, we apply Incognito algorithm to adult database, with K being 1, 2, 3, and 4. The R-U Maps of K-Anonymity and randomization schemes are shown in figures 1(a), 1(b), and 1(c).

As shown in the figures, the KL-distances of generalized databases are much bigger than those of randomized databases when $K > 1$. Note that there is a big gap between bound risks 1 and 0.5 because K can only take integer values.

From these figures, we can clearly see that randomization outperforms generalization used in K-Anonymity. It is mainly due to the fact that we can partially recover the original data distribution in the randomization scheme, whereas data distribution within each equivalence class is totally lost in generalization.

6.2 Comparisons of different Randomization schemes

To compare different randomization schemes, we generate a R-U map for each randomization scheme. As stated in section 5, we use the number of equivalence classes as the characteristic parameter of each R-U map, specifically, we choose the numbers to be 5, 8, and 11. We also draw R-U maps for three privacy definitions, i.e., Bound privacy, top 5-percentile privacy, and top 100-percentile privacy.

Figure 2(a), 2(b), and 2(c) show the R-U Map of Adult database, with different privacy definitions. From these figures, we can see clearly the conflicting trend of risk and utility. Obviously and naturally, there is a trade-off between minimizing risk and maximizing utility (in our figures, the smaller the KL-distance, the better the utility.) The risk is decreasing approximately linearly with the decreasing of the KL-distance of the disguised database. At the beginning, when the distortion of the original database is little, the utility of the disguised database is very high (KL-distance is zero), and the risk is also very high. With the decreasing of the utility of the disguised database, the speed of decreasing of risk slows down.

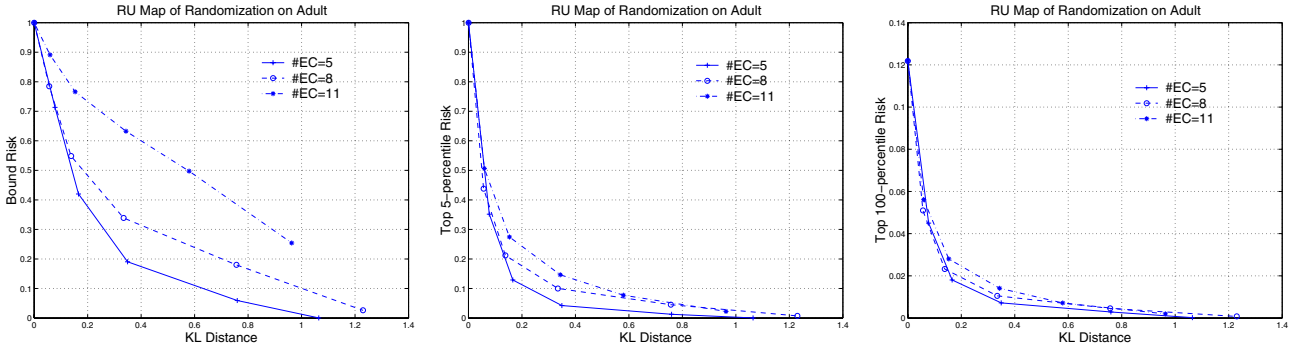
The smaller the total number of equivalence classes, the better the R-U map. By better we mean that with any given required utility, we may choose a parameter setup in the better randomization scheme that yields a less privacy breach.

For each total number of equivalence classes, the top 5-percentile risk is slightly smaller than the bound risk, which indicates that the worst case privacy is not a singular point, instead, there are some risks close to it.

7 Conclusions and Future Work

In this paper, we define a new privacy in the case of linking attacks, which makes it possible for us to compare K-Anonymity and randomization schemes in the same framework.

We promote to use the Risk-Utility Map to compare the effectiveness of different data disguising techniques. The curves in Risk-Utility Map clearly show the relations between disclosure risk and utility of the published database in



(a) Bound Risk (b) Top 5-percentile Risk (c) Top 100-percentile Risk

Figure 2. Comparison of various Randomization Schemes

the disguised form. The R-U maps shown in the evaluations clearly indicate the relative performances of the two data disguising schemes. In practical situations, these curves may help us decide which data disguising technique to use.

Generally speaking, randomization schemes outperform generalization in terms of achieving better privacy protection while yielding the same utility of the disguised database. As a result, we prefer randomization over generalization used in K-anonymity. However, generalization schemes do not require the modification of user's application software, since the disguised database contains the most accurate information given the way it is generated. On the other hand, randomization schemes require the modification of user's application software, of course, with the benefits of better estimated data distribution.

In randomization schemes, we prefer to use more equivalence classes in terms of maximizing data utility and minimizing privacy breaches. The extreme case of this is to let all domains of all attributes form a single equivalence class, as shown in the case of 5 equivalence class in evaluation section. However, the less the equivalence classes, the more the computation time. By forming only one equivalence class, the reconstructing of the original data distribution in the data mining or other data applications will take much longer time, slowing down the practical software.

In the practical applications of these techniques, whether randomization or generalization is the most appropriate data disguising technique depends on the context of the application. A trade-off must be made between privacy protection and data utility according to R-U maps.

In the future, we will evaluate generalization and randomization schemes in more details, for example, different ways to define equivalence classes. We will also apply the methodology to other data disguising techniques.

References

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD on Management of Data*, pages 439–450, Dallas, TX USA, May 15 - 18 2000.

[2] W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In *Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 505–510, Washington, DC, USA, August 24-27 2003.

[3] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 2002.

[4] G.T.Duncan, S.A.Keller-McNulty, and S.L.Stoke. Disclosure risk vs. data utility: The r-u confidentiality map. *Los Alamos National Laboratory Technical Report, LA-UR-01-6428*.

[5] K. LeFevre, D. J. Dewitt, and R. Ramakrishnan. Incognito:efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD*, June 12 - June 16 2005.

[6] S. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *Proceedings of the 28th VLDB Conference*, Hong Kong, China, 2002.

[7] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998.

[8] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, March 1965.

[9] Y. Zhu and L. Liu. Optimal randomization for privacy preserving data mining. In *the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2004.