12-11-2016

Assessing and Mitigating Disclosure Risk with Multiple Record Linkage

Hasan Kartal
University of Massachusetts Lowell, hasan_kartal@student.uml.edu

Xiao-Bai Li
University of Massachusetts Lowell, xiaobai_li@uml.edu

Assessing and Mitigating Disclosure Risk with Multiple Record Linkage

Completed Research Paper (Extended Abstract)

Hasan B. Kartal, Xiao-Bai Li
Department of Operations and Information Systems
Manning School of Business
University of Massachusetts Lowell
Lowell, MA 01854, U.S.A.
hasan_kartal@student.uml.edu
xiaobai_li@uml.edu

Abstract

This study examines privacy disclosure risks when multiple records in a dataset are associated with the same individual. Existing data privacy approaches typically assume that each individual in a dataset corresponds to a single record, which tends to underestimate the disclosure risks in the multiple-record problems. We propose a novel privacy approach, which uses a measure called g-balance to assess identity disclosure risk and another measure called h-affiliation to assess sensitive value disclosure risk in the multiple-record scenario. We develop an efficient algorithm based on the proposed measures to protect privacy disclosure due to multiple record linkage. An experimental study was conducted using real-world healthcare data with multiple records per person. The results of the experiments demonstrate that the proposed approach is more effective than traditional techniques in protecting privacy and preserving data quality.

Keywords: Privacy, *k*-anonymity, *l*-diversity, Gini index, kd-trees

Introduction

In analyzing privacy disclosure risk, it is well recognized that there are two types of disclosure (Duncan and Lambert 1989): (a) *identity disclosure* or *re-identification*, which occurs when a data intruder is able to match a record in a dataset to an individual; and (b) *attribute disclosure* or *sensitive value disclosure*, which occurs when an intruder is able to predict the sensitive value(s) of an individual record, with or without knowing the identity of the individual. To prevent re-identification, a well-known technique called *k*-anonymity (Sweeney 2002) generalizes the QI attribute values so that each record in a released dataset cannot be distinguished among at least *k* records. We call the group of records sharing the same QI values a *QI-group*. The *k*-anonymity method, as well as almost all of the existing privacy-preserving methods, assumes that each individual corresponds to a single record (Fung et al. 2010).

Many real-world datasets, such as patient diagnosis records, phone call records, credit card transactions, and web page clickstreams, often consist of multiple records for each individual (El Emam et al. 2009). When multiple records in a dataset are associated with the same individual, a QI-group of *k* records may contain a smaller number of individuals than *k*. Consequently, the occurrence of multiple records of an individual undermines the protection. To illustrate the problem, consider a hypothetical example shown in Table 1.

Table 1a shows identity, three QI attributes (Age, Gender, and ZIP) and one sensitive attribute (Disease) of a medical dataset from a hospital database. Each visit of a patient has a registration number (Reg. No), and some patients have multiple visits. Table 1b is the anonymized version of Table 1a, where records are grouped by age, gender and ZIP, and QI-groups are separated by lines. Because each group contains at least two records of indistinguishable QIs, the table satisfies the requirements of *k*-anonymity where $k = 2$. However, it is clear that Ashley, who appears alone in the first group twice, is not well protected.

Reg.No	Name	Age	Gender	ZIP	Disease
1	Ashley	32	Female	23000	Asthma
2	Ashley	32	Female	23200	Obesity
3	Bob	36	Male	21750	Obesity
4	Charlie	49	Male	22100	Diabetes
5	Charlie	49	Male	22100	Diabetes
6	Charlie	49	Male	22100	Gastritis
7	Charlie	49	Male	22100	Gastritis
8	Charlie	49	Male	22100	Diabetes
9	Diana	38	Female	24200	Ulcer
10	Diana	38	Female	24200	Gastritis
11	Edward	36	Male	23500	Diabetes
12	Edward	36	Male	23500	Gastritis
13	Fred	40	Male	23600	Diabetes
14	Greg	41	Male	24800	Ulcer
15	Harry	45	Male	24050	Ulcer
16	Harry	45	Male	24050	Asthma
17	Harry	45	Male	24050	Epilepsy
18	Harry	45	Male	24050	Asthma
19	Harry	45	Male	24050	Ulcer

a. Original Dataset

Reg.No	Age	Gender	ZIP	Disease
1	32	Female	23000-23200	Asthma
2	32	Female	23000-23200	Obesity
3	36-49	Male	21750-22100	Obesity
4	36-49	Male	21750-22100	Diabetes
5	36-49	Male	21750-22100	Diabetes
6	36-49	Male	21750-22100	Gastritis
7	36-49	Male	21750-22100	Gastritis
8	36-49	Male	21750-22100	Diabetes
9	36-38	*	23500-24200	Ulcer
10	36-38	*	23500-24200	Gastritis
11	36-38	*	23500-24200	Diabetes
12	36-38	*	23500-24200	Gastritis
13	40-45	Male	23600-24800	Diabetes
14	40-45	Male	23600-24800	Ulcer
15	40-45	Male	23600-24800	Ulcer
16	40-45	Male	23600-24800	Asthma
17	40-45	Male	23600-24800	Epilepsy
18	40-45	Male	23600-24800	Asthma
19	40-45	Male	23600-24800	Ulcer

b. Anonymized Dataset

Table 1. An Example Dataset Vulnerable to MRL Attacks

Traditional record-linkage attack models assume that the intruder knows the QI values of a target (Sweeney 2002; Fung et al. 2010). In multiple records scenarios, the intruder also knows that the target may have multiple records in the released dataset. For example, the intruder may know that his target was hospitalized several times in a certain period. So, his target's medical information would appear multiple times in the dataset released by the hospital. There are two records in the first QI group of Table 1b, both matching Ashley's QI values. If the intruder knew that Ashley visited the hospital twice, he would easily

conclude that both records must be Ashley. Regarding the second QI-group, Charlie has 5 out of 6 records. The intruder would be able to re-identify Charlie with a quite high probability, and Charlie has a higher disclosure risk than Bob. The intruder would be able to re-identify Charlie with a quite high probability, and Charlie has a higher disclosure risk than Bob. In many cases, a system generated patient ID may be provided by the data publisher in order for the data user to perform analysis at an individual level. This would make it even easier to identify the individuals by their occurrences in a QI-group. In this paper, we refer to the privacy attacks due to the existence and knowledge of one person having multiple records a *multiple record linkage* (**MRL**) attack.

QI-Group	Name	Number of Records	Probability of Re-identification
1	Ashley	2	100.0%*
2	Bob	1	16.7%
	Charlie	5	83.3%*
3	Diana	2	50.0%
	Edward	2	50.0%
4	Fred	1	14.3%
	Greg	1	14.3%
	Harry	5	71.4%*

*Under-protected if re-identification risk is not allowed to be larger than 50%.

Table 2. Probabilities of Re-identification for the Individuals in Table 1b

With the assumption of single record a person, the probability of linking a target to a specific individual using QI values is at most $1/k$ in a k-anonymized table. This is no longer true when an individual can have multiple occurrences. Instead, re-identification risk can be assessed based on individuals' occurrence frequencies. In a QI-group containing k individuals, let f_i be the number of records associated with the ith individual, the probability of re-identifying this individual can be defined by $f_i/\sum_{i=1}^{k} f_i$. Table 2 shows these probabilities for the example dataset in Table 1. QI-groups 1, 2 and 4, which satisfy (record-based) *2*-anonymity, fail to provide specified protection for some patients. That is, the probabilities of re-identification for some individuals in these groups are higher than 50% (marked with *).

The *k*-anonymity model considers re-identification risk but not attribute disclosure risk. Therefore, even when the re-identification risk of an individual is sufficiently limited in a QI-group, attribute disclosure may occur when there is little diversity in the values of a sensitive attribute. To address this problem, the *l*-diversity principle (Machanavajjhala et al. 2006) requires that each QI-group contains at least *l* well-represented values. However, *l*-diversity also assumes a single record per person; it is thus not effective in reducing attribute disclosure risk in MRL scenario. For example, in the first QI-group in Table 1b, even though the QI-group is 2-diverse, both sensitive values, Asthma and Obesity, are disclosed with certainty. This is because both values are affiliated with the same patient Ashley. Therefore, a more thorough approach is required to counter MRL attacks.

Essentially all of the existing approaches assume that each individual corresponds to a single record. To address the MRL attack problem, Tao et al. (2008) propose an approach that considers *K* distinct individuals instead of *k* records. We call this approach distinct *K*-anonymity (with a capital letter *K*). Distinct *K*-anonymity, however, does not consider risk caused by unbalanced frequency distribution in case of the MRL problem. Our study identifies a new identity disclosure problem based on occurrence distributions of the individuals. We investigate the problems of identity and sensitive value disclosures under MRL attacks. We propose a novel approach, which uses a measure called *g*-balance to assess identity disclosure risk and another measure called *h*-affiliation to assess sensitive value disclosure in MRL scenario. Proposed approach overcomes the drawbacks of existing measures and promises a better protection against privacy breaches. We construct a metric representing the trade-off between privacy measure and data quality. Using this trade-off metric, we develop an efficient algorithm for protection against privacy attacks. Furthermore, we conduct an experiment to evaluate our approach using real-world data.

Disclosure Risk Measures

The existence of multiple records causes individuals within a QI-group to have different disclosure risks. The basic idea of our approach for reducing identity disclosure risk is to create QI-groups that contain a sufficient number of individuals with relatively balanced occurrence distributions. The proposed measure is based on the classical Gini index (Breiman et al. 1984) and is called g-balance.

Definition 1 (*g-Balance*): Let n be the number of distinct PIDs in T, and c_i be the number of occurrences for the ith PID in T. The ***g-balance*** of T is defined as:

$$g = 1 - \sum_{i=1}^{n} \left(\frac{c_i}{\sum_{j=1}^{n} c_j} \right)^2 \quad (1)$$

A larger g value indicates a more balanced occurrence distribution in T, which suggests a better protection against re-identification after the QI values are generalized. The g-balance measure can be defined similarly for any subset of T, particularly for a QI-group of T. Our proposed method uses binary recursive partitioning to split table T into smaller sub-tables to form QI-groups. Our proposed method uses binary recursive partitioning to split table T into smaller sub-tables to form QI-groups. We denote the parent table for a split by T_p and the two child tables of T_p by T_1 and T_2. It can be shown that the g value before a split is always greater than or equal to the weighted average g value after the split (Breiman et al. 1984). Such a decrease in g-balance value implies an increase in re-identification risk. To measure this difference, we define g-balance reduction below.

Definition 2 (*g-Balance Reduction*): Let g_p, g_1 and g_2 be the g-balance values for T_p, T_1 and T_2, respectively. Let c_{i_p}, c_{i_1} and c_{i_2} be the number of occurrences of the i_pth PID in T_p, the i_1th PID in T_1 and the i_2th PID in T_2, respectively, where $\sum c_{i_p} = \sum c_{i_1} + \sum c_{i_2}$. The *g-balance reduction* from splitting T_p into T_1 and T_2 is:

$$\Delta g(T_p) = g_p - \frac{\sum c_{i_1}}{\sum c_{i_p}} g_1 - \frac{\sum c_{i_2}}{\sum c_{i_p}} g_2 \quad (2)$$

Next, we consider sensitive value disclosure risk. As discussed in the introduction, the traditional l-diversity principle is not applicable for MRL scenarios. Different individuals in a QI-group may have very diverse sensitive values. The attribute disclosure risk of a QI-group can be determined by the sensitive value that is affiliated with the largest proportion of the individuals in the group.

Definition 3 (*h-Affiliation*): Let n and m be the number of PIDs and sensitive values in T (or in any subset of T), respectively, and s_j be the number of PIDs affiliated with the jth sensitive value in T (or in any subset of T). The ***h-affiliation*** of T (or any subset of T) is defined as:

$$h = \max_{j=1,\ldots,m} \frac{s_j}{n} \quad (3)$$

Clearly, a larger h value suggests a higher sensitive value disclosure risk. It can be shown that *when a dataset is partitioned into subsets, h-affiliation for at least one subset will be greater than or equal to the h-affiliation of the original set.*

Our recursive partitioning method adopts the idea of the well-known kd-tree technique (Friedman et al. 1977), where each split is determined based on the variance of the QI attributes. Typically, the QI attribute with the largest variance at each iteration is used to split the data as this will result in the most significant reduction in variance in the partitioned data. A lower variance in a QI-group leads to a better data utility because it causes a smaller information loss after QI values within the group are generalized.

Proposed Algorithm

There are two objectives in our partitioning process: (1) to minimize identity and sensitive value disclosure risks, which means to keep reduction in g-balance and increase in h-affiliation as small as possible, and (2) to reduce the variance of the data as much as possible to minimize information loss after generalization. With g-balance and variance measures, we propose the following splitting criterion that represents the trade-off between privacy protection and data quality.

Definition 4 (*Balance-Variance ratio*): Let t be the T table or a sub-table of T, and v_j *be variance of* the jth QI attribute. The *balance-variance ratio* for splitting t on the jth QI attribute is defined as:

$$r_j(t) = \Delta g(t)/v_j \quad (4)$$

The balance-variance ratio represents the marginal decrease in g-balance per unit variance of a QI attribute. Since a small g-balance reduction and a large variance are preferred, the QI attribute that has the minimum balance-variance ratio should be selected for partitioning the data at each iteration. The proposed algorithm recursively splits data into two subsets at the median of the QI attribute having the minimum balance-variance ratio. If the QI attribute is an ordered categorical type, the split is made at the between-category point that is closest to the median among all between-category points. Table 3 describes the steps of the proposed algorithm.

	Input: Table T and user specified privacy requirement parameters values g^* and h^*.
Step 1	For the current table t, compute $r_j(t)$ for each QI attribute j. Let j^* be the QI with the minimum $r_j(t)$.
Step 2	(i) Split t into two sub-tables at the median of attribute j^*. (ii) If the g-balance value of any sub-table of t is smaller than g^* or h-affiliation value of any sub-table of t is greater than h^*, undo split and set j^* to the QI attribute with the next smallest $r_j(t)$ and go to (i). Stop splitting if no QI attribute can be assigned to j^*.
Step 3	Repeat Steps 1 and 2 for each sub-table until no further split can be made.
Step 4	Generalize the QI values using the ranges or means of the QI values in each sub-table.

Table 3. The Proposed Algorithm

Experimental Evaluation

Due to the page limitation, this section is omitted. It is available upon request.

References

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. 1984. *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Duncan, G. T., Lambert, D. 1989. "The Risk of Disclosure for Microdata," *Journal of Business and Economic Statistics* (7:2), pp. 201-217.

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., and Lysyk, M. 2009. "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records," *The Canadian Journal of Hospital Pharmacy* (62:4), pp. 307-319.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. 1977. "An Algorithm for Finding Best Matches in Logarithmic Expected Time," *ACM Transactions on Mathematical Software* (3:3), pp. 209-226.

Fung, B., Wang, K., Chen, R., and Yu, P. S. 2010. "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys* (42:4), pp.14.

Machanavajjhala A, Gehrke J, Kifer D, and Venkitasubramaniam M. 2006. "*l*-Diversity: Privacy beyond *k*-Anonymity," in *Proceedings of the 22nd IEEE International Conference on Data Engineering*, Washington, DC: IEEE Computer Society, pp. 24-35.

Sweeney, L. 2002. "*k*-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* (10:05), pp. 557-570.

Tao, Y., Tong, Y., Tan, S., Tang, S., and Yang, D. 2008. "Protecting the Publishing Identity in Multiple Records," *Data and Applications Security XXII*, pp. 205-218.