Probabilistic Privacy Analysis of Published Views

Hui (Wendy) Wang
Department of Computer Science
University of British Columbia
Vancouver, Canada
hwang@cs.ubc.ca

Laks V.S. Lakshmanan
Department of Computer Science
University of British Columbia
Vancouver, Canada
laks@cs.ubc.ca

ABSTRACT

Among techniques for ensuring privacy in data publishing, k-anonymity and publishing of views on private data are quite popular. In this paper, we consider data publishing by views and develop a probability framework for the analysis of privacy breach. We propose two attack models and derive the probability of privacy breach for each model.

Categories and Subject Descriptors: H.2 Database Management

General Terms: Security

Keywords: Published views, private association, privacy breach, probabilistic analysis.

1. INTRODUCTION

Many organizations are increasingly publishing relational data that contains personal information. However, since the data contains personal information, protecting individual privacy is an important issue. We assume that the private information takes the form of associations, which are the pairs of values in the same tuple. E.g., as shown in the base table in Figure 1 (a), "George" is associated with "HIV". Neither "George" nor "HIV" alone is a secret, but the association between them is. To protect such privacy information, we need to process the data before we publish it.

There are several approaches to process data so that the private association can be hidden in the published data. One possible solution is to suppress/generalize some of the entries in the table so that for each tuple in the modified table, there are at least k-1 other tuples in the modified table that are identical on the quasi-identifying attributes. Therefore the private association cannot be distinguished from the other k-1 associations. This approach is called k-anonymity [8]. This is one of the most dominant approaches for privacy preserving data publishing. Recently, this field has flourished with a number of research focusing on complexity issues and on extension of k-anonymity models (e.g., [3], [4], [5]).

ID	Name	Age	Job	Problem
1	Bill	30	Engineer	Cold
2	John	45	Professor	Diarrhea
3	George	45	Professor	HIV
4	Alan	42	Engineer	Cold
5	Sarah	45	Engineer	Cold

(a) Base Table T

Name	Age
Bill	30
John	45
George	45
Alan	42
Sarah	45

(b) V_1=$\Pi_{Name,Age}(T)$

Age	Job	Problem
30	Engineer	Cold
45	Professor	Diarrhea
45	Professor	HIV
42	Engineer	Cold
45	Engineer	Cold

(c) V_2=$\Pi_{Age,Job,Problem}(T)$

Figure 1: An Example

One of the drawbacks of k-anonymity is that modified data values in the published data will result in incorrect data analysis and mining result. E.g., from the medical table in Figure 1, the researchers want to study what are the ages that people are more likely to develop heart disease, or what diseases engineers are more likely to have. But if the data of `Age` and `Job` attributes are generalized in the released data for k-anonymity purpose, none of the results can be obtained anymore, i.e., the anonymized data becomes unusable for certain types of research.

Another possible solution for ensuring privacy of value associations in published data is to decompose the base table into several view tables. Thus the private association between the data values will be broken when they are put into different view tables [12]. No data value in the published views will be altered. From this view point, the published data is still useful for data analysis. However, as pointed out by [12], careless design of view tables still enables possibilities of privacy breach. A formal definition and a thorough study of the privacy breach is an important problem that has not received much attention. This is the main focus of our paper. Specifically, based on notions of *possible worlds* and *interesting worlds*, we propose a generic framework to measure privacy breach. We propose two different attack models and for every model, we analyze the likelihood of privacy breach by using probability theory.

Our contributions include: (1) we propose the *secret query* notion to specify private associations of a given database (Section 3.1). The secret query can be expressed in SQL. (2) We define two attack models and define the probability of privacy breach for each model (Section 3.2). (3) We propose a novel structure, *connectivity graph*, as the synopsis of the base table (Section 4). (4) By using the connectivity graph,

WPES'06, October 30, 2006, Alexandria, Virginia, USA.

we derive the formulas of quantifying the probabilities of privacy breach for each attack model (Section 5).

Related work is discussed in Section 2. We summarize the paper by Section 6. We assume that the base table is split into two projection views that contain no duplicates.

2. RELATED WORK

The problem of protecting privacy by published views has been studied from various aspects. The work by Chao et al. [12] is the closest related to ours. They studied the problem that for a set of secret associations, whether the released views violates any k-anonymity constraint. They use association cover to measure the information disclosure and by their definition, the released views violates k-anonymity requirement if there are association covers of size less than k. However, we consider a different measurement of information leakage by using probabilistic analysis. We define two attack models, the *unrestricted* and the *restricted* model, and measure the privacy breach of each model. We show that the probability of restricted model is equivalent to $1/n$, where n is the size of association cover by [12].

Miklau et al. [7] studied the problem that whether publishing view V logically discloses any information about a confidential query Q. Deutsch et al. [2] re-studied the problem by taking the additional correlations between tuples into consideration. For both [7] and [2], the published views V is secure if the posteriori probability of any possible answer to Q is the same as its priori probability. They both assume that the domain knowledge is available to the attacker so that he can reason of a priori probability, while we assume there is no such background knowledge to the attacker. Moreover, both [7] and [2] assume the attacker knows how to compute the probability. Thus they mainly focus on the complexity result for deciding whether a query Q is secure w.r.t. the published view V. However, we focus on the details of computation of the probabilities for different attack models. Furthermore, we focus on the protection of associations, while they consider the protection of the answer of a particular secret query.

Besides the probability model, *k-anonymity* is another approach of measurement of information disclosure. There has been much research done on *k-anonymity* problem (e.g., [1], [6], [8], [9]). Machanavajjhala et al. [5] and Xiao et al. [11] extend k-anonymity model by taking diversity and personalized preference into consideration. Most of the work focus on how to generalize and suppress the base table so that the released data will reach k-anonymity. None of them considers publishing views without altering the data values.

3. SECURITY MODEL

In this section, we define the private association definition and attack models.

3.1 Private Association

As in [12], we consider that the base table T contains an identifier attribute ID and a property attribute P, where ID uniquely identifies an entity (E.g., the attribute `Name` in Figure 1 (a)), and P is a private property of the entity (E.g., the attribute `Problem` in Figure 1 (a)). We specify the private association α as $(ID = i, P = p)$, which denotes that the fact that the entity i is associated with property p must be protected. Every private association α can be specified as a *secret query* q^{α}, which is a SQL query. ID and P are put in the `select` clause and value-based constraints in the `where` clause. E.g., the association `(Name=*, Problem=Cold)` can be specified as "`select Name, Problem from T where Problem='Cold'`". *We assume for any association (ID, P), each value on ID is associated with at most one value on P in the base table.*

We adapt the definition of *association cover* in [12] as follows:

Definition 3.1. **[Association Cover]** *For each association $\alpha(ID = i, P = p)$ and a given view $\mathcal{V}$, whose join result is $\mathcal{J}$, we say $\alpha'(ID = i', P = p')$ is in the* association cover *$C^{\alpha}_{\mathcal{V}}$ of α if: (1) $q^{\alpha'}(\mathcal{J}) \neq \oslash$, (2) $\alpha.ID = \alpha'.ID$, (3) $\forall \alpha'' \in C^{\alpha}_{\mathcal{V}}$ that $\alpha'' \neq \alpha$, $\alpha''.P \neq \alpha'.P$, i.e., the value of property attribute P is unique.* □

Note that it is possible that $\alpha \notin C^{\alpha}_{\mathcal{V}}$. E.g., from the join result of two views V_1 and V_2 in Figure 1 (b) and (c), the cover of association `α(Name='Alan', Problem='HIV')` is `(Name='Alan', Problem='Cold')`.

3.2 Probabilistic Models

Unrestricted model

By access to the published views, the attacker can try to reason about the *possible* base tables that would lead to the same tables as $\mathcal{V}$ by using the same view definitions of $\mathcal{V}$. Notice that hiding the view definitions from the attacker doesn't really help, so we should consider the case where the view definitions are known to the attacker. We first formalize this idea next, where we call each possible database an *unrestricted possible world.*

Definition 3.2. **[Unrestricted Possible Worlds]** *Given a base table T and a view $\mathcal{V}$={V_1, V_2, ..., V_n}, let q_i be the definition of view V_i. Then the unrestricted possible worlds of T UPW_T={T' | T' is a relation of the same schema as T, and $\forall$ view definition $q_i, q_i(T') = q_i(T)$.}* □

To understand the definition better, let's look at an example. Suppose the base table shown in Figure 2(a) is decomposed into two views, $V_1 = \Pi_{A,B}(T)$ and $V_2 = \Pi_{B,C}(T)$ (Figure 2 (b) and (c)). The unrestricted possible worlds UPW_T are shown in Figure 2 (d). By applying the same view definition $\Pi_{A,B}$ and $\Pi_{B,C}$ on each possible world, it yields the same view tables as V_1 and V_2.

Out of all *possible worlds*, there is only a subset that contain the private association. Only those can help the attacker infer the existence of a private association in the base table. We call those databases the *interesting worlds.* To be formal,

Definition 3.3. **[Unrestricted Interesting Worlds]** *Given a table T and an association α, whose corresponding query is q^{α}. The unrestricted interesting worlds of α w.r.t T UIW^{α}_T={T' | $T' \in UPW_T$, and $q^{\alpha}(T') \neq \oslash$}.* □

As an example, from the seven unrestricted possible worlds in Figure 2, for the association $\alpha(A = a_1, C = c_1)$, UIW^{α}_T= {w_1, w_3, w_4, w_5, w_7}.

Restricted Model

The attacker may be smarter: she may know that for any private association, each entry on its ID has at most one corresponding tuple in the base table. Thus she only considers the database candidates in which the association cover of α is of size 1. To be more precise,

Definition 3.4. **[Restricted Possible Worlds]** *Given a base table T and a view $\mathcal{V}$={V_1, V_2, ..., V_n}, let α be the private association. Then the restricted possible worlds of α w.r.t T RPW^{α}_T={T' |$T' \in UPW_T$, and | $C^{\alpha}_{\mathcal{V}}$ |=1.}* □

A	B	C
a1	b1	c1
a2	b1	c2

(a) base table T

A	B
a1	b1
a2	b1

(b) $V_1 : \Pi_{A,B}(T)$

B	C
b1	c1
b1	c2

(c) $V_2 : \Pi_{B,C}(T)$

ID	Possible World	ID	Possible World
w_1	{a1, b1, c1}, {a2, b1, c2}	w_2	{a1, b1, c2}, {a2, b1, c1}
w_3	{a1, b1, c1}, {a2, b1, c2} {a1, b1, c2}	w_4	{a1, b1, c1}, {a2, b1, c2}, {a2, b1, c1}
w_5	{a1, b1, c2}, {a2, b1, c1}, {a1, b1, c1}	w_6	{a1, b1, c2}, {a2, b1, c1}, {a2, b1, c2}
w_7	{a1, b1, c2}, {a2, b1, c1}, {a1, b1, c1}, {a2, b1, c2}		

(d)Possible Worlds

Figure 2: An example

Possible World	Interesting World	Probability Model	Probability
Unrestricted	Unrestricted	Unrestricted	$\frac{\lvert UIW\rvert}{\lvert UPW\rvert}$
Restricted	Restricted	Restricted	$\frac{\lvert RIW\rvert}{\lvert RPW\rvert}$

Figure 3: Various Probabilitis Models

For the same example, the restricted possible worlds $RPW_T^\alpha = \{w_1, w_2, w_4, w_6\}$, where in each world a_1 is associated with either c_1 or c_2, but not both of them at the same time. Note that the association (a_1, c_2) doesn't exist in the base table. However, it exists in RPW_T^α.

The attacker will apply the same reasoning on the interesting worlds. Formally,

Definition 3.5. **[Restricted Interesting World]** *Given a table T, a view $\mathcal{V}$, and an association α, whose corresponding query is q^α. The restricted interesting worlds of α w.r.t T $RIW_T^\alpha = \{T' \in RPW_T^\alpha$, and $q^\alpha(T') \neq \oslash$.$\}$* □

Still using the same example above, from the four restricted possible worlds $\{w_1, w_2, w_4, w_6\}$, $RIW_T^\alpha = \{w_1, w_4\}$.

We assume every possible world and interesting world are equally likely. Based on *unrestricted* and *restricted* versions of possible and interesting worlds, we defined two different models of probability (Figure 3) of which the attacker can infer the existence of a private association in the original base table.

For the running example, the probability of unrestricted model is 5/7. And the probability of restricted model is 2/4=1/2.

4. CONNECTIVITY GRAPH

The naive method of measuring the probability is to reconstruct both possible worlds and interesting worlds and count their numbers. However, since the numbers of both worlds are exponential in the number of tuples in the views, the computational cost may be expensive. Thus we use connectivity graph to help measure the probability of security breach. In this section, we explain the details of connectivity graph.

Given a view $\mathcal{V}$, the connectivity graph is constructed by the following steps:

Step 1: Locate Relevant Tuples

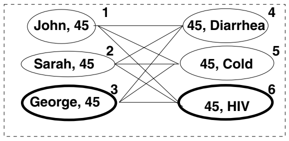

Figure 4: Connectivity Graph of 2 Views

For a given base table T and an association α, we first locate the tuples in T that contains α by applying q^α on T, where q^α is the corresponding query of the association α. Then we locate those tuples $t \in T$ s.t. $\exists t' \in q^\alpha(T)$ s.t t and t' have the same values on the join attributes of the views. For example, for the base table T in Figure 1 (a) and the association α(Name='George', Problem='HIV'), $q^\alpha(T)$={tuple 3}. Based on the two released views in Figure 1 (b) and (c) whose join attribute is Age, since tuple 3 has the same value 45 of attribute Age as that of tuples 2 and 5, tuples 2 and 5 are also located as the "relevant tuples" of α. After we locate the relevant tuples in the base table, we highlight the corresponding tuples in the view tables.

Step 2: Construct the Connectivity Graph

We construct a connectivity graph CG as follows: for every view V_i, let U_{V_i} be the set of the attributes of V_i that are involved in either the association or the join attributes. Then for every highlighted tuple $t \in V_i$, there is a node $n \in CG$ corresponding to $\Pi_{U_{V_i}}(t)$ (i.e., the projection on attributes U_{V_i} of tuple t). For every two nodes n_1 and n_2, there is an edge between n_1 and n_2. As an example, Figure 4 shows a connectivity graph for two views $V_1 = \Pi_{Name,Age}(T)$ and $V_2 = \Pi_{Age,Job,Problem}(T)$ in Figure 1 (b) and (c). In this example, U_{V_1}={Name, Age} and U_{V_2}={Age, Problem}.

One nice property of the connectivity graph constructed from two views is that it is always a 2-partite complete graph (e.g., the connective graph CG in Figure 4). We will show how to use this property to calculate the probability of privacy breach in Section 5.

5. PROBABILITY OF PRIVACY BREACH

In this section, we discuss how to calculate the probabilities of both models based on the connectivity graph. We first define *unrestricted cover* and *interesting unrestricted cover* of the connectivity graph.

Definition 5.1. **[Unrestricted Cover]** *Let $G = (V, E)$ be a 2-partite graph, where $V = V_1 \cup V_2$. Let $l \in V_1$ and $r \in V_2$ be two distinguished nodes. An* unrestricted cover *of G is a subset of edges $C \subseteq E$ such that for every node $v \in V$, there is some edge $e \in C$ such that e is incident on v. Furthermore, an unrestricted cover C is said to be* interesting *provided C contains at least one path from l to r.* □

Intuitively the *unrestricted cover* is an edge set s.t. every node is covered by at least one edge in it. E.g., one unrestricted cover of the connectivity graph of Figure 4 is $\{<1, 4>, <2, 5>, <3, 6>, <1, 6>\}$. If we specify l as node 1 and r as node 6, this cover is an *interesting* unrestricted cover.

By matching with the definition of unrestricted possible & interesting world, we have:

Lemma 5.1. **[Unrestricted Covers V.S. Unrestricted Worlds]** *Given a base table T and a view $\mathcal{V}$, let CG be the corresponding connectivity graph of $\mathcal{V}$. Then for any association $\alpha(ID = i, P = p)$, let n_i and n_j be the nodes in*

CG that correspond to the tuples that contain $ID = i$ and $P = p$. Then the number of unrestricted possible worlds is equal to the number of unrestricted covers *of CG. Furthermore, the number of unrestricted interesting worlds is equal to the number of* unrestricted interesting covers *of CG by treating n_i and n_j as the distinguished nodes l and r.* □

For example, for the 2-partite graph in Figure 4, Figure 5 shows a subset of unrestricted covers and the corresponding possible worlds. For the association α`(Name=''George'', Disease=''HIV'')`, whose corresponding nodes in Figure 4 are node 3 and 6, the unrestricted interesting worlds of α are the subset of unrestricted covers that contains the edge $< 3, 6 >$.

ID	Unrestricted Cover	Possible World
w_1	$< 1, 4 >$, $< 2, 5 >$ $< 3, 6 >$	{John, 45, Diarrhea}, {Sarah, 45, Cold}, {George,45, HIV}
...	...	...

Figure 5: Covers V.S. Possible Worlds

In general, we have:

LEMMA 5.2. **[Number of Unrestricted Possible & Interesting Worlds]** *For a 2-partite graph with two partitions consisting of n and m nodes respectively, the number of the unrestricted possible world*
$UPW(m, n) = \sum_{i=2}^{m} \sum_{j=2}^{n} \binom{m}{m-i} \binom{n}{n-j} (-1)^{m+n-i-j} 2^{i*j} - \sum_{i=2}^{m} \sum_{j=2}^{n} \binom{m}{m-i} \binom{n}{n-j} (-1)^{n+m-i-j} + (-1)^{m+1} m + (-1)^{n+1} n + (-1)^{m+n+1} (m * n)$.
The number of the unrestricted interesting world
$UIW(m, n) = \sum_{i=2}^{m} \sum_{j=2}^{n} \binom{m-1}{m-i} \binom{n-1}{n-j} (-1)^{m+n-i-j} 2^{i*j-1} + (-1)^{m+1} + (-1)^{n+1} + (-1)^{m+n+1}$. *Furthermore, when either m or n equals 1, $UPW(m, n) = UIW(m, n) = 1$.* □

For the above example, the 2-partite graph is partitioned into 2 disjoint sets, each of 3 nodes. Thus $UPW(3, 3) = 265$ and $UIW(3, 3) = 161$.

Then we come to the restricted model. We firstly define *restricted cover* and *interesting restricted cover* of a connectivity graph.

DEFINITION 5.2. **[Restricted Cover]** *Let $G = (V, E)$ be a 2-partite graph, where $V = V_1 \cup V_2$. Let $l \in V_1$ and $r \in V_2$ be two distinguished nodes. A* restricted cover *of G is a subset of edges $C \subseteq E$ such that for every node $v \in V$, there is some edge $e \in C$ such that e is incident on v. Furthermore, there is only one node in V_2 that l connects to. A restricted covering C is said to be* interesting *provided C contains the path from l to r.* □

Intuitively the restricted cover is an edge set that covers every node by at least one edge in it. Furthermore, there is only one property node that the ID node l is connected with. E.g., $\{< 1, 4 >, < 2, 5 >, < 3, 6 >\}$ is a restricted cover of the connectivity graph in Figure 4. If we specify l as node 3 and r as node 6, this cover is an *interesting* restricted cover.

By matching with the definition of restricted cover, we have:

LEMMA 5.3. **[Restricted Covers V.S. Restricted Worlds]** *Given a base table T and a view $\mathcal{V}$, let CG be the corresponding connectivity graph of $\mathcal{V}$. Then for any association $\alpha(ID = i, P = p)$, let n_i and n_j be the nodes in CG that correspond to the tuples that contain attribute $ID = i$ and $P = p$. The number of restricted possible worlds is equal to the number of* restricted covers *of CG. Furthermore, the number of restricted interesting world is equal to the number of* restricted interesting covers *of CG by treating n_i and n_j as two distinguished nodes l and r.* □

Probability Model	Probability
Unrestricted model	$\frac{UIW(m,n)}{UPW(m,n)}$
Restricted model	$\frac{UPW(m-1,n)+UPW(m-1,n-1)}{n\times(UPW(m-1,n)+UPW(m-1,n-1))} = \frac{1}{n}$

Figure 6: Probability Models

By reasoning, we have

LEMMA 5.4. **[Number of Restricted Possible & Interesting Worlds]** *For a 2-partite graph with two partitions consisting of n and m nodes respectively, the number of restricted interesting world $RIW(m, n) = UPW(m - 1, n) + UPW(m-1, n-1)$. And the number of the restricted possible world $RPW(m, n) = n \times (UPW(m-1, n) + UPW(m-1, n-1))$.* □

Figure 6 shows the probability of privacy breach for each attack model. Note that for the restricted model, the probability of privacy breach is $1/n$, where n equals to the size of the association cover of α by the k-anonymity model in [12]. E.g., for the view tables V_1 and V_2 shown in Figure 2, as we have already shown in Section 3.2, the probability of restricted model equals to 1/2, while by [12], the cover of association $\alpha(a_1, c_1)$ is of size 2 since a_1 will be associated with both c_1 and c_2 in the join result of V_1 and V_2.

6. CONCLUSION

In this paper, we defined a general framework to measure the likelihood of privacy breach for a set of published views. We proposed two attack models and for each model, we derive the formulas to calculate the probability.

In the future, we want to study the following problems: (1) for the database of very large size, the computation of the function $UPW(m, n)$ and $UIW(m, n)$ may be expensive. We want to study whether there exists any approximation to the probability calculations such that its computation complexity is not that high while the amount of numerical error is acceptable. (2) We will extend our work to the K-view-table case, where $k > 2$.

7. REFERENCES

[1] Roberto J. Bayardo, Rakesh Agrawal, "Data privacy through optimal k-anonymization", ICDE, 2005.

[2] Alin Deutsch, Yannis Papakonstantinou, "Privacy in Database Publishing", ICDT'05.

[3] Kristen LeFevre, David DeWitt, and Raghu Ramakrishnan, "Incognito: Efficient Full-domain K-anonymity", SIGMOD'05.

[4] Kristen LeFevre, David DeWitt, and Raghu Ramakrishnan, "Mondrian Multidimensional K-Anonymity", ICDE'05.

[5] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam, "l-Diversity: Privacy Beyond k-Anonymity". ICDE, 2006.

[6] Adam Meyerson, Ryan Williams, "On the complexity of optimal k-anonymity", PODS, 2004.

[7] Gerome Miklau, Dan Suciu, "A Formal Analysis of Information Disclosure in Data Exchange", SIGMOD'04.

[8] Pierangela Samarati, Latanya Sweendy, "Generalizing data to provide anonymity when disclosing information", PODS, 1998

[9] Pierangela Samarati, "Protecting Respondents' Identities in Microdata Release", IEEE TKDE 13(6): 1010-1027 (2001)

[10] Claude Elwood Shannon, "Communication theory of secrecy systems", Bell System Technical Journal, 1949.

[11] Xiaokui Xiao, Yufei Tao, "Personalized Privacy Preservation", SIGMOD, 2006.

[12] Chao Yao, X.Sean Wang, Sushil Jajodia, "Checking for k-Anonymity Violation by Views", VLDB'05.