Global Privacy Guarantee in Serial Data Publishing

Raymond Chi-Wing Wong[1], Ada Wai-Chee Fu[2], Jia Liu[2], Ke Wang[3], Yabo Xu[4]

[1]Hong Kong University of Science and Technology, [2]Chinese University of Hong Kong
raywong@cse.ust.hk, {adafu,jliu}@cse.cuhk.edu.hk
[3]Simon Fraser University, [4]Sun Yat-sen University
wangk@cs.sfu.ca, xuyabo@mail.sysu.edu.cn

***Abstract*— While previous works on privacy-preserving serial data publishing consider the scenario where sensitive values may persist over multiple data releases, we find that no previous work has sufficient protection provided for sensitive values that can change over time, which should be the more common case. In this work, we propose to study the privacy guarantee for such transient sensitive values, which we call the *global guarantee*. We formally define the problem for achieving this guarantee. We show that the data satisfying the global guarantee also satisfies a privacy guarantee commonly adopted in the privacy literature called the *local guarantee*.**

I. Introduction

Recently, there has been much study on the issues in privacy-preserving data publishing [5], [10], [4]. Most previous works deal with privacy protection when only one instance of the data is published. However, in many applications, data is published at regular time intervals. For example, the medical data from a hospital may be published twice a year. Some recent papers [6], [11], [3], [2], [8], [1] study the privacy protection issues for *multiple* data publications of multiple instances of the data. We refer to such data publishing *serial data publishing*.

Following the settings of previous works, we assume that there is a sensitive attribute which contains sensitive values that should not be linked to the individuals in the database. A common example of such a sensitive attribute is diseases. While some diseases such as flu or stomach virus may not be very sensitive, some diseases such as chlamydia (a sexually transmitted disease (STD)) can be considered highly sensitive. In serial publishing of such a set of data, the disease values attached to a certain individual can change over time.

A typical guarantee we want to achieve is that the probability that an adversary can derive for the linkage of a person to a sensitive value is no more than $1/\ell$. This is well-known to be a simple form of ℓ-diversity [5]. This guarantee sounds innocent enough for a single release data publication. However, when it comes to serial data publishing, the objective becomes quite elusive and requires a much closer look. In serial publishing, the set of individuals that are recorded in the data may change, and the sensitive values related to individuals may also change. We assume that the sensitive values can change freely.

Let us consider a sensitive disease chlamydia, which is a STD that is easily curable. Suppose that there exist 3 records of an individual o in 3 different medical data releases. It is obvious that typically o would not want anyone to deduce with high confidence from these released data that s/he has ever contracted chlamydia in the past. Here, the past practically corresponds to *one or more* of the three data releases. Therefore, if from these data releases, an adversary can deduce with high confidence that o has contracted chlamydia in one or more of the three releases, privacy would have been breached. To protect privacy, we would like the probability of any individual being linked to a sensitive value in one or more data releases to be bounded from the above by $1/\ell$. Let us call this privacy guarantee the *global guarantee* and the value $1/\ell$ the *privacy threshold*.

Id	Sex	Zip-code	Disease
o_1	M	65001	flu
o_2	M	65002	chlamydia
o_3	F	65014	flu
o_4	F	65015	fever

(a) T_1

Id	Sex	Zip-code	Disease
o_1	M	65001	chlamydia
o_2	M	65002	flu
o_3	F	65014	fever
o_5	F	65010	flu

(b) T_2

Fig. 1. A motivating example

Though the global guarantee requirement seems to be quite obvious, to the best of our knowledge, no existing work has considered such a guarantee. Instead, the closest guarantee of previous works is the following: for *each* of the data releases, o can be linked to chlamydia with a probability of no more than $1/\ell$. Let us call this guarantee the *localized guarantee*. Would this guarantee be equivalent to the above global guarantee ? In order to answer this question, let us look at an example.

Consider two raw medical tables (or micro data) T_1 and T_2 as shown in Figure 1 at time points 1 and 2, respectively. Suppose that they contain records for five individuals o_1, o_2, o_3, o_4, o_5. There are two kinds of attributes, namely *quasi-identifier (QID)* attributes and *sensitive* attributes. Quasi-identifier attributes are attributes that can be used to identify an individual with the help of an external source such as a voter registration list [7]. In this example, sex and zipcode are the quasi-identifier attributes, while disease is the sensitive attribute. Attribute id is used for illustration purpose and does not appear in the published table. We assume that each individual corresponds to at most one tuple in each table at each time point. Furthermore, we assume no additional background knowledge about the linkage of individuals to diseases [2], [1], and the sensitive values linked to individuals can be freely updated from one release to the next release.

Assume that the privacy threshold is $1/\ell = 1/2$. In a typical data anonymization [7], in order to protect individual

Sex	Zipcode	Disease
M	6500*	flu
M	6500*	chlamydia
F	6501*	flu
F	6501*	fever

(a) T_1^*

Sex	Zipcode	Disease
M	6500*	chlamydia
M	6500*	flu
F	6501*	fever
F	6501*	flu

(b) T_2^*

Fig. 2. Anonymization for T_1 and T_2

privacy, the QID attributes of the raw table are *generalized* or *bucketized* in order to form some *anonymized groups* ($\mathcal{AG}$) to hide the linkage between an individual and a sensitive value. For example, table T_1^* in Figure 2(a) is a *generalized* table of T_1 in Figure 1. We generalize the zip code of the first two tuples to 6500* so that they have the same QID values in T_1^*. We say that these two tuples form an *anonymized group*. It is easy to see that in each published table T_1^* or T_2^*, the probability of linking any individual to chlamydia or flu is at most 1/2, which satisfies the localized guarantee. The question is whether this satisfies the global privacy guarantee with a threshold of $1/2$.

For the sake of illustration, let us focus on the anonymized groups G_1 and G_2 containing the first two tuples in tables T_1^* and T_2^* in Figure 2, respectively. The probability in serial publishing can be derived by the possible world analysis. There are four possible *worlds* for G_1 and G_2 in these two published tables, as shown in Figure 3. Here each *possible world* is one possible way to assign the diseases to the individuals in such a way that is consistent with the published tables. Therefore, each possible world is a possible assignment of the sensitive values to the individuals at all the publication time points for groups G_1 and G_2. Note that an individual can be assigned to different values at different data releases, and the assignment in one data release is independent of the assignment in another release.

Consider individual o_2. Among the four possible worlds, three possible worlds link o_2 to "chlamydia", namely w_1, w_2 and w_3. In w_1 and w_2, the linkage occurs at T_1, and in w_3, the linkage occurs at T_2. Thus, the probability that o_2 is linked to "chlamydia" in at least one of the tables is equal to $3/4(= 0.75)$, which is greater than $1/2$, the intended privacy threshold. From this example, we can see that localized guarantee does not imply global guarantee.

In the full version [9] of this paper, we show that in order to ensure the global guarantee, the sizes of the anonymized groups typically need to be bigger than that needed for localized guarantee. In the above example, we can use size 4 anonymized groups as shown in Figure 4. There will be $4! \times 4!$ possible worlds. It is easy to see that $3/4$ of the possible worlds do not assign chlamydia to o_2 in the first release, $3/4$ of them do not assign chlamydia to o_2 in the second release, and $3/4 \times 3/4 = 9/16$ of the possible worlds do not assign chlamydia to o_2 in both releases. The remaining possible worlds assign chlamydia to o_2 in at least one of the two releases. Hence, the privacy breach probability = $1 - 9/16 = 7/16 < 1/2$. However, in [9], we show that the exact condition for privacy guarantee is not simply a size requirement, but a bound on a size ratio between that of the group and the sensitive value occurrences.

The contributions of this paper include the following: We point out the problem of privacy breach that arises with localized guarantee and propose to study the problem of global guarantee in privacy preserving serial data publishing. We formally analyze the privacy breach with transient sensitive values.

The rest of this paper is organized as follows. Section II surveys the previous related works. Section III contains our problem definition and show shows that the data satisfying the global guarantee also satisfies a privacy guarantee commonly adopted in the privacy literature called the *local guarantee*. Section IV describes a general formula for the breach probability. Section V concludes our work and points out some possible future directions.

In the full version [9] of our paper, useful properties related to the anonymization under the global guarantee are derived. These properties are related to the anonymized group sizes. Typically group sizes greater than that required for the localized guarantee will be needed to attain the global guarantee. These properties are then leveraged in the proposal of new anonymization strategies that can minimize the information loss. We have also conducted extensive experiments with a real medical dataset to verify our techniques. The results show that our methodology are very promising in real world applications. Details can be found in [9].

II. Related Work

Here, we summarize the previous works on the problem of privacy preserving serial data publishing. k-anonymity has been considered in [3] and [6] for serial publication allowing only insertions, but they do not consider the linkage probabilities to sensitive values. The work in [8] considers sequential releases for different attribute subsets for the same dataset, which is different from our definition of serial publishing.

There are some more related works that attempt to avoid the linkage of individuals to sensitive values. Delayed publishing is proposed in [2] to avoid problems of insertions, but deletion and updates are not considered. While [11] considers both insertions and deletions, both [2] and [11] make the assumption that when an individual appears in consecutive data releases, then the sensitive value for that individual is not changed. As pointed out in [1], this assumption is not realistic. Also the protection in [11] is record-based and not individual-based. This is quite problematic. As in our running examples, there are two *records* for one *individual* o_2, namely, t_1 in table T_1 and t_2 in table T_2 (note that T_1 and T_2 need not be consecutive releases, so that the sensitive value linked to o_2 can change even if we adopt the above unrealistic assumption in [2], [11]). If we consider just tuple t_1, then there are only 2 possible worlds where t_1 is linked to chlamydia in Figure 3, namely w_1 and w_2. If we just consider tuple t_2, there are also only 2 possible worlds linking it to chlamydia, namely w_1 and w_3. Hence, T_1^* and T_2^* satisfy the record-based requirement of [11] if the risk threshold is 0.5. In fact, these are possible tables generated by the mechanism proposed in [11]. However,

Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia

T_1

Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia

T_2

(a) Possible world 1 w_1

Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia

T_1

Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu

T_2

(b) Possible world 2 w_2

Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu

T_1

Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia

T_2

(c) Possible world 3 w_3

Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu

T_1

Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu

T_2

(d) Possible world 4 w_4

Fig. 3. Possible worlds for G_1 and G_2

Sex	Zipcode	Disease
M/F	650**	flu
M/F	650**	chlamydia
M/F	650**	flu
M/F	650**	fever

(a) T_1*

Sex	Zipcode	Disease
M/F	650**	chlamydia
M/F	650**	flu
M/F	650**	fever
M/F	650**	flu

(b) T_2*

Fig. 4. Anonymization for global guarantee

we have shown that this anonymization does not provide the expected protection for the individuals.

The ℓ-scarcity model is introduced in [1] to handle the situations when some data may be permanent so that once an individual is linked to such a value, the linkage will remain in subsequent releases whenever the individual appears (not limited to consecutive releases only). The major focus of [1] is the privacy protection for permanent sensitive values. However, for transient sensitive values, [1] and [11] adopt the following principle.

Principle 1 (Localized Guarantee): For each release of the data publication, the probability that an individual is linked to a sensitive value is bounded by a threshold.

However, we have seen in the example in the previous section that this cannot satisfy the expected privacy requirement. Hence, we consider the following principle.

Principle 2 (Global Guarantee): Over all the published releases, the probability that an individual has ever been linked to a sensitive value is bounded by a threshold.

III. Problem Definition

Suppose tables $T_1, T_2, ..., T_k$ are generated at time points, $1, 2, ..., k$, respectively. Each table T_i has two kinds of attributes, *quasi-identifier attributes* and *sensitive attributes*. For the sake of illustration, we consider one single sensitive attribute S containing $|S|$ values, namely $s_1, s_2, ..., s_{|S|}$. Assume that the sensitive values for individuals can freely change from one release to another release so that the linkage of an individual o to a sensitive value s in one data release has no effect on the linkage of o to any other sensitive value in any other data release. Assume at each time point j, a data publisher generates an anonymized version T_j^* of T_j for data publishing so that each record in T_j will belong to one anonymized group G in T_j^*. Given an anonymized group G, we define $G.S$ to be a multi-set containing all sensitive values in G, and $G.I$ to be the set of individuals that appear in G.

Definition 1 (Possible World): A series of tables $TS = \{T_1^p, T_2^p, ..., T_k^p\}$ is a possible world for published tables $\{T_1^*, T_2^*, ..., T_k^*\}$ if the following requirement is satisfied. For each $i \in [1, k]$,

1) there is a one-to-one correspondence between individuals in T_i^p and individuals in T_i^*
2) for each anonymized group G in T_i^*, the multi-set of the sensitive values of the corresponding individuals in T_i^p is equal to $G.S$.

Let $p(o, s, k)$ be the probability that an individual o is linked to s in at least one published table among published tables $T_1^*, T_2^*, ..., T_k^*$.

Let $t.S$ stand for the sensitive value of tuple t. We say that o is linked to s in a table T_i^p if for the tuple t of o in T_i^p, $t.S = s$. Following previous works, we define the probability based on the *possible worlds* as follows.

Definition 2 (Breach Probability): The breach probability is given by

$$p(o, s, k) = \frac{W_{link}(o,s,k)}{W_{total,k}} \quad (1)$$

where $W_{link}(o, s, k)$ is the total number of possible worlds where o is linked to s in at least one published table among $T_1^p, T_2^p, ..., T_k^p$ and $W_{total,k}$ is the total number of possible worlds for published tables $T_1^*, T_2^*, ..., T_k^*$.

For example, in our running example, for the tables shown in Figure 2, $p(o_2, s, 2)$ is equal to 3/4 where s is equal to "chlamydia". For the tables shown in Figure 4, $p(o_2, s, 2)$ is equal to 7/16. We will describe how A general formula to calculate $p(o, s, k)$ can be found in Section IV.

While privacy breach is the most important concern, the utility of the published data also needs to be preserved. There are different definitions of utility in the existing literature. Some commonly adopted utility measurements are described in Section II.

In this paper, we are studying the following problem.

Problem 1: Given a privacy parameter ℓ (a positive integer), a utility measurement, $k - 1$ published tables, namely $T_1^*, T_2^*, ..., T_{k-1}^*$ and one raw table T_k, we want to generate a published table T_k^* from T_k such that the utility is maximized, and for each individual o and each sensitive value s,

$$p(o, s, k) \leq 1/\ell$$

Note that the above problem definition follows Principle 2 for global guarantee as discussed in Section II. For example, for the tables shown in Figure 4, $p(o_2, s, 2)$ is equal to 7/16 ($\leq 1/2$), which satisfies the global guarantee.

A. Global versus Localized Guarantee

Here, we show that protecting individual privacy with Principle 2 (global guarantee) implies protecting individual privacy with Principle 1 (localized guarantee). Under Principle 1, let $q(o, s, j, k)$ be the probability that an individual o is linked to

a sensitive value s in the j-th table. Following the definition of probability adopted in most previous works [11], [1], we have

$$q(o, s, j, k) = \frac{L_{link}(o,s,j,k)}{W_{total,k}}$$

where $L_{link}(o, s, j, k)$ is the total number of possible worlds in which o is linked to s in the j-th table and $W_{total,k}$ is the total number of possible worlds for the k published tables.

In our running example, k=2 and from Figure 3, there are four possible worlds, $W_{total,k} = 4$. Consider published table T_1^*. There are two possible worlds where o_2 is linked to chlamydia (s), namely w_1 and w_2. Thus, $L_{link}(o_2, s, 1, k) = 2$ and $q(o_2, s, 1, k) = \frac{2}{4} = \frac{1}{2}$. Similarly, when $j = 2$, $q(o_2, s, 2, k) = \frac{1}{2}$.

In general, it is obvious that $W_{link}(o, s, k) \geq L_{link}(o, s, j, k)$ for any $j \in [1, k]$. We derive that

$$p(o, s, k) \geq q(o, s, j, k)$$

Hence we have the following lemma.

Lemma 1: If $p(o, s, k) \leq 1/\ell$ (under Principle 2), then for any $j \in [1, k]$, $q(o, s, j, k) \leq 1/\ell$ (under Principle 1).

Corollary 1: Principle 2 (global guarantee) is a strictly stronger requirement than Principle 1 (localized guarantee).

IV. Breach Probability Analysis

In this section, we give the general formula of the breach probability $p(o, s, k)$. Consider an anonymized group in T_j for individual o denoted by $\mathcal{AG}_j(o)$. Let n_j be the size $\mathcal{AG}_j(o)$. Let $n_{j,i}$ be the total number of tuples in $\mathcal{AG}_j(o)$ with sensitive value s_i for $i = 1, 2, ..., |S|$. Without loss of generality, we consider the privacy protection for an arbitrary sensitive value $s = s_1$. With the above notations, we have the following lemma for the general formula of $p(o, s, k)$.

Lemma 2 (Closed Form of $p(o, s_1, k)$):

$$p(o, s_1, k) = \frac{\prod_{j=1}^{k} n_j - \prod_{j=1}^{k} (n_j - n_{j,1})}{\prod_{j=1}^{k} n_j} \quad (2)$$

From Equation (1), $p(o, s_1, k)$ is defined with a conceptual terms with the total number of possible worlds. Lemma 2 gives a closed form of $p(o, s_1, k)$. Given the information of n_j (i.e., the size of the anonymized group in the j-th table) and $n_{j,1}$ (i.e., the number of tuples in the anonymized group with sensitive value s_1 in the j-th table), we can calculate $p(o, s_1, k)$ with its closed form directly.

Example 1 (Two-Table Illustration): Consider that we want to protect the linkage between an individual and a sensitive value s_1. Suppose o appears in both published tables T_1^* and T_2^*. Let $\mathcal{AG}_1(o)$ and $\mathcal{AG}_2(o)$ be the anonymized groups in T_1^* and T_2^* containing o. Suppose both $\mathcal{AG}_1(o)$ and $\mathcal{AG}_2(o)$ are linked to s_1.

By the notation adopted in this paper, n_k is the size of $\mathcal{AG}_k(o)$ and $n_{k,1}$ is the total number of tuples in $\mathcal{AG}_k(o)$ with sensitive value s_1.

By Lemma 2, we have $p(o, s_1, k) = \frac{n_1 n_2 - (n_1 - n_{1,1})(n_2 - n_{2,1})}{n_1 n_2} = \frac{n_{2,1} n_1 + n_{1,1} n_2 - n_{1,1} n_{2,1}}{n_1 n_2}$.

□

Example 2 (Running Example): In our running example as shown in Figure 2, consider the second individual o_2 and a sensitive value "chlamydia". We know that $n_1 = n_2 = 2$. Suppose s_1 is "chlamydia". Thus, $n_{1,1} = n_{2,1} = 1$. With respect to the published tables as shown in Figure 2, according to the formula derived in Example 1,

$$p(o_2, s_1, 2) = \frac{1 \times 2 + 1 \times 2 - 1 \times 1}{2 \times 2} = \frac{3}{4}$$

which is greater than $1/2$ (the desired threshold).

However, if we publish tables as shown in Figure 4, then $n_1 = n_2 = 4$ and $n_{1,1} = n_{2,1} = 1$.

$$p(o_2, s_1, 2) = \frac{1 \times 4 + 1 \times 4 - 1 \times 1}{4 \times 4} = \frac{7}{16}$$

which is smaller than $1/2$.

In this example, we observe that, since the published tables as shown in Figure 4 have a larger anonymized group size (compared with the published tables as shown in Figure 2), $p(o_2, s_1, 2)$ is smaller. □

We aim to publish table T_k^* like Figure 4 at each time point k such that $p(o, s, k) \leq 1/\ell$ for each individual o and each sensitive value s. How to anonymize the table can be found in [9].

V. Conclusion

In this paper, we propose a new criterion of *global guarantee* for privacy preserving data publishing. This guarantee corresponds to a basic requirement of individual privacy where the probability of linking an individual to a sensitive value in one or more data releases is bounded. We show that global guarantee is a stronger privacy requirement than localized guarantee which has been adopted in previous works.

Acknowledgements: The research of Raymond Chi-Wing Wong is supported by HKRGC GRF 621309 and Direct Allocation Grant DAG08/09.EG01.

References

[1] Y. Bu, A. W.-C. Fu, R. C.-W. Wong, L. Chen, and J. Li. Privacy preserving serial data publishing by role composition. In *VLDB*, 2008.

[2] J. Byun, Y. Sohn, E. Bertino, and N. Li. Secure anonymization for incremental datasets. In *Secure Data Management*, pages 48–63, 2006.

[3] B. C. M. Fung, K. Wang, A. Fu, and J. Pei. Anonymity for continuous data publishing. In *EDBT*, 2008.

[4] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.

[5] A. Machanavajjhala, J. Gehrke, and D. Kifer. l-diversity: privacy beyond k-anonymity. In *ICDE*, 2006.

[6] J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang. Maintaining k-anonymity against incremental updates. In *SSDBM*, 2007.

[7] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowldege based systems*, 10(5), 2002.

[8] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *SIGKDD*, 2006.

[9] R. C.-W. Wong, A. W.-C. Fu, J. Liu, K. Wang, and Y. Xu. Global privacy guarantee in serial data publishing. In *http://www.cse.ust.hk/∼raywong/paper/transPrivacy-technical.pdf*, 2009.

[10] R.C.W. Wong, J. Li, A. Fu, and K. Wang. (alpha, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In *KDD*, 2006.

[11] X. Xiao and Y. Tao. m-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, 2007.