A Unified RFID Data Anonymization Platform

Rui Chen
Concordia University
Montreal, QC
Canada H3G 1M8
Email: ru_che@cse.concordia.ca

Benjamin C. M. Fung
Concordia University
Montreal, QC
Canada H3G 1M8
Email: fung@ciise.concordia.ca

***Abstract*—RFID technology has been widely applied to various domains. However, the privacy concern embedded in RFID data becomes a major obstacle of its further application. In this paper, we report our recent achievement in developing a practical RFID data anonymization platform that supports both global and local suppressions to prevent RFID data from privacy linkage attacks without compromising the support of high-quality data analysis tasks. We also introduce our ongoing efforts to enhance the RFID data anonymization platform.**

I. Introduction

With the wide application of Radio Frequency IDentification (RFID) technology, RFID data, in form of $location$-$timestamp$ pair, have been collected in various sectors, such as public transit systems, hospitals, and supply chains. RFID data directly unveil the movement of a tag owner and, therefore, his identity. Moreover, since RFID data are often collected and stored with other sensitive information, for example, one's disease in the context of hospitals, RFID data could also be used to infer other sensitive information of a tag owner. Such privacy concerns have been becoming a major obstacle of sharing RFID data among different parties.

In spite of its importance, anonymizing RFID data for publishing has not been well-studied. The large number of approaches proposed for anonymizing relational data [4][5], unfortunately, do not apply to RFID data due to its high dimensionality, sparseness, and sequentiality. In this paper, we present a practical RFID data anonymization platform that is able to anonymize RFID data for different data mining tasks under a strong privacy model. In addition, we introduce our ongoing research work for enhancing the platform in terms of both anonymization mechanism and privacy model.

II. LKC-Privacy Model

More and more RFID data analysis tasks require publishing RFID data together with other personal information of data owners. In a typical RFID data table, a tuple contains all information of a data owner, including a *path* composed of all his location-timestamp pairs sorted by timestamps. Based on a RFID data table, an adversary can conduct two major types of attacks, namely identity linkage attack and attribute linkage attack. In an *identity linkage attack*, an adversary uses his background knowledge, a bounded number of location-timestamp pairs, to uniquely identify a data owner in the table. In an *attribute linkage attack*, an adversary infers an owner's sensitive information with relatively high confidence based on background knowledge. We proposed a privacy model called LKC*-privacy* [3] to thwart these privacy attacks.

Definition 2.1 (LKC-privacy): Let L be the maximum length of the background knowledge. Let S be a set of sensitive values. A RFID data table T satisfies LKC-privacy *if and only if for any sequence q in T with $0 < |q| \leq L$,*

1) *$|T(q)| \geq K$, where K is a positive integer specifying the anonymity threshold, where $T(q)$ is the group of records containing q,* and
2) *$Conf(s|T(q)) \leq C$ for any $s \in S$, where $0 \leq C \leq 1$ is a real number specifying the confidence threshold.* ■

III. Anonymization

Our anonymization platform aims at removing all "violations" from a RFID data table T, where a violation is a subsequence of a path in T that violates a given LKC-privacy requirement. A RFID data table usually contains a large number of such violations, and it is infeasible to directly enumerate all of them. Instead, the concept of *critical violation* was proposed in [3]. It is sufficient to eliminate all privacy threats by removing all critical violations in a RFID data table.

Definition 3.1 (Critical violation): A violation q is a *critical violation* if every proper subsequence of q is a non-violation. ■

The set of critical violations, denoted by $V(T)$, can be efficiently generated by recursively creating size-$(l+1)$ critical violations by pruning and self-joining size-l non-critical violations [3]. To eliminate all identified critical violations, our platform provides both global suppression and local suppression in order to accommodate different data analysis tasks. Global suppression eliminates all instances of a location-timestamp pair if it is selected to be suppressed, while local suppression allows some instances to remain intact. In general, global suppression requires less computational resources and guarantees data truthfulness, which is important if the data will be examined by human users for the purpose of auditing, data interpretation, or visual data mining; local suppression, in general, achieves better data utility.

Given a RFID data table T and a LKC-privacy requirement, it is NP-hard to find the optimal anonymization solution. Therefore, we propose greedy algorithms based on global and local suppressions to eliminate all identified critical violations in order to efficiently identify a reasonably good solution.

Generally, suppressing a pair p in $V(T)$ increases privacy and decreases data utility. To find the sub-optimal trade-off between privacy and utility, we define a greedy function, $Score(p)$, as follow:

$$Score(p) = \frac{PrivGain(p)}{InfoLoss(p)},$$

where $PrivGain(p)$ is the number of critical violations eliminated by suppressing p, and $InfoLoss(p)$ is the utility loss measured by a user's utility requirement, for example, the number of instances lost due to suppressing p. This simple design allows a user to easily incorporate his utility metric into our platform.

Global suppression. In each iteration of global suppression, the platform chooses the pair with the highest score, suppresses all its instances from the RFID data table T, and updates $V(T)$. The anonymization process ends when no pair is left in $V(T)$. Refer to [3] for a detailed discussion and experimental results on this approach.

Local suppression. Employing local suppression can significantly improve the resulting data utility; however, designing an efficient local suppression scheme raises new challenges. One nice property of global suppression is that the size of $V(T)$ monotonically decreases with respect to a global suppression. This property guarantees that the anonymization process takes at most $|V(T)|$ iterations to generate an anonymous output. However, local suppression does not share this property. A key to an efficient local suppression is to ensure that no *new* critical violation will be generated in the anonymizing process. A local suppression is termed a *valid* local suppression if it does not generate any new critical violation. In [1], we propose a novel approach, which does not calculate the values of new critical violations, if any, but is sufficient to foresee if any new critical violation will be generated. The approach significantly narrows down ! the search space to a very small set of pairs that may be affected by a suppression by carefully exploring the properties of critical violation.

Unlike global suppression, local suppression distinguishes different instances of the same pair because some instances can remain intact after a suppression. In each iteration of local suppression, we quickly identify the *instances* of pairs in $V(T)$ that can be eliminated with a valid local suppression. The instance with the highest score will be selected and suppressed in T. When no valid local suppression can be found, global suppression is used. In the following iteration, the algorithm seeks for new valid local suppressions.

The proposed global and local suppression methods are applied to anonymize various RFID data sets. The experimental results demonstrate that the methods can efficiently process large RFID data sets with desirable resulting utility.

IV. Ongoing Enhancement

There are several directions we are working on to enhance our RFID data anonymization platform. A reason of not employing *generalization*, another well-established anonymization mechanism, is that it is difficult to find a logical taxonomy tree for locations. Yet, it may be possible to find a well-defined taxonomy tree for timestamps. The possibility naturally stimulates the attempt of incorporating generalization into our platform. Generalization could be used as a standalone mechanism to achieve LKC-privacy and also be utilized in combination with suppression in order to obtain even better data utility.

Despite the fact that our platform has adopted a stronger privacy notion for RFID data than other existing works by taking into consideration the possibility of inferring data owners' sensitive information via RFID data, the specificity of RFID data enables adversaries to perform other kinds of privacy attacks with different extents of background knowledge. Recently, Dwork proposed a new privacy model, *differential privacy* [2], which is resistant to an adversary with arbitrary background knowledge and arbitrary computation power. We deem that differential privacy be a more stringent privacy model for RFID data anonymization. It requires that the outcome of any data analysis be insensitive to a single change in a data table. Differential privacy cannot be achieved by either suppression or generalization, but by randomization such as noise addition. Therefore, we are in the progress of developing new anonymization mechanisms to support differential privacy in our platform.

V. Conclusion

Anonymizing RFID data is a challenging task due to RFID data's high dimensionality, sparseness, and sequentiality. In the paper, we present a platform for anonymizing RFID data that supports both global and local suppressions under the LKC-privacy model while accommodating different data mining tasks. Preliminary experiments over various RFID data sets suggest promising performance.

VI. Acknowledgements

The research is supported in part by the Discovery Grants (356065-2008) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, under 2nd revision.

[2] C. Dwork. Differential privacy. In *International Colloquium on Automata, Languages and Programming (ICALP)*, 2006.

[3] B. C. M. Fung, K. Al-Hussaeni, and M. Cao. Preserving rfid data privacy. In *Proc. of the 2009 IEEE International Conference on RFID*, Orlando, FL, April 2009.

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

[5] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proc. of the 17th ACM PODS*, page 188, June 1998.