International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 02, February 2017
ISSN: 2455-3778
http://www.ijmtst.com

Auxiliary Information in Data Privacy and Attacks for Secure Multiparty Protocols

PON.Arivanantham[1] | Dr.M.Rama Krishnan[2]

[1]Research Scholar, Satyabama University, Chennai, India.
[2]Professor, Madurai Kamaraj, University, Madurai, India.

To Cite this Article
PON.Arivanantham and Dr.M.Rama Krishnan, "Auxiliary Information in Data Privacy and Attacks for Secure Multiparty Protocols", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 02, 2017, pp. 48-51.

ABSTRACT

Secure multiparty protocols have been proposed to enable non colluding parties to cooperate without a trusted server. Even though such protocols prevent information disclosure other than the objective function, they are quite costly in computation and communication. The high overhead motivates parties to estimate the utility that can be achieved as a result of the protocol beforehand. In this paper, we propose a look-ahead approach, specifically for secure multiparty protocols to achieve distributed k-anonymity, which helps parties to decide if the utility benefit from the protocol is within an acceptable range before initiating the protocol. The look-ahead operation is highly localized and its accuracy depends on the amount of information the parties are willing to share. Experimental results show the effectiveness of the proposed methods.

KEYWORDS: *multiparty protocol, SMC Protocol, Probabilistic model, look ahead approach, candidate Generation Algorithm.*

I. INTRODUCTION

SECURE multiparty computation (SMC) protocols are one of the first techniques used in privacy preserving data mining in distributed environments. The idea behind these protocols is based on theoretical proof that two or more parties, both having their own private data, can collaborate to calculate any function on the union of their data [8]. While doing so, the protocol does not reveal anything other than the output of the function or anything that can be computed from it in polynomial time. More-over, the protocol does not require a trusted third party. While these properties are Promising for privacy preserving applications, SMC may be prohibitively expensive. In fact, many SMC protocols for privacy preserving data mining suffer from high computation and communication costs. Furthermore, those that are closest to be practical are designed for the semi honest model, which assumes that parties will not deviate from the protocol. Theoretically, it is possible to convert protocols in the semi honest model into protocols in the malicious model. However, the resulting protocols are even more costly. To the best of our knowledge, this is the first work that looks ahead of an SMC protocol and gives an estimate for We state that an ideal look ahead satisfies the following:

1. The methodology is highly localized in computation, it is fast and requires little communication cost (at least asymptotically better than the SMC protocol).
2. The methodology relies on non sensitive data, or better, data that would be implied from the output of the objective function.

II. RELATED WORK

In this section, we outline a number of characteristics we consider crucial to the design of

a practical privacy criterion. At the same time, we review the literature, indicating how previous work does not match our desired characteristics. From our perspective, a practical privacy criterion should display the following characteristics:

1. Intuitive: The data owner (usually not a computer scientist) should be able to understand the privacy criterion in order to set the appropriate parameters.
2. Efficiently checkable: Whether a release candidate satisfies the privacy criterion should be efficiently checkable.
3. Flexible: In data publishing, the data owner often considered tradeoff between disclosure risk and data utility. A practical privacy criterion should provide this flexibility.
4. External knowledge: The privacy criterion should guarantee safety in the presence of common types of external knowledge.
5. Value-centric: Often, different sensitive values have different degrees of sensitivity (e.g., AIDS is more sensitive than flu). Thus, a practical privacy criterion should have the flexibility to provide different levels of protection for different sensitive values, not just uniform protection for all the values in the sensitive attribute. We call the latter attribute-centric. An attribute-centric criterion tends to over-protect the data. For example, to protect individuals having AIDS, the data owner must set the strongest level of protection, which is not necessary for individuals having flu. Instead, we take the more flexible value-centric approach.
6. Set-valued sensitive attributes: In many real-world scenarios, an individual may have several sensitive values, e.g., diseases. No existing privacy criterion fully satisfies our desiderata. The most closely-related work is that of Martin et al. While groundbreaking in the treatment of external knowledge, the approach has several important shortcomings:

• The knowledge quantification is not intuitive. It is hard to understand the practical meaning of k-implications.

• Martin et al. showed that their language can express any logic-based expression of external knowledge, when the number k of basic implications is unbounded. However, their language cannot practically express some important types of knowledge, e.g., simply Flu ∈ Bob[S] (a very common kind of knowledge that the adversary may obtain from a similar dataset). Expressing such knowledge in their language requires (|S|−1) basic implications, where |S| is the number of sensitive values. However, with this number of basic implications, no release candidate can possibly be safe. Thus, Flu ∈Bob[S] will never be used in their criterion.

• The privacy criterion is attribute-centric, and there is no straightforward extension of the proposed algorithm to the more flexible value-centric case. The reason is that the algorithm can only compute max {Pr(s ∈t[S] | K, D*)} for the sensitive value that is most frequent in at least one QI-group. However, the sensitive values that need the most protection (e.g., AIDS) are usually infrequent ones.

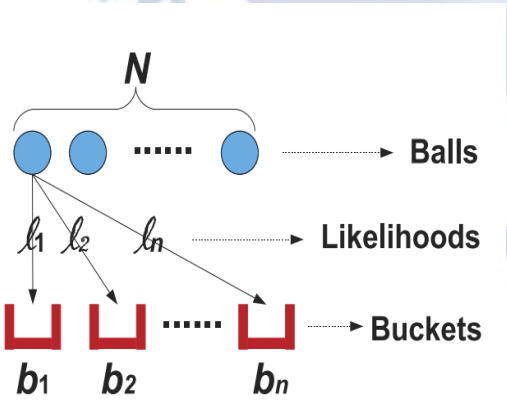
• Each individual is assumed to have only one sensitive value. Our work builds upon and addresses the above issues. Note that our language can express some knowledge (e.g., Flu∈ Bob[S]) that cannot be practically expressed in their language, and vice versa.

III. Discussion

The earlier section demonstrated the viability of our approach using an example with eight potentially identifying attributes. In general, the size of the solution space depends on the number of such attributes and the granularity at which they need to be considered. Determining which attributes should be considered as potentially identifying is based on an assessment of possible links to other avail-able data. This needs to be done with typical databases in each domain (e.g., retail). Clearly, as the number of potentially identifying attributes grows, identity disclosure risk in-creases. The corresponding increase in the number of unique combinations of potentially identifying values will have an impact on the k-anonymity approach. Also, the complex-it y of the optimization problem increases due to the larger solution space to be searched. Further experiments are needed to investigate the applicability of this approach to wider data sets. In each domain, in addition to the identifying attributes one needs to determine the sensitive attributes. It has been suggested that sensitive attributes be removed completely from data sets being publicly released [19]. Further work is needed to determine adequate ways of handling these at-tributes if they are included. Clearly, they cannot be targets of predictive modeling using our methods since that will result in their inferential disclosure. This is because the optimization we perform for predictive modeling would group together rows with similar values for the target attribute. This optimization improves the model

accuracy while satisfying the identity disclosure constraint, but it also increases the inferential attribute disclosure for the sensitive attribute being targeted. While this is an explicit issue with the k-anonymity approach to anonymization, further investigation is needed on issues related to the inferential disclosure of sensitive attributes even for other approaches (e.g., additive noise and swapping). In many cases only a sample of the data is released. The privacy protection due to sampling has been considered in various works (e.g., [6, 16, 3]). Applying the k-anonymity approach to the release of a sample opens up some new issues. One approach could be to require that the released sample satisfy the k-anonymity requirement. The choice of k would have to be made taking into account the sampling etc. Alternatively, the k-anonymity requirement could be rest applied to the entire population before a sample of the transformed table is released. The sizes of the groups in the released sample will depend on the form of sampling used (e.g., random, startled). Further work is needed to explore the k-anonymity approach in the context of sampling. For predictive modeling usage the metrics denned in consider predictability using only the potentially identifying attributes. This was done independent of the predictive capabilities of the other non-identifying attributes. Considering both identifying and non-identifying attributes during the data transformation process could lead to better solutions. Finding an effective way of doing this with potentially large numbers of non-identifying attributes needs further exploration.

IV. Probabilistic Model



A fast algorithm for distributed association rule mining is given in Cheung et. al. [2]. Their procedure for fast distributed mining of association rules (FDM) is summarized below.

1) Candidate Sets Generation: Generate candidate sets CGi(k) based on GLi(k−1) , item sets that are supported by the Si at the (k-1)th iteration, using the classic a-priori candidate generation algorithm. Each site generates candidates based on the intersection of globally large (k-1) item sets and locally large (k-1) item sets.

2) Local Pruning: For each X ∈ CGi(k) , scan the database DBi at Si to compute X.supi. If X is locally large Si , it is included in the LL i(k) set. It is clear that if X is supported globally, it will be supported in one site.

3) Support Count Exchange: LL i(k) are broadcast, and each site computes the local support for the items in ∪i LL i(k) .

4) Broadcast Mining Results: Each site broadcasts the local support for item sets in ∪i LL i(k) . From this, each site is able to compute L (k) .

V. Conclusion

Most SMC protocols are expensive in both communication and computation. We introduced a look-ahead approach for SMC protocols that helps involved parties to decide whether the protocol will meet the expectations before initiating it. We presented a look-ahead protocol specifically for the distributed k-anonymity by approximating the probability that the output of the SMC will be more utilized than their local anonymizations. Experiments on real data showed the effectiveness of the approach. Designing look ahead for other SMC protocols stands as a future work. A wide variety of SMC protocols have been proposed especially for privacy preserving data mining applications [12], [17], [28] each requiring a unique look-ahead approach.

As for the look-ahead process on distributed anonymization protocols, definitions of k-anonymity definitions can be revisited, more efficient techniques can be developed and experimentally evaluated.

References

[1] R.J. Bayardo and R. Agrawal, “Data Privacy through Optimal K-Anonymization, ”Proc. 21st Int’l Conf. Data Eng. (ICDE ’05), pp. 217-228, 2005.

[2] C. Blake and C.J. Merz, “UCI Repository of Machine Learning Databases,” http://www.ics.uci.edu/mlearn/ MLRepository.html, Univ. of California, Irvine, Dept. of Information and Computer Sciences, 2012.

[3] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, “Privacy Skyline: Privacy with Multidimensional

Adversarial Knowledge,"Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07),pp. 770-781, 2007.

[4] J. Domingo-Ferrer and V. Torra, "Ordinal, Continuous and Heterogeneous K-Anonymity through Micro aggregation, "Data Mining and Knowledge Discovery, vol. 11, no. 2, pp. 195-212, 2005.

[5] W. Feller,An Introduction to Probability Theory and Its Applications, vol. 1, Wiley, 1968.

[6] S.R. Ganta, S.P. Kasiviswanathan, and A. Smith, "Composition Attacks and Auxiliary Information in Data Privacy, "Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 265-273, http://doi.acm.org/10.1145/1401890. 1401926, 2008.

[7] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast Data Anonymization with Low Information Loss, "Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07),pp. 758-769, 2007.

[8] O. Goldreich, The Foundations of Cryptography, vol. 2, Cambridge Univ. Press, http://www.wisdom.weizmann.ac.il/ ~oded/ PSBookFrag/enc.ps, 2004.