ISSN : 2231-0711

Supriya S. Borhade et al | IJCSET(www.ijcset.net) | June 2015 | Vol 5, Issue 6,182-184

Defining Privacy for Data Mining- An Overview

Ms. Supriya S. Borhade[1], Dr. S. V. Gumaste[2]

[1]*M.E. Student, Computer Engineering Department, SPCOE , Dumberwadi, Pune Maharashtra, India*

[2] *Professor & Head, Computer Engineering Department, SPCOE , Dumberwadi, Pune Maharashtra, India*

Abstract: PPDM (Privacy preserving data mining) in receipt of valid data mining results without learning the original or essential data values. It has been receiving attention in the research scoiety and beyond. That time it was unclear what privacy preserving means. This paper provides a study framework and metrics for discussing the meaning of privacy preserving data mining, as groundwork for extra research in this field.

Keywords: Privacy, Security, Obscuring, SMC, PPDM

I. INTRODUCTION

The enormous amount of data available means that it is possible to learn a lot of information about individuals from public data like Purchasing patterns , Family history , Medical data and much more.

There has recently been a rise in interest in privacy preserving data mining [2][3][6][11][12]. Even accepted in press has picked up on this trend. Though after, the concept of what is meant by privacy was not clear. In this paper outline some previous work of the concepts that are addressed.

In this type of research and provide a path for defining and understanding privacy control or constraints. Generally when people talk of privacy, they say "information about me which feel as personal, confidential or private should not be unnecessarily distributed or publicly known, must be kept secure. This definition doesn't match the dictionary definition (Webster's), "freedom from unauthorized intrusion". It is this interruption or use of personal data in a way that harmfully impacts someone's life that causes concern. Providing that data is not misused, nearly all people do not feel their privacy has been violated. The problem is that once information is revealed, it may be impractical to prevent misuse. Utilizing this distinction – ensuring that a data mining project won't enable misuse of personal information – unlock opportunities that "complete privacy" would prevent. To do this, require technical and social solutions that ensure data will not be revealed. The same basic concerns also apply to collections of data. Given a collection of data, it is unlikely to learn things that are not revealed by any individual identity data item. An individual may not be care about someone perceiving or sharing their birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft. Such kind of privacy problem arises with large, multi-individual collections as well. A technique that guarantees no individual data or identity is revealed may still release information describe the collection as a whole. Such "commercial or corporate information" is generally the goal of data mining, but some results may still lead to concerns. The difference between such business privacy issues and individual privacy is not that important. If we view revelation of knowledge about an entity (information about an individual identity) as a likely individual privacy breach, then generalizing this to disclosure of information about a subset of the data captures both views.

First, let's study background on the two main classes of privacy preserving data mining.

II. APPROACHES TO PRIVACY PRESERVING DATA MINING

There are two papers entitled Privacy Preserving Data Mining appeared in 2000. Both paper addresses a similar problem, constructing decision trees from private training data, the concepts of privacy were relatively different. One paper was based on data obscuration, i.e., modifying the data values so real values are not disclosed by [3][2]. The other paper used Secure Multiparty Computation (SMC) to "encrypt" data values by [14], It ensures that no party learns anything about another's data values. In this paper Section A will describe Secure Multiparty Computation(SMC) and section B will give additional background on data obscuration.

A. **Secure multiparty computation** (SMC)

The thought of Secure Multiparty Computation (SMC) by [13] [1] is that the parties involved learn nothing but the results. Informally, imagine one may have a trusted third party to which all parties give their input or data. The trusted party manipulates the output and returns it to the parties. SMC facilitates this without the trusted third party. There may be considerable communication between the parties to get the final result, but the parties don't learn anything from this communication. The manipulation is secure if given just one party's input and output from those runs, One can imitate what would be seen by the party. In this case, to imitate means that the distribution of what is actually seen and the distribution of the imitated view over many runs are computationally indistinguishable. One may not be able to exactly imitate every run, but over time one cannot tell the imitation from the real runs. Since one could imitate the runs from knowing only our input and output, it makes sense to say that one can't learn anything from the run other than the output. This look like a strong guarantee of privacy, and has been used in privacy preserving data mining work [14][13]. One has to be careful when using Secure Multiparty Computation to define privacy. For example, suppose anyone use a SMC technique to build a decision tree from databases at two sites [14], categorizing people into high and low risk for a sensitive disease.

Assume that the non-sensitive data is public, but the sensitive data (needed as training data to build the classifier) cannot be revealed. The SMC computation won't reveal the sensitive data, but the resulting classifier will enable all parties to guess the value of the sensitive data. It isn't that the SMC was "broken", but that the result itself violates privacy.

B. Obscuring data

A different approach to privacy is to obscure data, making private data obtainable, but with adequate noise added that exact values cannot be determined. One approach, typically used in census data, is to collective items. Knowing the average earnings for a neighborhood is not enough to determine the actual income of a resident of that neighborhood. An alternative is to add random noise to data values, then mine the distorted data. While this lesser the accuracy of data mining results, research has shown that the loss of accuracy can be small relative to the loss of ability to estimate an individual item. Anyone can reconstruct the original distribution of a collection of obscured numeric values, enabling improved construction of decision trees [3]. This would enable data collected from a web survey to be obscured at the source – the correct values would never fade away the respondent's machine – ensuring that accurate data doesn't exist. Techniques have also been developed for association rules, enabling valid rules to be learned from data where items have been randomly added to or removed from individual transactions [12].

III. PERFECT PRIVACY

One difficulty with the above is the tradeoff between privacy and accuracy of the data mining results. Secure Multiparty Computation (SMC) does superior, but at a high computational and communication price. In the "web survey" example, the respondents could connect in a secure multiparty computation to obtain the outcome, and reveal no information that is not contained in the results. However, getting thousands of respondents to participate synchronously in a complex protocol is impractical. While useful in the corporate model, it is not appropriate for the web model. Here presented a solution based on somewhat trusted third parties – the parties are not trusted with exact data, but trusted only not to collude with the "data receiver".

Assume the existence of *k* un-trusted, non-colluding sites.

1. Un-trusted signifies that none of these sites should be able to gain any useful information from any of the inputs of the local sites.
2. Non-colluding signifies that none of these sites should collude with any other sites to acquire information beyond the protocol.

Then, all of the local parties can split their local inputs into *k* random shares which are then split across the *k* un-trusted sites. Each of these random shares are meaningless information by themselves. However, if any of the parties combined their data, they would obtain some meaningful information from the combined data. Because this, it requires that the sites be non-colluding. It is believed that this assumption is not unrealistic. Each site combines the shares of the data it has received using a secure protocol to get the required data mining result.

The following is a to the point description of this approach. Every party is assumed to have a single bit of information x_i, identified by some key i. Each party locally generates a random number r_i and then sends $(i, \bar{x}_i = x_i \oplus r_i)$ to one site and (i, r_i) to the second site. Note that neither site will be able to predict the x_i, . Due to the xor operation $\oplus$, the input they see is indistinguishable from any uniformly generated random sequence. Given any data mining task(f) defined on $X = [x_1, x_2, \ldots, x_n]$, it suffices to evaluate $f(\bar{X}^{-} \oplus \bar{R}) = f(X)$ since R = $[r_1, r_2, \ldots, r_n]$ and $\bar{X}^{-} \oplus \bar{R} = [\bar{x}_1 \oplus r_1, \bar{x}_2 \oplus r_2, \ldots, \bar{x}_n \oplus r_n]$. It is a known fact that with the assumption of existence of trapdoor permutations (RSA is assumed to be a trapdoor permutation), any functionality $g, (g : \{0,1\}* \times \{0,1\}* 7 \rightarrow \{0,1\}* \times \{0,1\}*)$ can be evaluated privately in the semi-honest model [1]. Since the initial xor operation can be easily represented as a circuit, given functionality f, one can define a functionality $g(X, R) = f(\bar{X}^{-} \oplus \bar{R})$. Thus, any data mining functionality can be evaluated privately without revealing any information other than the final result. (For a more complete action, see [11].) While this solution is not principally efficient, indeed not even necessarily very practical for large quantities of data, it does demonstrate a method of maintaining perfect privacy while computing the required data mining function.

IV. LIMITATIONS ON RESULTS

How can someone constrain the results of data mining? There has been work in this area, addressing specific problems such as hiding specific association rules [6][7] or limiting confidence in any data mining [6]. While these provide some specific techniques, the means available to constrain results are limited. What is needed is a general way to specify, what is and is not allowed.

One probable approach is constraint-based data mining. This kind of research study is concerned with improving the efficiency of algorithms and understandability of results through providing up-front constraints on what results would be of attention. Would the languages used to describe these constraints also serve to define what results are acceptable from a privacy standpoint? While the current approaches do not enforce that nothing outside the constraints can be learned, they could offer a starting point for further research study.

V. CONCLUSIONS

Privacy preserving data mining (PPDM) has the potential to raise the reach and benefits of data mining technology. However, someone must be able to justify that privacy is preserved. For this, one needs to be able to communicate what we mean by "privacy preserving". The current mixture of definitions, with each paper having its own definition of what privacy is maintained, will lead to confusion among potential adopters of the technology. Here, presenting some suggestions for defining, measuring, and evaluating privacy preservation. Showed how these relate to both privacy policy and practice in the wider group of people, and to techniques in privacy preserving data

mining. The key point to remember is that privacy preserving data mining is possible. Technology has been, and is being, developed to consent to data mining without disclosing private information or sensitive data. There are legal and historical definitions of privacy that can be used to justify that this technology does preserve privacy. This is by no means the definitive word on the subject. While some measures, such as the differential entropy metric of Agrawal [2], have clear mathematical foundations and applications, others have strong potential for further development. Accepting a common framework for discussion of privacy preservation will enable next generation data mining technology to make significant advances in alleviating privacy concerns.

REFERENCES

[1] Goldreich, O.; Micali, S.; and Wigderson, A. 1987. How to play any mental game - a completeness theorem for protocols with honest majority. In *19th ACM Symposium on the Theory of Computing*, 218.229.

[2] Agrawal, D., and Aggarwal, C. C. 2001. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 247.255. Santa Barbara, California, USA: ACM.

[3] Agrawal, R., and Srikant, R. 2000. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, 439.450. Dallas, TX: ACM.

[4] Atallah, M.; Bertino, E.; Elmagarmid, A.; Ibrahim, M.; and Verykios, V. 1999. disclosure limitation of sensitive rules. In *Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, 25.32.

[5] Clifton, C., and Estivill-Castro, V., eds. 2002. *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14. Maebashi City, Japan: Australian Computer Society.

[6] Clifton, C. 2000. Using sample size to limit exposure to data mining. *Journal of Computer Security* 8(4):281.307.

[7] Delugach, H. S., and Hinke, T. H. 1996.Wizard: A database inference analysis and detection system. *IEEE Transactions on Knowledge and Data Engineering* 8(1).

[8] Du, W., and Atallah, M. J. 2001a. Privacy-preserving cooperative scientific computations. In *14th IEEE Computer Security Foundations Workshop*, 273.282.

[9] Du, W., and Atallah, M. J. 2001b. Privacy-preserving statistical analysis. In *Proceeding of the 17th Annual Computer Security Applications Conference*.

[10] Kantarcioglu, M., and Clifton, C. 2002. Privacy preserving distributed mining of association rules on horizontally partitioned data. In *The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02)*, 24.31.

[11] Kantarcioglu, M., and Vaidya, J. 2002. An architecture for privacy-preserving mining of client information. In Clifton, C., and Estivill-Castro, V., eds., *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14, 37.42. Maebashi City, Japan: Australian Computer Society.

[12] Rizvi, S. J., and Haritsa, J. R. 2002. Maintaining data privacy in association rule mining. In *Proceedings of 28th International Conference on Very Large Data Bases*, 682.693. VLDB.

[13] Yao, A. C. 1986. How to generate and exchange secrets. In *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*, 162.167. IEEE.

[14] Lindell, Y., and Pinkas, B. 2000. Privacy preserving data mining. In *Advances in Cryptology . CRYPTO 2000*, 36.54. Springer-Verlag.