Privacy Preserving Data Mining: An Analysis of Current Methods and Research

Peter Clark, Gus Anthon, Ahmet Oğuz Ozturk

Sound and Music Computing
peter.clark01@estudiant.upf.edu
gustav.anthon01@estudiant.upf.edu
ahmetoguz.ozturk01@estudiant.upf.edu

Abstract. This paper examines the state of privacy preservation techniques involved in data mining. A literature review of these techniques is included, in which the most noteworthy and effective methods are described. Further, an analysis of the efficacy of such techniques is included, in addition to a description of their respective advantages and disadvantages. Lastly, a summary of current methodologies of privacy preserving data mining is presented, along with guidelines and future possibilities to improve these methods.

Keywords: Data Privacy, Data Mining, PPDM, Anonymization-Based, Randomization-Based.

1 Introduction

With the development of the Internet, the amount of data produced by humanity is increasing every year. According to IBM research, 90% of all data obtained by the entire human civilization is generated in the last two years. [1] It has become important to make plans and forecasts for the future using this data. Extraction of the necessary data in a large data heap and storing them in a structured way can be explained as data mining. In other words data mining can be explained as the procedure to generate structured big data.

Source of data mining can be divided into 4 parts: websites, social media, mobile devices and internet of things (IoT). As data mining can be done manually in small scales, mostly automatic crawlers are used to gather and structure large amounts of data. Results of data mining have many uses. Big data is used in many fields, but especially for e-commerce and logistics areas. With the concept of big data, consumer behaviour can be predicted by obtaining massive amounts of data and then analyzing it. [1]

While these data mining methods have yielded large profits for businesses and resulted in positive contributions to information science and systems, the downsides are undeniable. Data breaches, leaks and misuse of personal information collected in data mining operations have proliferated as this industry has grown. Oftentimes, these instances occur as a result of poor planning and are easily preventable. For example, Schweiger and Ladwig [2] describe the missteps that led to the infamous breach of Target's customers' credit card information in 2013.

They cite a report produced by the Sans Institute that covered the breach, which stated that the attackers utilized simple google searches to investigate the state and architecture of Target's online database infrastructure. Much of the information retrieved from this and used to carry out the attack was found in a publicly available case study examining how Target handles its data.

One of the principal issues concerning privacy in data mining and big data analytics is the improvement of the classical data mining methods and techniques to become more *privacy-aware* [3]. Current improvements to these methods include reducing the degradation of data utility resulting from the data transformations used in privacy preserving data mining, and increasing the ability to make better predictions of the underlying patterns in the privacy concerned data.

The collection and processing of a lot of data led to the formation of the concept of "personal data" all over the world, especially in Europe, and the binding of data collection to the rules. Therefore different Privacy Preserving Data Mining (PPDM) methods are generated. PPDM mainly considers two aspects. First, how to guarantee that the information such as ID card number, name, address etc is not revealed in the data analysis process. The next is how to make more advantageous applications in light of privacy-preserved data. [2]

The structure of our paper is as follows: Chapter 2 includes a literature review focused on identifying the methods and techniques used for privacy preservation in data mining and big data analysis, followed by the respective applicability and current applications of the highlighted methods. Chapter 3 includes our analysis of the efficacy of the proposed methods and their fundamental limitations, followed by the author's recommendations for privacy preservation in future research through data mining and big data analysis. Conclusion part takes place in Chapter 4.

2 Literature Review

For this paper, we sought to identify the main methods and techniques used for privacy preservation in data mining and big data analysis, their applicability and current applications. This was done by aggregating and investigating proposed algorithmic approaches, critical analyses of the proposed methods, and of the state of privacy preservation in the field of data mining and big data analytics. In later sections we will provide our critical analysis of the described methods, looking at their advantages, disadvantages, and the limitations of these proposed methods. Praminik et al. [4] highlight two principal methods among others for manipulating data with the purpose of privacy preservation: randomization and anonymization.

2.1 What is data privacy?

Data privacy is not data security. Data privacy concerns the production, handling, and protection of personally identifiable information, and other confidential data. Personally

identifiable information is defined as “any information which can be used to distinguish the identity of an individual alone or when combined with other personal or identifying information which is linked or linkable to a specific individual” [5]. Maintaining privacy in data mining therefore implies that no piece of information within the dataset or gained through analysis can lead to identification of an individual. However, there is a fundamental problem with the nature of privacy when it comes to data mining algorithms that has to do with data utility. Data utility is the amount of useful information that can be extracted from a set of data.. Quoting researchers Dwork and Pottenger [6], “If the database said nothing, or produced only random noise unrelated to the data, then privacy would be absolute, but such databases are, of course, completely uninteresting.”, there appears to be an inverse relationship between privacy and utility of a dataset. To combat this issue, researchers have sought to develop methods to increase dataset utility while maintaining individual privacy of the dataset constituents.

2.2 Classification of PPDM Methods

From a top to bottom approach, PPDM methods can be classified into three groups [7], while the topic of this paper will focus on techniques found in one of the groups:

1) Heuristic-based techniques: In heuristic-based techniques, only the predetermined private or sensitive data is modified. The loss of effectiveness is minimized in these techniques by focusing only on a set of values, rather than all available values.

2) Cryptography-based techniques: Cryptography-based methods include encryption implementation, which makes the operation very secure but complex and requires more computation time according to other techniques.

3) Reconstruction-based techniques: In reconstruction-based techniques, the original distribution of the data is reconstructed from the randomized data. Both of the techniques mentioned in this article fall into this category.

2.3 Detailed Explanation of the Methods Mentioned:

The basic form of data in a table consists of the following four types of attributes [4]:
(i) Explicit Identifiers: are a set of attributes that contain information that clearly identifies a record owner, such as name, identification number, address, etc. They are mostly unique, or there are only several more in the set.
(ii) Quasi Identifiers: are a set of attributes that, when combined with publicly available data, can potentially identify a registrant.

(iii) Sensitive Attributes: are a set of attributes that contain sensitive personal information such as disease, bank account, salary, etc.
(iv) Non-Sensitive Attributes: are a set of attributes that are not problematic when disclosed even to untrusted parties. In most cases it is really hard to distinguish quasi identifiers and non-sensitive attributes, since identifying the record owner from attributes is related to size and distribution of the data set.

2.3.1. Anonymization based PPDM Methods:

In anonymization based PPDM, private or sensitive data is transformed such that the data becomes less specific than the original version. This anonymized data can be given as input to the data miner. In this way, reconstruction of sensitive data is prevented.The 3 main types of anonymization methods are: k-anonymity, l-diversity, and t-closeness methods. For all practical purposes, k-anonymity is the base method, and the other two methods are extensions of it.

k-anonymity: k-anonymity method can be explained as creating a modification in which a record owner cannot be distinguished from (k-1) record owners based on the information in a given data set. [8]

l-diversity: k-anonymity method has some weaknesses, and record owners can be identified when the people in the given data set have some common identifiers. To overcome this weakness, the l-diversity method is proposed to guarantee that sensitive data is not only k-anonymous, but also diverse from each other. [9]

t-closeness: t-closeness method is proposed when the l-diversity also has some limitations and requires that the distribution of a sensitive attribute in any equivalence class be close to the distribution of the attribute in the overall table. [10]

Anonymization based PPDM methods are simple and scalable, easy to implement, do not require much computation time and preserve the semantics of the data for data mining algorithms.

2.3.2. Randomization based PPDM Methods

In the randomization based methods, noisy signals are given to the data set to hide the true values of the individual records. The noise added to the data is significantly large to preserve the individual value of the recordings. However, the aggregate behavior of the data distribution can be reconstructed by removing noise from the data. The restructured distribution is usually sufficient for various data mining tasks. Adding A and B creates a new distribution of

C. Given that the distribution of B is known to everyone, we can estimate the distribution obtained by subtracting B from C. [4]

In the randomization method, data collection is performed using two steps. In the first step, the data providers randomly select their data and transfer the random data to the data receiver. In the second step, the data importer reconstructs the original distribution of the data using a distribution reconstruction algorithm.

The randomization based methods are relatively simple and do not require information about the distribution of other records in the data. Therefore, the randomization method can be applied at data collection time. It does not require a trusted server containing all the original records to perform the anonymization.

3 Critique of PPDM Methods

3.1 Critique of Anonymization Methods

Narayanan and Shmatikov [11] demonstrate that with little background information (quasi identifiers), a method of de-anonymizing supposedly anonymized data can fairly easily reveal potentially sensitive information about individuals. In their 2008 publication at the IEEE Symposium on Security and Privacy, "Robust De-anonymization of Large Sparse Datasets", they describe how they were able to perform a de-anonymization technique using little background information of an individual, as well as public information from the International Movies Database in order to decode the records of Netflix users, and detect sensitive information such as their apparent political affiliations. This data set belonging to Netflix was anonymized using the 'k-anonymity', and ultimately what Narayanan and Shmatikov proved here was that this method of anonymization becomes less effective on larger scales of data sets. [12]

Narayanan & Shmatikov [11] also demonstrate that k-anonymization methods *completely* fail on high dimensional sets of data, which contain most modern real-world datasets of individuals. Not only, but that these methods do not guarantee privacy for the individual, as the values of particularly sensitive attributes using quasi-identifiers may not be diverse enough to maintain data anonymization.

Backstrom, Dwork, and Kleinberg [13] demonstrated the fragility of using data anonymization methods. They invented several techniques for attacking anonymized data sets with the result of identifying anonymized data nodes in social networks. Shmatikov [14] concludes that anonymization alone cannot be relied on to effectively preserve individual privacy, in other words, that simple removal of identifiers is insufficient to maintain non-identifiability.

3.2 Critique of Randomization Methods

With the development of their de-anonymization algorithm, Narayanan & Shmatikov [11] conjecture that the amount of randomization through noise (perturbation) that would need to be applied to the datasets in question to not be defeated by their algorithm would effectively remove all data utility. This presents a problem for collaborative filtering, which is the prediction of a user's future choices based on past behavior and behavior or similar users, and thus destroys the usability of the dataset. Oliveira & Zaiane [15] discovered that they ran into misclassification issues when a noise randomization method was used alongside clustering large datasets. Aggarwal [12] discovered that not only is randomization unable to achieve data privacy in datasets with high dimensionality, but it is still vulnerable to individual node identification through natural properties of datasets, such as outliers and clusters.

Another fundamental issue with perturbation is that it is computationally expensive and slow for datasets of large size. Analyses of noise-randomized datasets are unable to provide desired results in meaningful run times, further decreasing the data utility. Dwork [6] mentions that discovering techniques for yielding sufficient data utility on datasets of a large enough size is difficult, and entirely not straightforward.

3.3 Guidance and Future Needs for PPDM

Moving forward, an important step would be to reach an agreed upon approach to effectively analyzing methods of privacy protection data mining, as many papers on the subject note the need for firm and clear definitions and ways of quantitatively examining data privacy. Clifton, Kantarcioglu and Vaidya [16] assert the need for an agreed upon method of ensuring privacy is upheld, as they cite that many different papers and sources offer their own descriptions of ways to uphold privacy. They specifically cite Agrawal & Aggarwal 2001's differential entropy metric and classification accuracy as potentially strong starting points for reaching a consensus on the preservation of privacy. They write that what is needed is a computational toolkit for building and assessing data mining techniques, and assert that although difficult, this is a feasible goal.

Malik et al. [17] explore the state of PPDM techniques in a paper published at the International Conference on Computer and Communication Technology, titled "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects". They note that each PPDM approach has its respective advantage, but one thing they emphasize is the fact that these approaches must be improved as far as quantitatively determining their efficacy. A potentially effective way of achieving this would be defining PPDM as its own engineering science - this can be achieved by optimizing the tradeoffs between disclosure, utility and costs for efficient PPDM algorithms, and integration of PPDM solutions to the industry as opposed to

merely having academics research them, Ultimately, what this would amount to is a cohesive toolkit for privacy preserving data mining, as proposed by Clifton, Kantarcioglu and Vaidya.

Pramanik et al. [4] additionally summarize four parameters that ought to be integral to determining the overall efficacy of a method of privacy preserving data mining. They list first *Data utility*: i.e how useful a data set can still be once a method of privacy preservation has been applied to it, then *Robustness*: which refers to how effective the method of privacy preservation is, in other words how resistant the method is to attacks. Then *Complexity*: in terms of computational power of the privacy preservation method, and *Efficiency*: as far as how efficiently data can be analyzed and transformed into data sets that preserve privacy. Vitaly Shmatikov affirms the need for utility of privacy-preserved data when he writes, "*Although developing an ideal privacy-preserving method that will provide zero privacy and zero utility losses is practically impossible, the maximum utility of the data should be maintained without compromising the underlying privacy constraints. For assessing any privacy method, only the quantification of privacy is insufficient without quantifying the utility of the data created by the privacy method.*".[14] It does appear that the future for research of privacy preservation in data mining as we understand it will focus on maximizing data utility and privacy, despite the nature of its inverse relationship.

4 Conclusion

Overall, while promising advances have been made in the area of privacy preserving data mining, they currently still come with faults that can and have been exploited and must be improved. What this field requires is further robust research, funding and acceptance within the industry of data mining. The final goal, though realistically impractical, would be to reach a consensus approach of sourcing and compiling data securely and privately, and a significant step towards this would be achieving a consensus method of analyzing how effective a PPDM technique is in protecting its participants' data while still providing useful information.

Due to the fundamental limitations of current PPDM techniques and methods highlighted in this paper, much of future research will be directed at a) discovering new methods and techniques with increasing data privacy and utility simultaneously and b) understanding the nature of adversaries and their capabilities in their plight to compromise datasets and obtain private information. As we continue the path forward in time, the amount of data that we as humans will generate will only increase, highlighting the dire need for measures to ensure privacy for each and every individual whose data was gathered.

References

[1] Yuguang, Wang et al. (2021). *Review of Data Scraping and Data Mining Methods*. J. Phys.: Conf. Ser. 1982 012161

[2] Schwieger, D., & Ladwig, C. (2016). Protecting privacy in big data: A layered approach for curriculum integration. Information Systems Education Journal, 14(3), 45-54. http://isedj.org/2016-14/n3/ISEDJv14n3p45.pdf

[3] Qi X, Zong M. (2012). *An Overview of Privacy Preserving Data Mining. Procedia environmental sciences*. 12:1341–7.

[4] Pramanik, MI, Lau, RYK, Hossain, MS, et al., (2021). Privacy preserving big data analytics: A critical analysis of state-of-the-art. *WIREs Data Mining Knowl Discov*. 11:e1387. https://doi.org/10.1002/widm.1387

[5] Storage Networking Industry Association. (2021). *What is data privacy?* SNIA. Retrieved November 21, 2021, from https://www.snia.org/education/what-is-data-privacy.

[6] Dwork, C., & Pottenger, R. (2013). Toward practicing privacy. *Journal of the American Medical Informatics Association*, *20*(1), 102–108. https://doi.org/10.1136/amiajnl-2012-001047

[7] Vaghashia, H., & Ganatra, A.P. (2015). A Survey: Privacy Preservation Techniques in Data Mining. *International Journal of Computer Applications, 119*, 20-26.

[8] Samarati, P. & Sweeney, L. (1998). *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression* (). SRI International

[9] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L -diversity: Privacy beyond k -anonymity. ACM transactions on knowledge discovery from data. 2007;1(1):3–es.

[10] Ninghui Li, Tiancheng Li, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: 2007 IEEE 23rd International Conference on Data Engineering. IEEE; 2007. p. 106–15.

[11] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*. https://doi.org/10.1109/sp.2008.33

[12] Aggarwal, C. C., (2007). On randomization, public information and the curse of dimensionality. IEEE 23rd International Conference on Data Engineering; Istanbul, Turkey, pp. 136–145.

[13] Backstrom L, Dwork C, Kleinberg J. (2007). Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In: Proceedings of the 16th international conference on world wide web. ACM; 2007. p. 181–90.

[14] Shmatikov, V. (2011). Anonymity is not Privacy. *Communications of the ACM*, *54*(12), 132–132. https://doi.org/10.1145/2043174.2043198

[15] Oliveira, S., & Zaiane, O. Data perturbation by rotation for privacy-preserving clustering, 2004.

[16] C. Clifton, M. Kantarcioglu, and J. Vaidya. Defining privacy for data mining. In H. Kargupta, A. Joshi, and K. Sivakumar, editors, National Science Foundation Workshop on Next Generation Data Mining, pages 126–133, Baltimore, MD, Nov. 1-3 2002

[17] Malik, Majid Bashir, et al. “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects.” *2012 Third International Conference on Computer and Communication Technology*, 2012, https://doi.org/10.1109/iccct.2012.15.

[18] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren. Information Security in Big Data: Privacy and Data Mining. IEEE access. 2014;2:1149–76.

[19] Samarati, P. & Sweeney, L. (1998). *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression* (). SRI International

[20] Kreso, I, Kapo, A, Turulja, L. Data mining privacy preserving: Research agenda. *WIREs Data Mining Knowl Discov*. 2021; 11:e1392. https://doi.org/10.1002/widm.1392