Data Privacy in Its Three Forms – A Systematic Review

Amen Faridoon(✉) and Mohand Tahar Kechadi(✉)

University College Dublin, Dublin, Ireland
amenjadoon2@gmail.com, tahar.kechadi@ucd.ie

Abstract. The constant growth of large amounts of data have gifted many fields, such as healthcare, business, e-commerce, education, social sites, and others, to make timely decisions and improve their services for their users. However, the considerable amount of these applications' data is of personal nature. Thus, the sensitive information of each individual should be protected to gain the trust of users how their private information is shared with the organizations. The privacy and security of the data are the dominating challenges and attracted much attention in recent times. They explored privacy threats and also introduced many privacy-preserving techniques to deal with a variety of data threats. In this paper, we present a systematic review of the techniques that have been used to tackle these threats depending on the state of the data – whether it is at rest (in data stores), in transit (over the network), or in-use (during the analysis). This has shown very interesting conclusions about the data privacy and security with regard to big data characteristics.

Keywords: State-of-the-art privacy techniques · Data at rest · Data in transit · Data in use

1 Introduction

In recent years, big data has gained significant attention from researchers and industrial experts as the world has faced challenges related to big data storage, transmission, management, processing, analysis, visualization, integration, architecture, security, quality, and privacy. It has been noted that 90% of the world's data was collected in the last two years. Moreover, this newly emerging area, called data science faces a major challenge that is, the collected data is usually private. It contains sensitive information, such as person-specific private and sensitive data; age, gender, zip code, disease, caste, shopping cart, religion, etc., and data analytics is prone to privacy violations. The most common case is when this data is released to third party who can access it and analyze it, they may extract valuable knowledge, and lead to inference attacks and violation of individuals' privacy [7].

The privacy of individuals is on the stake in each category of data that is; at rest, in transit or in-use. The Data at rest category is governed by the inactive

T. K. Dang et al. (Eds.): FDSE 2020, CCIS 1306, pp. 425–433, 2020.
https://doi.org/10.1007/978-981-33-4370-2_30

data that is physically stored in a device. Due to the large amount and valence dimension of big data, data holders utilize new platforms for storing their data such as cloud data centers [16]. The use of these platforms high alerts the data holders that they require some new privacy-protection methods to make sure the privacy and security of their data are not violated. Data in transit is governed by data sharing. The challenge here is how to share a large volume of data without violating the privacy of individuals. This is usually called privacy-preserving data sharing. On the other hand, velocity and variability dimensions of the big data cause greater hindrance to monitor the traffic in real-time, and inconsistencies in data types, speed, and formats lead to privacy and security risks [28]. Data in use category is governed by the analysis of the data to extract useful knowledge. The objective here is how to analyse (or mine) the data without revealing individuals identifiable and sensitive information. This is called privacy-preserving data analytics. The data scientists usually have direct access to the dataset during the process of big data mining. In this case, there are two types of possible privacy violations: 1) intentional or unintentional leakage of personal information to an unauthorized party. 2) results of the analytic algorithm can violate the privacy of individuals such as linkage, re-identification, and other attacks are possible.

1.1 Need for Privacy-Preserving Techniques

Some common privacy-preserving threats that an individual may face after sharing their private information with an organization are: **Surveillance:** Many organizations use surveillance tools to observe the behavior of their customers and make product suggestions. But this is a serious privacy threat because surveillance can lead to more serious matters [17]. **Disclosure:** After removing identity attributes, data holders publish it or hand it to third-party for analysis. However, with the help of quasi-attributes, data analysts can match this data to other available datasets and disclose the sensitive information of a person [8]. **Embarrassment and abuse:** Big data analytics models not only contribute positively but also have some negative implications in the life of a person. **Discrimination:** is a serious noticeable privacy threat. When sensitive information of an individual is disclosed, discrimination can happen. Statistical analysis of electoral results is an example of discrimination. These forms of attacks are extremely dangerous, and one needs to deal with them urgently [4].

1.2 Our Contribution

Many regulations were put in place to protect personal data in domains. However, most of these regulations do not enforce absolute confidentiality, which would cause more harm than good, but rather protect individually identifiable data that can be traced back to an individual with or without external knowledge. This gave rise to a wide range of studies primarily focusing on the privacy-preserving data sharing and analytics techniques at a larger scale with the objective of keeping the data private while extracting useful knowledge from it.

We have conducted a wide review of various data privacy-preserving techniques. After deeply analyzing each of them, we come up with the classification of the state of the art techniques that best ensure the privacy of data in its three forms. The main objectives of this systematic review are as follows:

- Classify of privacy-preserving techniques into three forms of data.
- Study the existing data privacy-preserving techniques in its three forms.
- Highlight the gaps or limitations and possible attacks faced by the data privacy techniques.
- Perform critical analysis of data privacy techniques in relation to big data characteristics.

2 Classification of Privacy-Preserving Techniques

In this section, we present categorisation of privacy-preserving data mining techniques along with a brief analysis of existing techniques for all the three forms (at rest, in transit and in use).

2.1 Data at Rest

De-identification. Replacing sensitive values with more general values is called de-identification [15]. Any recognisable information should be first anonymised with semantically consistent values of generalisation and suppression before the data is analysed. The reverse process of de-identification is called data re-identification. The key de-identification techniques are K–anonymity, L–diversity, and T–Closeness. There are some common terms used in these techniques.

- **Personal Identity Attributes:** Personal identifying attributes are type of attributes that uniquely or directly identify the identity of a person like name, national identity card number, phone number, etc.
- **Quasi Attributes:** Type of attributes whose values match with external data to re-identify a person like gender, zip code, age, etc.
- **Sensitive Attributes:** Sensitive attributes contain the sensitive information of an individual. A person does not want to share it with others like salary, disease, etc.
- **Insensitive Attributes:** General information of an individual.

K–Anonymity: Sweeney et al. introduced a technique called K–anonymity in 1998 [14] to tackle the problem of privacy violation. The records that are present in the dataset are anonymised if the values of the attributes of each record cannot be distinguish from (K-1) other data records. So, personal identification columns are omitted. Suppression and Generalisation are the two methods that are most commonly used for data distortion to reach anonymity. Many algorithms are employed for creating attribute hierarchy to achieve anonymity the algorithms proposed in [23,26]. Hence, K–anonymity provides the primary outline for privacy-preserving. However, this technique has limitations. It cannot protect the attribute disclosure, and background knowledge, temporal, homogeneity attacks are also possible [6].

L-Diversity: To overcome the flaws of the k–anonymity approach, the L–diversity approach was proposed [19]. Distance, entropy and recursive l-diversity models represent the extension of k-anonymity. In this technique, sensitive attributes present in each equivalence group must have L diverse number of values for their representation. Suppose that the values of a sensitive attribute are positive and negative and only 1% of them are positive. This may allow adversaries to gain significant knowledge about individuals of a particular class. However, l–diversity has some drawbacks. If the original dataset has more than one sensitive attribute then it would be difficult to achieve l–diversity. It is even more challenging if the sensitive attributes have not diverse values. Therefore, it is also not sufficient to prevent attribute disclosure. Due to the velocity and variety dimensions of big data sometimes L–diversity is not possible.

T-Closeness: approach was proposed to enhance the l–diversity technique [15]. The threshold is decided to reduce the gaps between the equivalence classes. The dataset is considered to be t–closeness if all the equivalence classes have t–closeness, while the distance among the distribution of sensitive attribute values in an equivalence class and the distribution of attributes in the whole dataset is less than the threshold. There are two common distance measures; earth mover's distance and kullback-leibler are used to maintain the distance among the values of the sensitive attribute within a class and in the middle of the equivalence groups. T–closeness technique removes the quasi identifier attributes by replacing them by the most general value present in the hierarchy tree. However, the appropriate data distribution of a sensitive attribute is not always possible. Separate protection of quasi-identifiers is also deficient in t–closeness [13].

Randomisation consists of adding random noise in an original data. The new distorted data values are usually generated by probability distribution [1]. Randomisation is simple than the other privacy protection techniques and it does not need to know the information of other data records. It is applicable at the time of data collection and pre-processing. However, it also has some weaknesses such as, the randomisation process is not scalable with the increasing sizes of the datasets. The accuracy of the results is also affected by the additional noise. Moreover, the adversary can gain access to sensitive information of individuals with the help of faraway points present in the dataset because randomization cannot have a significant impact on outliers.

Cryptographic Techniques. There are a number of cryptographic techniques that are employed to preserve the privacy in a distributed environment. It solves the problem of an untrusted environment. Cryptographic technique, such as homomorphic encryption can encrypt the data while it is sharing among different parties during processing in a collaborative environment [27]. This technique enables the data holder to process their encrypted data without privacy violation. Moreover, cryptographic approaches empower the parties to compute their results from aggregated input rather than sharing their original data [22,24].

However, the involvement of a large number of participants can slow down the computation process. The use of data encryption during data analytics is difficult and it can also reduce the accuracy of the analysis.

2.2 Data in Transit

Distributed Privacy. Distributed privacy is popularized along with the attractiveness of distributed data mining. The goal of this technique is to perform data analysis within a distributed environment without violating the privacy of the original data [12]. Many algorithms, such as Naïve Bayes, ID3 decision tree, K-nearest neighbours, Support vector machine have been implemented in a distributed environment. However, distributed privacy has some limitations. Some participants are fully or partially adversaries and they do not wish to share their local data records with others.

Secure Multi-party Computation (SMC). The concept of two-party computation for solving the problem of two millionaires introduced and extended to secure multi-party computation problems [20]. Secure multi-party computation is a sub-field of cryptography with the aim to allow a number of parties to jointly compute some known functions on private data in a distributed environment without revealing their individual data sites. Multi-party computation (MPC) system has three basic roles: 1) Set of input sites: provide input data to the trusted computation. 2) The result sites: received the results from the trusted computation. 3) The computing sites: mutually computing the trusted computation. Each member of the MPC computation may have more than one role. In real world, multi-party computation has many applications; Jana system that provides MPC-secure database developed by Galois Inc., Cybernetica also developed MPC-secure database, Partisia used MPC for their commercial activities since 2009, ... [2]. However, MPC protocols need each pair of parties to communicate with one another but in many cases it is not feasible for all pairs of parties to exchange messages like an application running between a web server and a number of clients, communication rounds between parties depend on the depth of the network. Moreover, the involvement of large number of parties makes the implementation of MPC more complex.

2.3 Data in Use

Differential Privacy. In 2006, differential privacy was introduced to protect the information of individuals [10]. The goal of this approach is to give roughly equal privacy to each entity of the dataset. The analyst does not have direct access to the dataset. Four-step process is completed without violating the privacy of entities: 1) analyst can create a query on database, 2) privacy mechanism accepts the query to calculate privacy risk, 3) and execute analyst's query on the database, 4) at the end, the privacy mechanisms adds noise component (according to the calculated privacy risk) to the original results and give it back to the

analyst. Noise component depends upon the privacy risk, [9]. In recent years, many organisations have used the differential privacy during the analysis stage, to protect the private and sensitive information; U.S Census Bureau in 2008 [18], Google's RAPPOR in 2014 [11], Google in 2015, Apple in 2016, Microsoft in 2017 [5], Privitar Lens in 2019 and LinkedIn in 2020. **Limitations:** one major challenge encountered by them is related to auxiliary information. For instance, a person "A" height is a very delicate piece of information and the disclosure of "A's" height is considered as privacy violation. If anyone only has access to the auxiliary information that "the person "A" has a height three inches shorter than the average height, they can deduce the person "A's" height.

Privacy-Preserving Machine Learning. The involvement of machine learning in big data analytics introduces the need for privacy on four stages i.e.; training, input, output, and model privacy. Big data analytics (or data in use) category has further divided the privacy approaches into two categories: data-centered and model-centered approaches. Data-centered privacy-preserving techniques are directly used on big data with the goal to protect sensitive data of individuals, such as homomorphic encryption [3], differential privacy [25], federated learning, secure multi-party computation [21], etc. Conversely, the combination of technologies can by some means protect the model from black-box (adversaries gain access to the functions of the model without having the internal knowledge about the model) and white-box (adversaries gain access to the individual's contributing information by reaching the internal knowledge of the model) attacks like homomorphic encryption plus Differential privacy, Secure Multiparty Computation plus Differential privacy, etc. Techniques used in side-channel attacks using machine learning could be used as a starting point for building countermeasures for real hardware-software systems (e.g., AI-enabled Security Watchdog).

3 Critical Analysis

The existing privacy-preserving approaches did not consider explicitly big data dimensions. Table 1 summarises the investigation that whether the particular technique is capable to cover 3Vs (volume, velocity, and variety) or not.

K-anonymity can deal with the volume and velocity dimensions, because data size and speed do not affect its fundamental principle. It needs to make quasi-identifier (QID) classes with at least k members that are moderately simple for huge volume of data. However, it does not support the variety, because the attribute categorisation (PID, QID, sensitive and non-sensitive attributes) is difficult when the data is of unstructured or semi-structured nature. L-diversity is more suited for large amounts of data. However, it cannot deal with the data velocity, because it is hard to designate a new record and balance equivalency class. If the data has stream and heterogeneous nature then the use of t-closeness is not possible, because it is difficult to find the closeness using variational, earth mover's distance and kullback-leibler distance measures.

Randomisation technique cannot stand with any of the 3Vs dimensions of the big data because of time complexity. Like randomisation, distributed privacy also cannot deal with any of the 3Vs of the big data, because of time complexity. The key purpose of this technique is to mine shared data records. If multiple sides contain large amount and increasing speed of data generation then its time complexity. Cryptographic techniques are a suitable tool for the big data features like volume and velocity because the primary purpose of cryptographic algorithms is to encrypt the data.

Differential privacy is the only contender for the variety dimension. The reason for its suitability is that it is not based on attributes. The key role of differential privacy is to add random noise to the result of the query and this does not affect with large quantity and heterogeneity of the data.

Table 1. Comparison between data privacy-preserving techniques and the characteristics of big data

Techniques	Volume	Velocity	Variety
K–Anonymity	✓	✓	✗
L–Diversity	✓	✗	✗
T–Closeness	✗	✗	✗
Randomization	✗	✗	✗
Distributed privacy	✗	✗	✗
Cryptographic techniques	✓	✓	✗
Differential privacy	✓	✗	✓

4 Conclusion

Manipulating and analysing private data is a very critical issue. Various techniques have been proposed to overwhelm the problem of privacy violation of individuals. But the blemishes of the already existing methods enforce to continue the research efforts in this area. Whereas, homomorphic encryption, secure multiparty computation and differential privacy are performing well while data is at-rest, in-transit and in-use respectively. Moreover, using machine learning in improving the security of devices and systems should also be explored. Techniques used inside-channel attacks employing machine learning could be used as a starting point to build counter measures. However, involvement of machine learning in big data analytics has introduce more challenges for researchers. In addition, most of the existing privacy protection techniques are not able to cope with the dominating characteristics of the big data (volume, velocity and variety). The differential privacy is the only contender for the variety dimension. Privacy-preserving data mining still requires advancement.

References

1. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 439–450 (2000)
2. Archer, D.W., et al.: From keys to databases-real-world applications of secure multi-party computation. Comput. J. **61**(12), 1749–1771 (2018)
3. Aslett, L.J., Esperança, P.M., Holmes, C.C.: Encrypted statistical machine learning: new privacy preserving methods. arXiv preprint arXiv:1508.06845 (2015)
4. Chabot, Y., Bertaux, A., Nicolle, C., Kechadi, T.: An ontology-based approach for the reconstruction and analysis of digital incidents timelines. Digit. Invest. **15**, 83–100 (2015)
5. Ding, B., Kulkarni, J., Yekhanin, S.: Collecting telemetry data privately. In: Advances in Neural Information Processing Systems, pp. 3571–3580 (2017)
6. Domingo-Ferrer, J., Torra, V.: A critique of k-anonymity and some of its enhancements. In: 2008 Third International Conference on Availability, Reliability and Security, pp. 990–993. IEEE (2008)
7. Ducange, P., Pecori, R., Mezzina, P.: A glimpse on big data analytics in the framework of marketing strategies. Soft Comput. **22**(1), 325–342 (2017). https://doi.org/10.1007/s00500-017-2536-4
8. Duncan, G.T., Lambert, D.: Disclosure-limited data dissemination. J. Am. Stat. Assoc. **81**(393), 10–18 (1986)
9. Dwork, C.: Ask a better question, get a better answer a new approach to private data analysis. In: Schwentick, T., Suciu, D. (eds.) ICDT 2007. LNCS, vol. 4353, pp. 18–27. Springer, Heidelberg (2006). https://doi.org/10.1007/11965893_2
10. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
11. Erlingsson, Ú., Pihur, V., Korolova, A.: RAPPOR: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 1054–1067 (2014)
12. Fukasawa, T., Wang, J., Takata, T., Miyazaki, M.: An effective distributed privacy-preserving data mining algorithm. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 320–325. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28651-6_47
13. Helma, C., Gottmann, E., Kramer, S.: Knowledge discovery and data mining in toxicology. Stat. Methods Med. Res. **9**(4), 329–358 (2000)
14. Khanaa, V., Thooyamani, K.: Protecting privacy when disclosing information: k anonymity and its enforcement through suppression. Int. J. Comput. Algorithm **1**(1), 19–22 (2012)
15. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, pp. 106–115. IEEE (2007)
16. Liu, C., Yang, C., Zhang, X., Chen, J.: External integrity verification for outsourced big data in cloud and IoT: a big picture. Future Gener. Comput. Syst. **49**, 58–67 (2015)
17. Liu, Y., Guo, W., Fan, C.I., Chang, L., Cheng, C.: A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. IEEE Trans. Ind. Inform. **15**(3), 1767–1774 (2018)

18. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 277–286. IEEE (2008)
19. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data (TKDD) **1**(1), 3-es (2007)
20. Micali, S., Goldreich, O., Wigderson, A.: How to play any mental game. In: Proceedings of the Nineteenth ACM Symposium on Theory of Computing, STOC, pp. 218–229 (1987)
21. Mohassel, P., Zhang, Y.: SecureML: a system for scalable privacy-preserving machine learning. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 19–38. IEEE (2017)
22. Pinkas, B.: Cryptographic techniques for privacy-preserving data mining. ACM SIGKDD Explor. Newslett. **4**(2), 12–19 (2002)
23. Samarati, P.: Protecting respondents identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
24. Sgaras, C., Kechadi, M.T., Le-Khac, N.A.: Forensics acquisition and analysis of instant messaging and VoIP applications. In: Garain, U., Shafait, F. (eds.) IWCF 2012, IWCF 2014. LNCS, vol. 8915, pp. 188–199. Springer, Cham (2012). https://doi.org/10.1007/978-3-319-20125-2_16
25. Song, S., Chaudhuri, K., Sarwate, A.D.: Stochastic gradient descent with differentially private updates. In: 2013 IEEE Global Conference on Signal and Information Processing, pp. 245–248. IEEE (2013)
26. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **10**(05), 571–588 (2002)
27. Tran, N.H., Le-Khac, N.A., Kechadi, M.T.: Lightweight privacy-preserving data classification. Comput. Secur. **97**, 101835 (2020)
28. Xu, L., Jiang, C., Wang, J., Yuan, J., Ren, Y.: Information security in big data: privacy and data mining. IEEE Access **2**, 1149–1176 (2014)