Research on big data privacy protection technology

Tianyou Lu

Wuhan Sannew School, Wuhan, Hubei, China, 430090

377813780@qq.com

Abstract. The extensive application of big data technology makes data burst with unprecedented value and vitality. However, due to the large amount of data, many data sources, and complex data access relationships, the current development of privacy protection technology is seriously lagging behind the development of big data technology, which restricts the application and promotion of big data. At present, it is urgent to sort out the development status of big data privacy protection technology, so as to provide a reference for the research and breakthrough of key issues of big data privacy protection. This paper analyzes k-anonymity and differential privacy protection, and points out that a protection method that can reduce the cost of high-dimensional data processing and ignore the background knowledge requirements of attackers is urgently needed for privacy protection in big data scenarios. At the same time, this paper also puts forward suggestions for the application and future development of these technologies.

Keywords: Big Data, Privacy Protection, Data Privacy

1. Introduction

With the rapid development of information technology, we are ushering in the era of big data. Computer networks provide convenient conditions for people's work and lives, break the restrictions formed by space for people to transmit information, and change the lives and work of the public. At present, the application of big data is being promoted to many fields, such as e-commerce, national defense, and scientific research.

While big data has brought changes to production and lifestyles, its security problems have become increasingly prominent. In 2017, the data leakage of Equifax, a US credit reporting agency, resulted in the personal sensitive information of almost half of the US population being in the hands of hackers. In 2018, Cambridge Analytica illegally collected Facebook user information and interfered in the U.S. election. These security incidents show that improvement in big data privacy protection has become an urgent requirement to ensure the steady advancement of the digital economy.

However, different from traditional data, big data has the characteristics of large volume, variety and speed, which brings great challenges to the application of privacy protection technology in big data environments. The research on big data privacy protection is still in its infancy, researchers still have differences in the core cognition and key characteristics of big data privacy protection, and there is still a gap between theoretical results and practical application requirements. This paper reveals the scope of application and shortage of current privacy protection services by analyzing four commonly-used privacy protection technologies, giving a reference for future adjustments and the appearance of new privacy protection technologies.

2. Publishing data anonymity

Data anonymity publishing can be applied to three categories of data: structured data, graph data, and location data.

For structured data, a typical anonymity scheme is k-anonymity proposed by Sweeney [1]. This scheme generalizes and compresses the quasi-identifier values recorded in the data set, so that all records are divided into several equivalence classes, each of which contains at least k records with the same quasi-identifier, so that the identification information can be hidden. This solution is mainly oriented to static data. In view of the dynamic nature of big data release, Bu et al. [2] proposed an anonymous technology that supports data insertion, deletion and modification. This anonymous technique can resist attackers by combining historical data analysis and reasoning.

Graph data anonymization not only hides the user's identity and attribute information, but also hides the relationship between users. It mainly includes super node-based methods and structural transformation-based methods. The supernode-based anonymous scheme uses supernodes to segment and cluster the graph structure. Fu et al.segmented the attribute connection and social connection of nodes, which made up for the shortcomings of the existing methods for insufficient perturbation of the association between attribute distribution and social structure [3]. The most typical structural transformation anonymity scheme is subgraph k-anonymity, but this purely structural anonymity method cannot completely conceal the node-attribute association. Yuan et al. [4] proposed a “k-degree-l-diversity” anonymous model to solve this problem by adding noise nodes to the original graph.

Location data includes location data and trajectory data. Gruteser and Grunwald first introduced the concept of k-anonymity into the field of location privacy protection, and proposed using location k-anonymity to ensure that the location queried by users contains at least k different users [5]. However, location-only anonymity techniques cannot effectively solve the problem of trajectory leakage. Trajectory k-anonymity prevents trajectory-based identification attacks by making any trajectory in an area containing at least k trajectories at any time.

It can be seen that the current data anonymization technology is mainly improved on the basis of the original k-anonymity scheme [1]. However, the privacy protection effect of the k-anonymity scheme is easily affected by the distribution of data. Availability may be seriously degraded after processing. On the other hand, k-anonymity schemes usually assume background knowledge possessed by attackers, but in big data scenarios, attackers can obtain unknown background knowledge from various sources [6]. These are the key problems that need to be solved in future data anonymization research.

3. Data publishing based on differential privacy

Differential privacy is a privacy protection method first proposed by Dwork based on a solid mathematical foundation, without considering any background knowledge the attacker may have. Depending on the implementer of data privacy processing, differential privacy can be divided into centralized differential privacy (CDP) and local differential privacy (LDP).

Centralized differential privacy requires data to be collected in the data center first, and then processed by the data center for privacy. The representative scheme of CDP was proposed by Dwork [7]. The principle of CDP is that for two data sets that differ by one record, after the processing function processes them, a random function is used to add noise to the result, so that the two data sets have almost the same probability of producing the same result. In this case, missing any record in the data set has almost no effect on the final output result. The current CDP mainly includes two noise mechanisms: the Laplace mechanism for numerical data and the exponential mechanism for non-numerical data [8].

Localized differential privacy is proposed for untrustworthy third-party data managers. After the user performs data perturbation that satisfies differential privacy locally, the data is sent to the collector. The LDP perturbation mechanism mainly includes random response, information compression and distortion, where random response is the mainstream perturbation mechanism in LDP, and its basic principle is to answer its own real situation with a certain probability for a sensitive problem, protect individual privacy information through this uncertainty, and then correct the collected data by correcting The function restores the true statistics [9].

CDP and LDP face some common challenges in the big data environment. First, big data is complex and heterogeneous, and there may be correlations between the records of the data set, especially for graph data. The current CDP and LDP methods cannot guarantee the data comparison. Secondly, the release of high-dimensional data has become a major bottleneck of differential privacy methods due to the increase in disturbance error and computational complexity. The current solution is mainly to reduce the high-dimensional data, but there is no solution now to effectively solve the huge communication cost problem of high-dimensional data publishing.

4. Conclusion

As a new research topic, big data privacy protection still needs more attention. Although privacy protection technology has begun to take shape, there are still many problems with every protection method. Aver years of development, the classic k-anonymity scheme can now be applied to graph data and location data, besides common structured data. But in the big data scenario, it gradually becomes obsolete because attackers can obtain unknown background information through big data easily. On the other hand, differential privacy protection doesn't require attackers' background knowledge, and provides a solution for privacy leakage on the data collectors' side. But it isn't sensitive to correlation between data, which means attackers may bypass the protection through their reasoning. It is also unsuitable for publishing high-dimensional data because of its high cost and inaccurate when processing high-dimensional data. To sum up, as big data technology becomes mature, its characteristics of huge volume, combined sources, and complex relationships post serious threats to privacy protection, requiring more effort to adapt present privacy protection methods to this new big data scenario.

References

[1] Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst, 2002, 10: 557 – 570

[2] Bu Y, Fu A W C, Wong R C W, et al. Privacy preserving serial data publishing by role composition. Proc VLDB Endow, 2008, 1: 845–856

[3] Fu Y Y, Zhang M, Feng D G, et al. Attribute privacy preservation in social networks based on node anatomy. J Sotfw, 2014, 25: 768–780

[4] Yuan M X, Chen L, Philip S Y, et al. Protecting sensitive labels in social network data anonymization. IEEE Trans Knowl Data Eng, 2013, 25: 633–647

[5] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, San Francisco, 2003. 31–42

[6] Huo Z, Meng X F. A trajectory data publication method under differential privacy. Chin J Comput, 2018, 41: 400–412

[7] Dwork C. Differential privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, Venice, 2006. 1–12

[8] Friedman A, Schuster A. Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD Interna- tional Conference on Knowledge Discovery and Data Mining, Washington, 2010. 493–502

[9] Warner S L. Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc, 1965, 60: 63–69