Overview of Teaching Cases of "Big Data Privacy Protection" Oriented to Practical Ability Training

Kaiyue Hu, Wenju Liu, Ze Wang

Abstract—**Under the background of the new era, the society has put forward higher requirements for computer professionals. The traditional professional practice and innovation teaching link is faced with many problems, which is difficult to meet the needs of the society for innovative and comprehensive talents. In order to promote big data teaching reform, promote the integration of scientific research and teaching, improve students' practical ability and innovation ability, and meet the requirements of the new era of professional talent training goals, this paper analyzes and introduces the location of big data in big data. Location big data not only makes people's lives more convenient, but also brings the risk of personal sensitive information leakage. This paper summarizes the location privacy protection methods and trajectory privacy protection methods, introduces two privacy protection models, and provides a certain basis for the reform and development of space-time big data in big data courses.**

Index Terms—**big data, location big data, privacy protection, teaching reform**

I. INTRODUCTION

The rapid development of big data, an emerging discipline, has brought us many conveniences. Under the background of the big data era, the use of big data science and technology to collect useful data information from various data sources, and to conduct data preprocessing, analysis, modeling and display has been widely used in various disciplines. Innovating the teaching mode in colleges and universities is the objective need of knowledge dissemination and social development. In order to promote the combination of teaching and practice, and improve students' ability to analyze and practice big data, the project team summarized the research progress in the field of "big data", and provided teaching cases for the training of students' practical ability for the "big data" related in class and extracurricular talent training links.

In recent years, with the rapid development of mobile network and location technology, a large number of mobile devices and mobile applications based on location-based services have been produced. People can use location-based services more conveniently, making location-based services cover all aspects of people's daily life, greatly improving people's quality of life, and people's dependence on location-based services is gradually increasing [1]. These location-based applications can be roughly divided into two categories: real-time and non real-time. Real time applications refer to mobile users' access to services by providing real-time locations to servers, including location queries, navigation functions, advertising, mobile social network services, entertainment services, road condition queries, etc. Non real time application refers to the location service provider's collection, processing and analysis of mobile users' movement patterns, or the release of relevant data for commercial use. For example, after the transportation department obtains track data, it can be used to analyze urban road construction or traffic management strategies, and the commercial department can make business decisions by analyzing users' movement patterns.

However, while enjoying and utilizing the convenience brought by these applications, a series of security problems have gradually emerged. In the general location service, users send the request information containing their own location information and query content to the location server for processing. Once the information is eavesdropped by malicious attackers during the transmission to the server, the malicious attackers will obtain the user's location information. If the user requests continuous query services at this time, the malicious attacker can easily infer the user's movement track during this period of time by continuously eavesdropping on the user's location information and mastering the user's location information at different times. It is usually possible to infer a lot of privacy information from the user's track information, such as his/her workplace, home address, some personal habits, etc. The user track privacy disclosure may even make the user subject to the personal harassment and personal attack of malicious attackers, seriously threatening the user's security [2]. Therefore, how to protect users' location privacy is a problem worthy of in-depth study.

II. CONCEPT INTRODUCTION

A. LBS Overview

Location service refers to obtaining the location information of mobile users through wireless communication devices and positioning technology, and providing this information to the mobile users themselves, others or application systems to achieve various services related to the current user's location [3].

In a system, the communication subjects are usually mobile users, base stations and location database servers. Mobile users are the requesters of network location services, usually mobile devices with location function. A base station is an entity that provides location application services to users. It collects users' personal information and determines whether to grant access to the location server based on users' identification information, and then sends legal location

Manuscript received October 05, 2022.

Kaiyue Hu, School of Software, Tiangong University, Tianjin, China.
Wenju Liu, School of Computer Science and Technology, Tiangong University, Tianjin, China.
Ze Wang, School of Computer Science and Technology, Tiangong University, Tianjin, China.

service requests to the location server. The location server uses the location system to obtain the location information of users, and provides users with different quality of service and different levels of location information.

The three communication subjects exchange by sending request response information to each other. When the base station receives the requester's service request for location information, it requests the location server to find the relevant location information. At the same time, it uses some communication protocols to negotiate the quality of service, service cost, and other relevant application parameters. The location server processes relevant requests according to the request. After receiving the location information from the location server, the base station sends the information to be queried to the requester.

B. Location privacy protection technology

The main purpose of location privacy [4] protection is to prevent or reduce the ability of users to identify their specific location information in location-based service systems. There are two methods to protect location privacy: one is to protect the identity information of users, so that the server can not determine the true identity of the requester. The other is to protect the user's location information, that is, the user does not send his or her real location to the server, but sends a region that protects the user's location to the server, and the server searches for the corresponding results to the user according to this region. Common basic methods for location privacy protection are as follows:

1) False position anonymity: create false position to achieve the effect of confusing the real with the false. The main idea of the algorithm is to generate virtual non-existent users around users requesting services according to some rules, and these users also send application requests at the same time, so that the attacker can determine which is the target.
2) Temporal and spatial anonymity: A user's location is expanded into a spatiotemporal area, and information is sent by delaying until K users have visited the same location, and then K user requests are sent at the same time. The anonymous area contains K users, achieving the effect of anonymity.
3) Location k anonymity: This method requires that the location information sent by the mobile client to the location server is indistinguishable from the location information of other K-1 mobile users, that is, K users' requests are sent at a time, and the location information of these K users is indistinguishable.

C. Track privacy protection technology

Track privacy protection methods are developed with the development of location privacy protection methods. As a special kind of personal privacy, track privacy has received extensive attention in recent years, and research on track privacy protection methods has also emerged in endlessly. The existing track privacy protection methods include track generalization, track suppression and false track methods.

1) Track generalization: the track information is generalized to the corresponding anonymous area to achieve the purpose of privacy protection. In this kind of technology, the most commonly used is the track K-anonymity technology. Track K-anonymity refers to that the anonymous server selects K-1 tracks that are indistinguishable from the user's real track before the track is published, and then uses the positions corresponding to the same time in the k tracks to form an anonymous area, so as to achieve track privacy protection.
2) Trajectory suppression: according to the conditional release of location data, sensitive data defined by users and locations frequently accessed by users are divided into sensitive data. For sensitive data, the method of non release or partial release is adopted to achieve privacy protection. The suppression method has the advantages of simple implementation and high privacy protection, and also has the disadvantage that too many suppressed points will cause serious data distortion.
3) False trajectory: There are two methods to generate false trajectory: random method and rotation method. The random method first selects the start point and end point of the false track, and then moves randomly from the start point to the end point based on the speed and movement type of the track to form a false track. The main idea of the rotation method is to rotate the real track by a certain angle through a sample point of the real track to form a false track.

D. Common attack models

1) Homogeneous attack: The existing privacy protection methods are usually supported by K-anonymity technology, and an anonymous group is formed by constructing K-1 false trajectory data and the original trajectory data through a specific method. When the frequency of a sensitive location is too high, the attacker is easy to obtain fragmented sensitive information. Combined with other background information, the attacker will easily narrow the scope of attack.
2) Background knowledge attack: The main purpose of the attacker is to infer the real track data belonging to the target moving object from the anonymous track data. When the attacker has the information associated with the target object, it can easily achieve the goal. In short, background knowledge attack refers to that the attacker infers the real track of the target object through the original background knowledge.
3) Similarity attack: The similarity between tracks is one of the important factors that affect the anonymity of tracks. The similarity attack is mainly due to the high similarity between the attributes of the anonymous track and the real track, such as the overall direction, close distance, and coincidence of location points. At this time, the attacker can mine sensitive information without inferring the real track of the moving object.

III. RELATED RESEARCH PROGRESS

A. Privacy protection mechanism based on probability inference

Probability based location privacy protection assumes that the malicious attacker knows all the background knowledge, calculates the risk probability of privacy disclosure of each published location information, and determines whether the published location information meets the privacy

requirements according to the risk probability. This method can quantitatively protect the user's location privacy under the attack model in which the attacker has complete background knowledge.

Tsai et al.[5] proposed a simple MaskSensitive method to suppress sensitive locations marked by users. Although this method can protect the privacy data of sensitive locations from being published, when attackers collect a lot of background knowledge, they can speculate that the currently suppressed location points are sensitive location points, which will cause privacy disclosure. Therefore, G ö tz M et al.[6] calculated the release probability of each location, that is, each location has a release probability, and the system will suppress or release each location according to the release probability. For context sensitive information privacy disclosure, Wang [7] considers context dynamics and the ability of malicious attackers to adjust attack strategies to identify context privacy issues, and describes the interactive competition between users and opponents as a competitive Markov decision process MDP. This method uses an efficient minimax learning algorithm to obtain users' optimal policies. Li [8] cited K-anonymity technology to probabilistic speculative privacy protection, and proposed a Maskk privacy protection algorithm based on K-anonymity. When the location was published, K suppressed locations were published, making attackers unable to identify sensitive locations.

B. Protection Mechanism Based on K-anonymity

The K-anonymity idea refers to combining the user's real location with other K-1 locations, and finally publishing the combined k-location data, so that attackers can not infer the user's real location from the published data, thus protecting the user's location information.

K-anonymity model was first proposed by L.Sweeney[9] and used in relational databases. Marco first introduced K-anonymity into location privacy protection. Domingo et al. [10] first gave track similarity indicators considering time and space factors on the basis of K-anonymity, then used aggregation algorithm to cluster tracks, and arranged track distance and position to achieve track K-anonymity. Gao [11] used the angle to determine the direction similarity of tracks when calculating the track similarity, converted the selection of K-anonymous clusters to the minimum spanning tree model, found the best K-anonymous cluster, and proposed a personalized anonymous model to select the track K-anonymous set. On the basis of traditional anonymous methods, Xu et al.[12] took the four characteristics of direction, speed, time and space as the basis of track similarity measurement, moved the tracks in the same cluster set in space, and realized K-anonymization of the tracks in the same cluster set. Xin et al. [13] adopted the Gibbs sampling clustering method to detect the identified regions, and further promoted the detected representative regions according to the rationality of the equivalence class.

C. Privacy protection method based on false data

The privacy protection technology based on false data forms false data by using false identification information and location information to replace the real information or by disturbing the real data, while ensuring that the disturbed data does not have serious distortion.

Song[14] considers the unreachability of location points, and proposes an effective false location technology to protect users' privacy. Gao [15] also proposed a method to generate virtual trajectories. By selecting a certain number of user partners, several trajectories similar to user trajectories are constructed at intervals to protect user location privacy and provide accurate location information, so that users can have high-quality service quality. In addition, the location privacy of users includes not only the real location information, but also other privacy information of users (identity, hobbies, work units, etc.). Yang et al. [16] considered the problem of auxiliary information, semantic diversity and physical dispersion of location, and proposed a false location selection algorithm (SPDDS) based on location semantics and physical distance. By combining three objectives (location semantic diversity, query probability and physical dispersion of location), they solved the single objective optimization problem. Hara [17] proposed a location anonymization method, which mixes false locations that move according to user preferences with those that do not.

IV. CONCLUSION

In the age of big data, users' location data will be collected from different perspectives and channels. While the location big data brings huge benefits to people's lives, business operations and scientific research, it also brings serious privacy threats to people because it contains privacy information such as people's behavior patterns, habits and preferences, and sensitive information.

This paper introduces the concepts related to location privacy and trajectory privacy, and reviews and summarizes the research results of location big data privacy protection technology from the perspective of probability based conjecture privacy protection mechanism, K-anonymity based privacy protection mechanism and pseudo data based privacy protection method. For the teaching reform, it provides teaching cases for the cultivation of students' practical ability in the talent training link of the "big data analysis" course.

REFERENCES

[1] F. Xu, P. Zhang, and Y. Li, “Context-aware real-time population estimation for metropolis,” in Proc. ACM UbiComp, Heidelberg, Germany,2016, pp. 1064–1075.

[2] H. Jiang, P. Zhao, C. Wang, and J. C. Lui, “Roblop: Towards robust privacy preserving against location dependent attacks in continuous lbs queries,” IEEE/ACM Transactions on Networking, vol. 26, no. 2, pp.1018–1032, 2018.

[3] Zhou A Y, Yang B, Jin C Q, Ma Q. Location Based Services: Architecture and Development. Chinese Journal of Computers [J],2011,34(7):1155-1171

[4] M.Duckham, L.Kulik. Dynamic and Mobile GIS: Investigating changes in space and time[M].Drummond. Boca Raton: CRC Press,2006,35-52

[5] Tsai J Y, Kelley P G, Cranor L F, et al. Location-sharing technologies: Privacy risks and controls[J]. Isjlp, 2010, 6: 119.

[6] Götz M, Nath S, Gehrke J. Maskit: Privately releasing user context streams for personalized mobile applications[C]. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. 2012: 289-300.

[7] Wang W, Zhang Q. Privacy preservation for context sensing on smartphone[J]. IEEE/ACM Transactions on Networking, 2016, 24(6): 3235-3247.

[8] Li J, Bai Z H, Yu R Y, et al. Mobile Location Privacy Protection Algorithm Based on PSO Optimization [J]. Chinese Journal of Computers[J], 2018, 41(5): 1037-1051.

[9] Sweeney L A . k-anonymity[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002.
[10] Gruteser M , Grunwald D . Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking[C]// Proceedings of the First International Conference on Mobile Systems, Applications, and Services (MobiSys 2003), San Francisco, CA, USA. 2003.
[11] Gao S, Ma J, Sun C, et al. Balancing trajectory privacy and data utility using a personalized anonymization model[J]. Journal of Network and Computer Applications, 2014, 38: 125-134.
[12] Xu H J, Wu Q H, Hu X M. Privacy protection algorithm based on multicharacteristics of trajectory[J]. Comput. Sci, 2019, 46: 190-195.
[13] Xin Y, Xie Z Q, Yang J. The privacy preserving method for dynamic trajectory releasing based on adaptive clustering[J]. Information Sciences, 2017, 378: 131- 143.
[14] Song D, Song M, Shakhov V, et al. Efficient dummy generation for considering obstacles and protecting user location[J]. Concurrency and Computation: Practice and Experience, 2021, 33(2): e5146.
[15] Gao S, Ma J, Shi W, et al. LTPPM: a location and trajectory privacy protection mechanism in participatory sensing[J]. Wireless Communications and Mobile Computing, 2015, 15(1): 155-169.
[16] Yang D, Ye B, Chen Y, et al. A Dummy Location Selection Algorithm Based on Location Semantics and Physical Distance[C]. International Conference on Information Security Practice and Experience. Springer, Cham, 2021: 283-295.
[17] Hara T. Dummy-based Location Anonymization for Controlling Observable User Preferences[C]. 2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019: 1-7.