Big Data Analysis Course Case Survey

Xiaoqi Wang, Ze Wang

Abstract— **In the era of big data, the development of location sensing technology such as mobile communication and sensing devices has formed location big data, which has brought huge benefits to people's life, business operation methods and scientific research. Due to the diverse uses of location big data and the overlapping redundancy of contents, the classical privacy protection methods based on "informed consent" and anonymity can not fully protect user privacy. The privacy protection technology of location big data measures the user's location privacy and protects the user's sensitive information in the sense of information theory. This paper introduces the concept of location big data and the privacy threats of location big data, summarizes the unified metric-based attack model for location big data privacy, and summarizes the existing research achievements in the field of location big data privacy protection. The user's knowledge of text content and online behavior can be used to carry out reasoning attacks on users. Social relational reasoning and attribute reasoning are two basic attacks on user privacy in social networks. The research on the protection mechanisms and methods of inference attacks is also increasing. This paper classifies and summarizes the research and literature related to privacy inference and protection techniques, and finally discusses and looks into the future.**

Index Terms—**Recommendation algorithm, Attribute Reasoning, Teaching case.**

I. Introduction

In the era of big data, the development of location sensing technology such as mobile communication and sensing devices has digitized the geographical location of people and things. The sensor chip in the moving object collects the location data of the moving object in a direct or indirect way: on the one hand. GPS, WiFi and other positioning devices embedded in mobile devices such as mobile phones and vehicle-mounted navigation devices can directly obtain the accurate location information of moving objects at any time, and publish the collected location information through various ways. For example, some new applications of mobile social networks can publish the location information of the user at any time [1]. On the other hand, the accelerations, optical images and other data collected by widely used sensor devices such as wearable devices can also accurately determine the user's location information after processing [2-4].

Artificial intelligence, big data analysis and Internet + application are important development fields related to the training of IT professionals in China's development planning based on new engineering and new infrastructure. In order to promote the integration of teaching and science research, higher education workers should actively explore the demonstration effect of high-level research results of disciplines and industries on talent training, and carry out the reform practice of hierarchical and progressive practical teaching based on the integration of production and education. In order to promote the mutual integration of frontier academic progress and education and teaching work, mutual support. This project team summarizes the research progress in the field of "big data analytics", and provides teaching cases for the cultivation of students' analytical ability for the in-class and extracurricular talent training links related to "big data analytics".

The speed and scale of automatic location information collection by sensors far exceed the processing capacity of existing systems. According to statistics, each moving object submits its current position once every 15s on average. In this way, hundreds of millions of mobile phones, vehicle-mounted navigation devices and other moving objects submit more than million pieces of position information every second [5]. In the future, the progress of mobile sensing devices and the improvement of communication technology will generate location information more frequently. In the era of big data, such production speed and data scale bring great changes to people's life, enterprise operation and scientific research [6]. This kind of data is called location big data because it contains location information and has the characteristics of large scale, fast production speed and high value to meet the widely recognized big data [7].

Location big data not only brings huge benefits to people, but also brings harm to the disclosure of personal information. This is because location-based big data not only directly contains users' private information, but also contains users' personality habits, health status, social status and other sensitive information. The improper use of location-based big data will bring serious threats to users' privacy in all aspects. Some existing cases illustrate the harm of privacy leakage. For example, a well-known mobile application does not pay attention to the protection of location big data, which leads to the triangulation method to infer the user's home address and other sensitive locations, resulting in a number of criminal cases [8]. At the same time, a well-known mobile device manufacturer collected a large number of users' location data without the permission of the users [9-10], and the attackers could infer the users' personal sensitive information such as their physical condition through these location data [11-13]. After proper location privacy protection is provided for users, more people are willing to submit their mobile data to intelligent transportation, smart city and other analysis systems, thus providing more convenience for People's Daily life.

II. Related Work

A. Social Networks

Online social networking platforms have become an indispensable part of people's lives in modern society, and these

Manuscript received October 04, 2022.

Xiaoqi Wang, School of Computer Science and Technology, Tiangong University, Tianjin, China.
Ze Wang, School of Computer Science and Technology, Tiangong University, Tianjin, China.

enterprises have gained a large number of users. As of January 2020, Facebook had 2.4 billion users, ranking No. 1 among all social networking applications. The social network has the advantages of instant messaging, information sharing and Posting comments for users.

At first, people mainly used social networks to express some of their thoughts. As time goes by, online activities become more complex and diverse. The booming development of social networks has brought a large amount of user-generated content, with 66% of user tweets being about users themselves, most of which are free and publicly available [14].

In addition, more and more users are joining location-based Social networks (LBSN) to enjoy different location-related services, such as friend finding, Location of interest search, check-in, geo-tagged photo sharing, etc. Location information not only represents a person's geographical location, but also reveals their lifestyle, lifestyle and personal information, which leads to high privacy risks for users.

In social networks, users always want to share some information to gain revenue, while others are hidden to protect their privacy. Unfortunately, with the rapid development of machine learning, various powerful inference attacks may infer its hidden information [15].

B. Privacy Reasoning

Similar to the general definition of privacy, we believe that the privacy of location big data is the control of moving objects over their location data. In the era of big data, location data comes from a wide range of sources. The combination of location information of mobile objects at different times and background knowledge in location big data will reveal sensitive information such as health status, behavior habits and social status of users. For example, when a user is observed near a hospital, the general health status of the user can be inferred; The user's home address and other information can be inferred by considering the location where the user's trajectory starts and ends [16]. In addition, the information collected by the acceleration sensor, etc. only contains part of the location, which can also enable the attacker to effectively speculate the behavior pattern of the user [17].

The attacker who performs privacy reasoning can be any party interested in user privacy, such as cyber criminals, social network providers, advertisers, data brokers or surveillance agencies [18].Cyber criminals can use user privacy information to carry out targeted social engineering attacks; Social network providers and advertisers can use user data to target ads; Data brokers can profit by selling user information to advertisers, banking companies and other parties such as the insurance industry; Surveillance agencies can use this information to identify users and monitor their activities [19].

C. Inferential Attack Classification

According to the purpose of the attack, that is, to obtain the user's private information, the existing inference attacks can be roughly divided into two categories according to the purpose of the attack: inference on attributes and inference on social relations. In attribute reasoning, the location-specific reasoning is also a major research focus in the field, so it is introduced separately in this paper.

Attribute-based reasoning can be divided into several types of attribute-based reasoning methods according to the technology and different types of data used, such as content-based, social link-based and user behavior-based. Location-based reasoning includes social graph based and social behavior based methods. While the reasoning for social relationships is mainly divided into two methods: location-based and topic-tag-based.

III. Attribute-based inference attack

A. Sensitive Attribute Definition

There exists a dichotomous concept of user attributes, which can be divided into two categories: public attributes and private sensitive attributes. Users should determine which category their attributes belong to. Certain attributes (political leanings and race, for example) can be publicly displayed because a user's followers may follow him because of his public attributes. Others (gender and location, for example) are private and sensitive, and users do not want them displayed.

Attribute inference can be viewed as a method of reasoning about a set of sensitive attributes that a user does not want to be known to others from the information they post and interact with online.

The reasoned user attributes can be used for a variety of security-sensitive activities, such as spear phishing and authentication of personal information. In addition, an attacker can use the inferring attributes to identify the same user on multiple sites or form a comprehensive personal portrait of the user using offline records (for example, publicly available voter registration records), posing greater security and privacy risks to the user.

B. Content-based Attribute Reasoning

Content-based attacks mainly use topics, personal information and tweet text to reason about sensitive attributes of users.

Georgiou et al. [20] introduced an attribute inference attack based on community trending topics. From a statistical point of view, these public community-aware trending topics were used to infer the sensitive attributes of online social network users, because participating users in each topic formed homogeneous groups (communities), even if they did not have direct links.

A trending topic refers to a set of words or phrases related to a temporarily popular topic, which is used to understand and explain how information and memes spread through a huge social network with hundreds of millions of nodes [21].

The trending topic algorithm returns to the attacker a set of users who mention the offered topic. The attacker has general knowledge of the previous distribution of each attribute, such knowledge might include, for example, location distribution based on census, age distribution based on statistics published by social media services, gender distribution based on users who made this information public, etc. This increasing knowledge allows an attacker to gradually increase his inference confidence against a given user's sensitive attributes.

Thomas et al. [22] used the multi-label classification method to reason attributes, and proposed multi-party privacy to defend against attribute inference. Zhang et al. [23] showed that the subject tag in a user's tweet can be used alone to accurately infer the user's location, with an accuracy of 70% to 76%.

Otterbacher[24] studied gender reasoning using user writing styles. Narayanan et al. [25] showed a stronger result that authorship can be de-anonymized through writing style analysis. Adali and Gol-Beck et al. [26-27] used a user's tweets to study how personality can be inferred.

C. Attribute Reasoning Based on Social Links

He et al. [28] transformed attribute inference into inference on Bayesian networks constructed using social links between users, and evaluated their approach using a LiveJournal social network dataset with synthetic user attributes. The effects of prior probability, influence and social openness on attribute inference were also discussed.

Lindamood et al. [29] modified the Naive Bayes classifier to reason about some attributes based on social links and other public attributes of users. For example, other attributes of users (employer, city where users live, social friends of users and their attributes) were used to reason about users' majors. However, their method is not applicable to users who do not share any attributes at all.

Bhagat et al. [30] used the K-nearest neighbor algorithm based on ICA framework to reason the attributes of LiveJournal dataset, and proposed a local iterative algorithm to reason the attributes by selecting the most frequently occurring value in the local neighbors of a user node, which can be called majority voting of local neighbors.

Macskassy and Provost[31] proposed a neighbor relationship model and proposed two algorithms, namely, iterative neighbor and probabilistic neighbor, for attribute inference.

Mo et al. [32] proposed a graph-based attribute inference model, which used friend relationship, group membership and network relationship for similarity calculation, and used them as transition matrix to perform label propagation.

Yin et al. [33] used random walk and restart Social Attributes Network (SAN) to perform attribute ranking. They modeled attributes as nodes and established links between user nodes and attribute nodes. However, attribute relevance is not considered in the inference process, and a random walk causes labels to propagate in the network and stop at the nearest node. The methods based on vote allocation are similar in that they both use transition matrices to propagate labels among labels and finally select the closest attribute value.

Misolve et al. [34] proposed an attribute inference method based on community attributes. They infer the sensitive attributes of users according to the common attributes of users in the same community. Experiments were conducted on Facebook datasets to reason about the user's department of work, among other things.

Traud et al. [35] compared the community structure with Facebook-based partitions of a given category to examine the effect of common attributes on binary level data.

D. Attribute Reasoning based on user behavior

User behavior includes behaviors such as liking, following, and retweeting comments, which can be used to reason about attributes.

The method proposed by Kosinski[36] can easily use Facebook Likes in user behavior to automatically and accurately predict a series of highly sensitive personal attributes, including sexual orientation, race, religion and political views, personality traits, intelligence, parental divorce, age and gender, etc. Users and their likes are represented as a sparse user similarity matrix. If there is an association between users and likes, the item is set to 1, otherwise it is set to 0. The dimension of user-like matrices can be reduced using Singular Value Decomposition (SVD).Linear regression models are used to predict numerical variables such as age or intelligence, while logistic regression is used to predict dichotomous variables such as gender or sexual orientation.

The study of Chaabane et al. [37] proved that user behavior data can also be pages or lists that users like or share. Attackers (e.g., social platform providers, advertisers, or data brokers) can use machine learning classifiers to reason about private attributes of targeted users (e.g., gender, city of residence, and political leanings).

IV. Location-based Reasoning Attack

A. Location-based relational reasoning

The widespread popularity of location-based social networks such as Foursquare and location-based online services such as Uber has brought a wealth of human trajectory data. Understanding basic human trajectory patterns has proven valuable for a variety of applications, such as predicting the location of the next visit.

Hsieh et al. [38] used offline geographic activities of users (such as check-in records and meeting events) to reason about online social relationships. First, a co-address graph is constructed, where nodes are users, edges are co-addresses among users, and edge weights are combined eigenvalues. Two nodes with high closeness, probability and common location similarity have a high probability of knowing each other. Secondly, if the location of the meeting activity is more meaningful or important for both nodes, higher weight should be assigned to such co-addresses, and two people with higher meeting frequency tend to have social relations.

Zhang et al. [39] studied the problem of social relationship reasoning in a given LBSN by treating the spatial, temporal and social attributes of user pairs as different views of effective user links.

Backes et al. [40] deduced social relationships from users' locations, and used deep learning methods to learn users' mobile functions and apply them to social relationship reasoning. Works such as literature [41-45] can infer social ties from the same space and time, for which two users share a common friend or location.

Olteanu et al. [46] studied the influence of the same location information on location privacy. Recently, Zhou et al. [47] inferred social connections from friends and mobility data.

B. Others

Rahman et al. [48] proposed a multimodal approach to reasoning about social relationships, evaluating a real dataset of 22 million user posts collected from Instagram using five different dimensional features of users, namely images, tweet text, subject tags, geo-location, and (incomplete) social relationships. The results prove that when multiple patterns are combined, the success rate of inference attacks on social relationships is greatly improved.

Gupta et al. [49] studied the reasoning of people's social relationship in videos posted by social network users. They used audio-visual features and motion trajectories to calculate the measure of social relationship in each scene of the video, and used face recognition to calculate the appearance of people in each scene.

V.SUMMARY AND PROSPECT

Inference attack and protection technologies in social networks are in constant confrontation, and both technologies are improving. At present, the attacker has more and more knowledge, and the attack ability is stronger and stronger. The content of social network data is also more and more complex, including various attributes of users, including the relationship between users and other sensitive information.

In the aspect of attribute inference, attackers can get more powerful classifiers through adversarial machine learning in the future and use them to reason. Collect more user information, including cross-platform data, and use the correlation between attributes to perform better attribute inference. In location-specific reasoning, computer vision technology can be used to better identify the location of photos in tweets, and more spatio-temporal correlations between continuous social behaviors can be considered. For social relationship reasoning, some directions of future work include strengthening the learning of link weights of social graph models and extending the vote assignment attack to infer hidden social relationships between users.

REFERENCES

[1] Jabeur N, Zeadally S, Sayed B. Mobile social networking applications. Communications of the ACM, 2013, 56 (3) : 71-79. [doi:10.1145/2428556.2428573]

[2] Sousa M, Techmer A, Steinhage A, Lauterbach C, Lukowicz P. Human tracking and identification using a sensitive floor and wearable accelerometers. In: Proc. Of the IEEE Int 'l conf. on Pervasive Computing and Communications (PerCom). San Diego,2013. 166-171. [doi: 10.1109 / PerCom 2013.6526728]

[3] Ugolotti R, Sassi F, Mordonini M, Cagnoni S. Multi-Sensor system for detection and classification of human activities. Journal of Ambient Intelligence and Humanized Computing, 2013,4(1):27−41. [doi:10.1007/s12652-011-0065-z]

[4] Anguelov D, Dulong C, Filip D, Frueh C, Lafon S, Lyon R, Ogale A, Vincent L, Weaver J. Google street view: Capturing the world Computer, 2010,43(6):32−38. [doi: 10.1109/ Mc.2010.170]

[5] Civilis A, Jensen CS, Pakalnis S. Techniques for efficient road-network-based tracking of moving objects. IEEE Trans. On Knowledge and Data Engineering, 2005,17(5):698−712. [doi:10.1109/ tkde.2005.80]

[6] Mayer-schonberger V, Cukier K. Big Data: A Revolution that Will Transform How We Live, Work, and Think.eamon Dolan/Houghton Mifflin Harcourt, 2013. 102−105.

[7] Dijcks JP. Oracle: Big Data for the Enterprise. White Paper. Oracle, 2012.

[8] Chi HB. Three circle at three location: Weibo can locate. 2013. http://www.fawan.com.cn/html/2013-07/03/content_442649.html

[9] Williams C. Apple under pressure over iphone location tracking. 2011. http://www.telegraph.co.uk/technology/apple/8466357/Apple-underpressureover-iPhone-location-tracking.html

[10] Cheng J. How apple tracks your location without your consent and why it matters. 2011. http://arstechnica.com/apple/news/2011/04/how-appletracks-your-location-without-your-consent-and-why-it-matters.ars

[11] Hansell S. AOL removes search data on vast group of Web users. 2006. http://query.nytimes.com/gst/fullpage.html?res=9504E5D81E3FF93BA3575BC0A9609C8B63

[12] Wicker SB. The loss of location privacy in The cellular age. ommunications of The ACM, 2012,55(8):60−68. [doi: 10.1145 /2240236.2240255]

[13] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proc. of the IEEE Symp. on Security and Privacy.oakland, 2008.111 −125. [doi: 10.1109/ sp.2008.33]

[14] Ding S H, Fung B C, Iqbal F, et al. Cheung. learning stylometric representations for authorship analysis[J]. IEEE Transactions on Cybernetics, 2019,49 (1) : 107-121.

[15] Wang G, Wang B, Wang T, et al. Ghost Riders: Sybil attacks on crowd sourced mobile mapping services.IEEE/ACM Transactions on Networking, 2018,26 (3) : 1123-1135.

[16] Wicker SB. The loss of location privacy in The cellular age. Communications of The ACM, 2012,55(8):60−68. 10.1145 /2240236.2240255]

[17] Fitzpatrick M. Mobile that allows bosses to snoop on staff developed. BBC News. 2010. http://news.bbc.co.uk/2/hi/technology/8559683.stm

[18] Dwork C, Naor M, Pitassi T, Et al.Pan-private Streaming algorithms[C]//Proceedings of The First Symposium on Innovations in Computer Science, 2010.

[19] Gupta P, Gottipati S, Jiang J, et al.Your Love is Public Now: Questioning the use of personal information in Computer science Authentication [C]//Proceedings of the 8th ACM SIGSAC Symposium on Information Computer and Communications Security, 2013.

[20] Georgiou T, Abbad A E, Yan X.Privacy-preserving community-aware trending topic detection in online social media[C]//Proceedings of the IFIP Annual Conference on Data and Applications Security and Privacy, July 19-21, 2017:205-224.

[21] Al-kharji S, Al-Rodhaan M.Anovel (K, X) -Isomorphism Method for Taxation Privacy in Weighted Social Network [C]//Proceedings of the 21st Saudi Computer Society National Computer Conference (NCC), April 25-26, 2018.

[22] Thomas K, Grier C, Nicol D M. Friendly: Multi-party privacy risks in social networks[C]//Proceedings of the 10th International Conference on Privacy Enhancing Technologies, 2010:236-252.

[23] Zhang Y, Humbert M, Rahman T, et al.Tagvisor: A Privacy Advisor for Sharing Hashtags [J]. ArXiv: 1802.04122, 2018.

[24] Otterbacher j.nferring Gender of Movie reviewers: Exploiting writing style, The content and metadata 19 th [C] / / Proceedings of the ACM Conference on Information and Knowledge Management, October 26 to 30, 2010.

[25] Narayanan A, Paskov H, Gon Zhenqiang, et al.On the feasibility of Internet-scale Author Identification [C]//Proceedings of the IEEE Symposium on Security & Privacy 2012, May 20-23, 2012.

[26] Adali S, Golbeck J.Predicting personality with social behavior[C]//Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Aug 26-29, 2012:302-309.

[27] Golbeck J, Robles C, Turner K. Personality with Social Media [C]//Proceedings of the International Conference on Human Factors in Computing Systems, May 7-12, 2011:253-262.

[28] He J, Chu W, Liu Z.I Nferring Privacy Information from Social Networks [C]//Proceedings of the International Conference on Human Factors in Computing Systems, May 23-24, 2006.

[29] Lindamood J, Heatherly R, Kantarcioglu M, et al.Inferring private information using social network data[C]//Proceedings of the 18th International Conference on World Wide Web, April 20-24, 2008:1145-1146.

[30] Bhagat S, Cormode G, Rozenbaum i.Applying link-based classification to Label Blogs [C]//Proceedings of the 1st International Workshop on Social Networks Analysis, August 12-15, 2007:97-117.

[31] Macskassy S A, Provost F.A simple relational classifier[C]// Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 27, 2003:64-76.

[32] Mo M, Wang D, Li B, et al.Exploit of online social networks with semi- supervised learning[C]//Proceedings of the 2010 International Joint Conference on Neural Networks, May 23, 2010:1-8.

[33] Yin Z, Gupta M, Weninger T, et al.A unified framework for link recommendation using random walks[C]// Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, Aug 9-11, 2010:152-159.

[34] Misolve A, Viswanath B, Gummadi K P, et al.You Are Who You Know: Inferring user profiles in online social networks[C]//Proceedings of the Third ACM International Conference on Web Search and Data Mining, Feb 4, 2010:251-260.

[35] Traud A L, Mucha P J, Porter M A. Siocial Structure of Facebook Networks [J].Tatistical Mechanics and Its Applications, 2010 (in Chinese) Journal of Social Sciences, 2012, 391:4165-4180.

[36] Kosinski M, Stillwell D, Graepel T.Private traits and attributes are predictable from digital records of human behavior[J].National Academy of Sciences, 2013, 110:5802-5805.

[37] Chaabane A, Acs G, Kaafar M A.You are what you like! Information Leakage through users' interests[C]//Proceedings of the 19th Annual Network & Distributed System Security Symposium, Feb 5-8, 2012.

[38] Hsieh H P, Li C T.Inferring online social ties from offline geographical activities[C]//Proceedings of the ACM Transactions on Intelligent Systems and Technology.2019

[39] Zhang W, Lai X, Wang J.Social link inference via Multiview matching network from spatiotemporal trajectories[J].Transactions on Neural Networks and Learning Systems (Early Access), 2020 (4) : 1-12.

[40] Backes M, Humbert M, Pang J, et al. Walk2friends: Inferring social links from mobility profiles[C]// Proceedings of the ACM Conference on Computer and Communications Security 2017 (CCS 2017), Nov 3,2017.

[41] Eagle N, Pentland A S, Lazer d. yferring friendship network structure by using mobile phone data[J]. National Academy of Sciences, 2009,106:451. 15274-15278.

[42] Wang H, Li Z, Lee W C.PGT: Measuring Mobility Relationship using Personal, Global and temporal factors[C]//Proceedings of the 2014 IEEE International Conference on Data Mining, Dec 14-17, 2014: 570-579.

[43] Crandall D J, Backstrom L, Cosley Dan, 43.Inferring Social ties from geographic coincidences[J]. National Academy of Sciences, 2010, 107:22436-22441

[44] Scellato S, Noulas A, et al. Mascolo C.Exploiting place features in link prediction on location-based social networks[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 21-24, 2011:1046-1054.

[45] Pham H, Shahabi C, Liu Y.EBM: An entropy-based model to infer social strength from spatiotemporal data[C]//Proceedings of the ACM Special Interest Group on Management of Data, June 22-27, 2013:265-276.

[46] Olteanu Huguenin K, who R, et al., Quantifying interdependent privacy risks with the location data [J].IEEE the Transactions on Mobile Computing, 2017 (3) : 829-842.

[47] Zhou F, Wu B, Yang Y, et al. Vec2Link: Unifying heterogeneous data for social link prediction[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Oct 22-26201-8:1843-1846.

[48] Rahman T, Fritz M, Backes M, et al.Everything About You: A multimodal approach towards friendship inference in Online Social Networks [J]. ArXiv: 2003.00996, 2020.