Privacy in Location-based Data Mining

Bo Liu, Tianqing Zhu, Philip S. Yu

1 Definition

In a location-based service (LBS), privacy means that an individual user can enjoy the services offered with the expectation that their location will not be revealed to a degree that exceeds what is deemed acceptable.

From a privacy perspective, the location information in LBSs is not just a set of coordinates or the name of a place. It may also include the user's identity, spatial information (position), and temporal information (time). Hence, a user's location information can be defined as a tuple $< identity; position; time >$ [1], and location privacy can be defined as protecting the elements of this tuple. Blumberg et al. [2] define location privacy as:

Location privacy "*the ability of an individual to move in public space with the expectation that under normal circumstances their location will not be systematically and secretly recorded for later use*".

By this definition, location privacy has two main components: 1) the individual's expectation of "normal circumstances", which implies that location privacy is a relative thing and should be understood in terms of degree; and 2) the manner of "systematically and secretly", which pertains to the way information is collected and used. Consequently, in the context of location-based data mining, privacy concerns arise from two issues: a) location information collected for data mining purposes can be directly used to disclose a person's identity, behavior, and activities; and b) even more private information can be inferred from that data after it is mined (e.g., one's home or work address).

Bo Liu
University of Technology Sydney e-mail: bo.liu@uts.edu.au

Tianqing Zhu
University of Technology Sydney e-mail: tianqing.zhu@uts.edu.au

Philip S. Yu
University of Illinois at Chicago e-mail: psyu@uic.edu

2 Background

Global positioning systems (GPSs), now a standard component of most cell phones, are the driving force behind the rapid growth of LBSs. Since LBSs fulfill many useful and interesting needs in a wide range of fields, they should continue to be a fixture in our digital world for the foreseeable future. Mobile social networks, navigation, finding places of interest (POI), and augmented reality (AR) games are just a few of the practical applications that have benefited from LBSs [3]. However, with growth comes competition, and to ensure their services have the widest appeal, companies are beginning to amass information on an enormous number of individuals for behavioral analysis and pattern mining. Given these datasets contain extensive knowledge about a user's daily behavior, mobility patterns, and personal preferences, there are severe concerns over the vulnerability of people's privacy if this data were to be breached with location-based data mining. Here are just a few examples of how location information can be exploited to potentially violate privacy.

- A social network user shares check-ins to show the places they have visited. For most people, the distribution of these check-ins is a function of the distance between their home and their workplace. Therefore, an adversary can use this data coupled with when the check-ins occurred to predict the location of a user's home or office.
- A person regularly visits a hospital. Mining the dates and times of their visits, along with other information like their medical bills, will disclose sensitive information about their identity and illness.

The above examples illustrate the privacy risks associated with location-based data mining, providing a clear demonstration of how and why protecting the rich knowledge contained in location data against rapidly advancing data-mining techniques must be treated seriously.

3 Theory

3.1 Privacy Attacks

According to the current privacy laws like GDPR [4], a breach of privacy requires that a personal identifier is revealed. In the realm of location privacy, a personal identifier means any location data that could be used to reveal the identity of a specific person. Understanding this concept leads to some definitions of the types of privacy attacks seen in location-based data mining:

- Context linking attack: Location attacks may involve some contextual knowledge, such as a person's routine, which is easy to combine with observed locations to conduct a localization attack. Context can also be combined with precise

location information to conduct an identity attack. For example, if an adversary knows someone's home address and finds that address in a hospital's check-in list, the adversary can conclude that their target was in hospital at that specific time.

- Probability-based attack: This attack [5] is based on collected statistics about an environment. Using this method, an adversary can either perform a localization attack (location prediction) or an identity attack (de-anonymization). Strictly speaking, statistical information is actually a type of contextual information. However, since exploiting probability theory is an important category of attacks, it warrants a separate discussion.
- Machine/deep learning-based attack: Data mining techniques, especially recent advanced machine/deep learning techniques, can be used to breach privacy. For example, Li et al. [6] proposed a method for inferring user demographics, such as gender and education level, from mobile social networks based on machine learning. Their experiments showed a 70% success rate on a large real-world dataset. In another recent work [7], Weyand et al. showed that a convolutional neural network can determine the location of a photo just from its pixels.

3.2 Privacy Protection Schemes

Location-based data mining typically comprises two major processes: 1) collecting and preprocessing the location data; and 2) using the location data for data mining tasks. Accordingly, location-privacy protection schemes (LPPMs) can be discussed in two groups: 1) schemes that protect by manipulating the data (data LPPMs); and 2) schemes that protect through the design of the data-mining model or algorithm (LPPM algorithms).

3.2.1 Privacy Protection through Data Manipulation

Data LPPMs can be categorized into four different groups: cryptographic, anonymization, obfuscation, and information reduction.

Cryptography-based methods use encryption to protect a user's location, which can be applied to either the communications between peers or between the user and the server. For example, Mascetti et al. [8] designed a framework in which each user shares a secret with each of his friends through symmetric encryption. Then, when a friend is nearby, the user can be notified without needing to disclose their own location to the server. This can be an effective means of preventing privacy leaks by the server, but does not necessarily prevent leaks at the user level. Another approach proposed by Ghinita et al. [9] is to use private information retrieval (PIR) to provide location privacy. With PIR, the LBS server can answer queries without having to know or reveal any information about the query. PIR relies on the assumption of a quadratic residual, which states that for the product of two large prime numbers, it

is difficult to find a quadratic residual in the modular operation of a large composite number. The main problem with cryptographic mechanisms is their computational complexity and/or the need for a collaboration server.

Anonymization mechanisms aim to break the links between identity and location information. They fall into two main categories: *k*-anonymity and mix-zone.

k-anonymity [10] grants location privacy protection through generalization and suppression algorithms that ensure one location cannot be distinguished among all $(k-1)$ other records. Hence, a location is considered to be *k*-anonymous if it is indistinguishable from the locations of $k-1$ other users. Note that *k*-anonymity requires that the location privacy server be operated by a trusted third party (TTP), which knows all the precise user locations and acts as an anonymizer. Whenever a user needs to transfer a location with a query, the TTP calculates a set of *k* users and reports an obfuscated area containing *k* positions – one of which is the position of the querying user. *k*-anonymization works well for applications that do not require real or pseudo-identity, such as finding a nearby gas station or advertising the price of items while walking through a mall. However, spatially cloaked regions are vulnerable to inference attacks, so, if an LBS relies on conveying specific identities to provide its services, such as Exxon gas stations, *k*-anonymization techniques become ineffective [11]. Unlike *k*-anonymity, mix-zones can be used without user identity information. Beresford [12] proposed the first mix-zone approach, where the privacy of the user is maintained by constantly changing the user's name or pseudonym within the mix-zone. This method has since been studied in the context of several different applications, including vehicular networks and mobile crowdsensing.

Obfuscation mechanisms encompass a range of methods that reduce the precision of location information, through dummy locations or perturbation (adding noise).

The goal of dummy locations is to mask a user's true position by sending multiple false positions to the LBS server (i.e., "dummies"), along with the real position [13]. The dummies could be generated randomly or with additional contextual knowledge, such as the physical constraints of the real environment. Spatial obfuscation methods protect privacy by deliberately reducing the accuracy of the location information sent by the user to the LBS server and then to the client. Ardagna et al. [14] proposed a classical spatial obfuscation approach, where the user sends a circular area to the LBS server instead of the user's exact location.

Differential privacy, which is another form of obfuscation, and its applications in location protection have been studied in several recent papers. Geo-indistinguishability [15] is the key concept here, which is formally defined as protecting a user's location within a radius of r with a level of privacy that depends on *r*. Privacy levels are met by adding controlled random noise to the user's location. In general, obfuscation schemes sacrifice user utility. Although there are always trade-offs between utility and privacy, there are some special cases. For example, Soma et al. [16] investigated location-privacy protection in trip planning queries, devising a method to protect location privacy by sending false or hidden location information to the service provider while still providing accurate trip plans in response to queries. In

cases like this, obfuscation is a good choice for privacy protection because it does not degrade performance.

Reducing location information sharing, as the name states, simply means increasing privacy by reducing the amount of information shared. To date, cache systems have been one of the most effective ways to reduce shared data. For example, Niu et al. [17] proposed a cache-aware dummy selection algorithm that combines k-anonymity, caching, and side information to achieve both a higher degree of privacy and a better caching hit ratio. Liu et al. [18] proposed a framework that enhances the privacy of LBSs in wireless vehicular networks through active caching. Alternative approaches to information reduction incorporate aspects of game theory as a solution. For example, Liu et al. [19] proposed a framework that enhances the location privacy of mobile crowdsensing applications by reducing the number of bidding and assignment steps in the crowdsensing cycle.

3.2.2 Privacy Protection through the Algorithms and Models

Beyond data LMMPs, location privacy can also be protected in the way that models are designed and built. There are two main approaches in this category: differentially private machine learning [20] and federated learning [23].

Differentially private machine learning [20] means using a machine learning algorithm to build a model that provides a differential privacy guarantee. Since its inception, this strategy has been extended and refined by many researchers. For instance, Goodfellow et al. [21] introduced a simpler version of the differentially-private SGD (DPSGD) algorithm that ensures differential privacy by cutting the gradients in each layer down to a maximum l_2 norm. With this approach, a high-quality model can be trained with only a moderate privacy budget. The downside is that the DPSGD method offers differential privacy at the record-level, so privacy can suffer at higher levels, e.g., the user level. To solve this problem, Mcmahan et al. [22] introduced a user-level differentially-private algorithm called the DP-FedAvg algorithm to protect all of a user's data. Instead of limiting the contribution of a single record to training the model, DP-FedAvg limits the contribution of each user's entire dataset to training the model.

The other approach is federated learning [23], which provides a framework to mine data in a distributed manner. With this method, the privacy of individual users can be protected against adversarial collaborators or a central server. In federated learning, each device downloads the current model, improves it by learning from data on a local device, and then sums up the changes in a small centralized update. Only the model update is sent to the cloud, and it is sent in an encrypted communication. The shared model is then updated according to the average of all the users' updates. The risk to privacy is greatly reduced since all the training data stays on local devices and no updates from individual users are stored in the cloud.

It is worth noting that these two approaches can be used together. For example, Shokri et al.'s [20] early and well-known method of implementing differential privacy involves training a machine learning model in a distributed manner by updating

the selected local gradients and then adding noise to them within the privacy budget of each parameter.

4 Applications

With the rise of the smartphone, LBSs have boomed and, today, they are commonplace in a variety of contexts – health, entertainment, work, our personal lives, etc. At a base level, an LBS is a collection of location information with some personal identifiers, such as usernames and email addresses, that is mined by a service provider or third-party analyzer and sent to users in response to queries.

Typically, the only way a user can control the release of their own location information from an LBS is by choosing when and where they use it. Therefore, it largely falls to the service provider to offer privacy protection options and technologies to its users. This raises the question of "when during the process" should privacy protection techniques be implemented. There are several stages to choose from:

- during the collection stage, when the customer uses the LBS. During collection, a caching scheme can be used to reduce the amount of location data that is shared. For example, in a POI application, a user might use their own or a peer's previous query results instead of transmitting a fresh query, or the communication process could be encrypted to prevent eavesdropping by adversaries. Alternatively, if offered as an option in the application, a user might choose to obfuscate their precise location or join an anonymity group (k-anonymity).
- when preprocessing the data. Service providers can use anonymity to remove any personal identifiers from the dataset. Differential privacy, for example, guarantees that all users are indistinguishable from each other.
- at the data mining stage. The service provider could adopt a private data mining or federated learning framework. With a federated learning framework, data preprocessing becomes a distributed process, and the differential privacy mechanism would be applied to each user individually, i.e., local differential privacy.

References

1. M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Personal and Ubiquitous Computing*, vol. 18, no. 1, pp. 163–175, 2014.
2. A. J. Blumberg and P. Eckersley, "On locational privacy, and how to avoid losing it forever," *Electronic frontier foundation*, vol. 10, no. 11, 2009.
3. B. Liu, W. Zhou, T. Zhu, L. Gao, and Y. Xiang, "Location privacy and its applications: A systematic study," *IEEE access*, vol. 6, pp. 17 606–17 624, 2018.
4. EU, "The EU General Data Protection Regulation ," https://eugdpr.org/, 2019, [Online; accessed 19-July-2019].

5. R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. IEEE Symposium on Security and Privacy*, 2011, pp. 247–262.
6. H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense," *IEEE Transactions on Dependable and Secure Computing*, no. 1, pp. 1–1, 2016.
7. T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," in *Proc. European Conference on Computer Vision*. Springer, 2016, pp. 37–55.
8. S. Mascetti, D. Freni, C. Bettini, X. S. Wang, and S. Jajodia, "Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 20, no. 4, pp. 541–566, 2011.
9. G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: anonymizers are not necessary," in *Proc. ACM SIGMOD*, 2008, pp. 121–132.
10. B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proc. IEEE ICDCS*, 2005, pp. 620–629.
11. B. Palanisamy and L. Liu, "Attack-resilient mix-zones over road networks: architecture and algorithms," *IEEE Transactions on Mobile Computing*, vol. 14, no. 3, pp. 495–508, 2015.
12. A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive computing*, vol. 2, no. 1, pp. 46–55, 2003.
13. H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Proc. IEEE ICPS*, 2005, pp. 88–97.
14. C. A. Ardagna, M. Cremonini, S. D. C. di Vimercati, and P. Samarati, "An obfuscation-based approach for protecting location privacy," *IEEE Transactions on Dependable and Secure Computing*, vol. 8, no. 1, pp. 13–27, 2011.
15. M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. ACM SIGSAC conference on Computer & communications security*, 2013, pp. 901–914.
16. S. C. Soma, T. Hashem, M. A. Cheema, and S. Samrose, "Trip planning queries with location privacy in spatial databases," *World Wide Web*, vol. 20, no. 2, pp. 205–236, 2017.
17. B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Enhancing privacy through caching in location-based services," in *Proc. IEEE INFOCOM*, 2015.
18. B. Liu, W. Zhou, T. Zhu, L. Gao, T. H. Luan, and H. Zhou, "Silence is golden: Enhancing privacy of location-based services by content broadcasting and active caching in wireless vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9942–9953, 2016.
19. B. Liu, W. Zhou, T. Zhu, H. Zhou, and X. Lin, "Invisible hand: A privacy preserving mobile crowd sensing framework based on economic models," *IEEE Transactions on Vehicular Technology*, 2016.
20. R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. the 22nd ACM SIGSAC conference on computer and communications security*. ACM, 2015, pp. 1310–1321.
21. M. A. A. C. I. Goodfellow, "Deep learning with differential privacy," *CCS*, 2016.
22. H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private language models without losing accuracy," *arXiv preprint arXiv:1710.06963*, 2017.
23. J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.