Privacy-Preserving Inference of Social Relationships from Location Data: A Vision Paper

Cyrus Shahabi, Liyue Fan, Luciano Nocera
Integrated Media Systems Center
University of Southern California
shahabi, liyuefan, nocera@usc.edu

Li Xiong
Department of Math&CS
Emory University
lxiong@mathcs.emory.edu

Ming Li
Department of ECE
University of Arizona
lim@email.arizona.edu

ABSTRACT

Social relationships between people, e.g., whether they are friends with each other, can be inferred by observing their behaviors in the real world. Thanks to the popularity of GPS-enabled mobile devices or online services, a large amount of high-resolution location data becomes available for such inference studies. However, due to the sensitivity of location data and user privacy concerns, those studies cannot be largely carried out on individually contributed data without privacy guarantees. Furthermore, we observe that the actual location may not be needed for social relationship studies, but rather the fact that two people met and some statistical properties about their meeting locations, which can be computed in a private manner. In this paper, we envision an extensible framework, dubbed Privacy-preserving Location Analytics and Computation Environment (PLACE), which enables social relationship studies by analyzing individually generated location data. PLACE utilizes an untrusted server and computes several building blocks to support various social relationship studies, without disclosing location information to the server and other untrusted parties. We present PLACE with three example social relationship studies which utilize four privacy-preserving blocks with encryption and differential privacy primitives. The successful realization of PLACE will facilitate private location data acquisition from individual devices, thanks to the strong privacy guarantees, and will enable a wide range of applications.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*Security, integrity, and protection*; H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Social Relationship, Location Privacy

1. INTRODUCTION

SIGSPATIAL'15 November 03-06, 2015, Bellevue, WA, USA

DOI: http://dx.doi.org/10.1145/2820783.2820880.

For decades, anthropologists and social scientists have been studying people's social behaviors by utilizing sparse datasets obtained by observations [5] and interviews [16]. These studies received a major boost in the past decade due to the availability of web data (e.g., social networks, blogs and review web sites) [10, 13]. However, due to the nature of these datasets, these studies were confined to behaviors that were observed in the online world. Recently, due to the availability of high-resolution spatiotemporal location data collected by GPS-enabled mobile devices through mobile apps, e.g., Google Maps, Facebook, Foursquare, and WhatsApp, or through online services, such as geo-tagged contents (tweets from Twitter, pictures from Instagram, Flickr or Google+ Photo), etc., it has become possible to study social behaviors by observing people's behaviors in the real world, especially via location history [17, 15]. For example, if two people came close in contact with each other, we can infer that an item could be exchanged or a disease could be transmitted from one to the other. Another example is if two people were seen at the same places and at the same time, i.e., *co-occurred*, we can infer that they are socially connected [15]. Those inference studies can enable various applications, such as new member identification of organized crime, target advertising/recommendation, and contact tracing for epidemics.

The main impediment to utilize these location datasets to infer real-world social behaviors is the sensitivity of the raw location data. The downside of public location sharing can be illustrated by the website of "Please Rob Me" [1]. By publicly sharing location via check-ins, tweets, etc., attackers may infer when one is not at home. Furthermore, the anonymity of movement data is hard to achieve. In fact, De Montjoye et al. [6] studied fifteen months of human mobility data for one and a half million individuals and concluded that human mobility patterns are highly unique. In a dataset where location is specified every hour and the spatial resolution is coarsely given by antennas, four spatiotemporal points are enough to uniquely identify 95% of the individuals. Given the sensitivity of location data and the fundamental constraints to individual privacy, data holders may not be willing to share these datasets for social good.

However, our main observation is that to make such inferences about people's social behavior, we do not require the specific location information, e.g., the semantic of the location, but only the knowledge that two people have met (i.e., have been at the vicinity of each other for some period of time) and some statistics about how often they meet and the popularity of the locations at which they meet. For

example, to infer social connection [15], we do not need to know at which exact restaurant two people meet as long as we know the popularity of the location they meet at (e.g., quantified by location entropy) – the more popular the location the less the chance that they are socially connected and vice versa. Based on this core observation, our vision is that these social behaviors can be studied in a privacy-preserving manner if we can simply capture the "meeting" event, for example by using *encryption* on locations, and collect the frequency of meetings and the popularity of the meeting locations, for example by using *differential privacy* on location statistics.

To this end, we envision an extensible framework, dubbed Privacy-preserving Location Analytics and Computation Environment (**PLACE**), which enables social relationship studies by analyzing individually generated location data. PLACE utilizes an untrusted server and performs location analytics to support various social relationship studies, without disclosing location information to the server and other untrusted parties. Three use cases of PLACE will be presented: *Reachability* use case answers the question whether one person can be reached by another through a sequence of pairwise meetings during a period of time. *Social Strength* use case infers whether one person is socially connected (e.g., friends) with another. *Spatial Influence* use case estimates whether one person's behavior influences another.

In order to support the three use cases without revealing people's location information, we design four privacy-preserving building blocks: *Location Proximity*, *Co-Occurrence Vector*, *Location Entropy*, and *Followship*. These blocks are designed based on deep understanding of people's social behaviors and generic such that they can be utilized across use cases as well as to define new blocks. With encryption and differential privacy primitives, we envision a systematic approach in PLACE that simultaneously provides strong privacy guarantees for both location and statistics, and achieves high accuracy and efficiency for computing all the blocks over large-scale spatiotemporal data.

2. RELATED WORKS

A plethora of works has been developed to protect location privacy. Here we briefly review related works on Location Obfuscation, Cryptographic approaches, and Differential Privacy.

Some location obfuscation techniques hide the actual location among a set of dummy and send redundant queries to the server [23], while others adopt the concept of k-anonymity [19] and use a Cloak Region (CR) which includes the actual location as well as $k-1$ other users [8]. The privacy guarantee of obfuscation based techniques is weak. The actually location hidden among dummy locations can be disclosed by brute-force attacks. Cloaking approaches are prone to semantic disclose, when the CR region contains only one type of locations. More recent obfuscation methods [21, 2] extend differential privacy [7] for location protection and provide more rigorous semantic privacy protection. While the perturbed locations may be useful for applications that do not require exact locations such as k-nearest neighbor queries, they are not suitable for our problems that require more exact locations, e.g. contact tracing for epidemic studies.

Cryptographic approaches, such as Private Information Retrieval (PIR), Private Proximity Testing (PPT), and Searchable Encryption approaches offer strong privacy guarantee. PIR protocols, e.g., in [9] allow individual users retrieve their nearest neighbors, e.g., the nearest gas station, through an untrusted server, while the server learns nothing about the requesting user's location. However, such protocols protect only the querier's location, meaning other people's location data is disclosed to the server. PPT protocols [14, 24] enable a pair of mobile users to be notified through an untrusted server when they are within a threshold distance of each other, but otherwise reveal no information about their locations to anyone. Since peer-to-peer model is adopted, PPT protocols are not directly applicable in our setting where anonymized data and computation is centralized on an untrusted server. Furthermore, it is not straight-forward to utilize those schemes for computation complex blocks, such as Location Entropy. To enable common search over encrypted data on an untrusted server, many Searchable Encryption schemes, e.g., [18, 4], have been proposed. However, they are not suitable in our problem setting as most search schemes only provide equality checking over encrypted data, while the computation of spatial data normally require compute-then-compare operations.

Differential privacy [7] has become the state-of-the-art privacy paradigm for statistical databases. It guarantees that an adversary is not able to decide whether a particular individual is included or not in the published dataset, regardless of the amount of additional information available to the adversary. Many techniques have been developed to publish static datasets indexed in a hierarchical structure [22] and trajectories [3]. However, Differential Privacy sanitizes statistical information only and assumes the data aggregator is trusted (data records are disclosed). The privacy challenges in our setting, where the computation is done on an untrusted server and location data is held by individuals, cannot be fully addressed by Differential Privacy alone.

3. PLACE FRAMEWORK

To show the practicality of our vision, we envision a system dubbed Privacy-preserving Location Analytics and Computation Environment (**PLACE**), which will ingest location data from a large number of mobile devices. PLACE will enable location analytics using an untrusted server, e.g., cloud, and privacy will be ensured by encryption and statistical inference control. As in Figure 1, PLACE will provide four essential blocks, *Location Proximity*, *Co-Occurrence Vector*, *Location Entropy*, and *Followship*, all computed on an untrusted server without compromising the privacy of data providers. On the other hand, data consumers, e.g., epidemiologists, criminologists, intelligent analysts, and policy makers will utilize the analytics provided by PLACE to infer social relations from location data. Similarly, third party application developers can utilize the blocks and/or use cases to build their applications.

Use Cases. We showcase the applicability of PLACE with three following example social relationship studies:
1) *Reachability*: If two people have come in close contact or there exists a contact path between them through other people, we can infer one is "reachable" from the other for delivering a package, or contracting a disease [17].
2) *Social Strength*: If two people have been to the same places at the same times, i.e., *co-occurred*, we can infer that they are socially connected. Depending on the places vis-

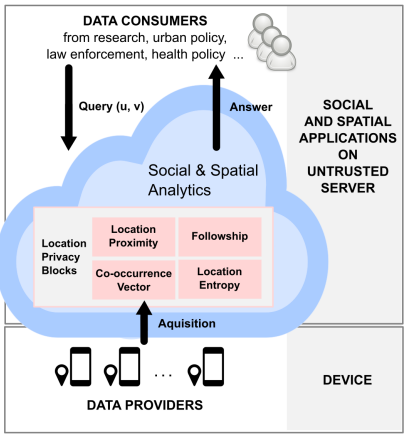

Figure 1: The Vision of PLACE

ited, we can also infer how strong their connection is [15].
3) *Spatial Influence*: If one person "follows" the other through a sequence of places, i.e., visits the same locations shortly after the other person's visits, we can infer he/she is under the influence of the other. Depending on the spatial and temporal properties of the "followship", we can also quantify the amount of influence one exerts on the other.

The three use cases are generic enough to empower many real-world applications including all the applications enabled by online social networks such as marketing applications (e.g., target advertising, recommendation engines such as friendship suggestions), social studies (e.g., identifying influential people) and cultural studies (e.g., to examine the spreading patterns of new ideas, practices and rumors). In addition, they also have their own unique applications due to the geo-spatial properties. For example, the inferred social connections can be used to identify the new (or unknown) members of a criminal gang or a terrorist cell or it can be used in epidemiology to study the spread of diseases through human contacts. The inferred social influence also has several applications, specifically by inducing local influence in real-world applications bounded to a specific location. Examples include healthcare (when we need to inform the residents of a suburb about the outbreak of a contagious disease), in local advertisements (local restaurants, cafes, events), in a local political campaign (selecting a district's representative), or simply disseminating information (ideas, rumors) related to a geographically contained community, e.g., students at a university campus.

Input Data. We define the individually generated data record as $d = (u, l, t)$, where u is a pseudo user identifier, l is an exact location (e.g., Latitude=22.3130, Longitude=114.0406), and t is a time stamp (e.g., April, 2, 2015, 3:20pm). Pseudonyms are used to prevent the disclosure of social relationships between actual user pairs from the intermediate computation results. The real user IDs will only be revealed to a third party upon proof of authorization, e.g., search warranty, court subpoena, etc.

Building Blocks. We design four building blocks in PLACE based on deep understanding of people's social behaviors and will be utilized by various social relationship studies. Our block definitions are generic and they can be built on top of each other. Specifically,

1) *Location Proximity* block tests whether two locations are within close distance. It is a fundamental procedure for computing other blocks. Formally, the proximity test is to return a binary answer (*Yes* or *No*) for the following inequality in the physical space: $dist(l_1, l_2) \leq r$, where l_i is a location and r is the range threshold.
2) *Co-occurrence Vector* block computes the frequency of two individual co-occurring at each place, as defined in [15]. Co-occurrence is an important block to measure Social Strength and Spatial Influence between two people. Formally, the co-occurrence vector between two individuals u and v is defined as $C_{uv} = (c_{uv,1}, ..., c_{uv,m})$, where $c_{uv,l}$ is the frequency of their co-occurrence at location l.
3) *Location Entropy* block computes/maintains the popularity of a location based on how frequently people visit it. It is another important block to measure Social Strength and Spatial Influence. To compute Location Entropy of a given location l as in [15], we need a list of frequencies: $F_l = (f_1, f_2, ...)$, where f_i represents the number of visits to l by user u_i.
4) *Followship* block is a novel concept in Spatial Influence. It computes the co-location between two people at different times, in order to quantify influence. This temporal aspect of followship is measured as the time delay between the visits by u and v at a location. The spatial aspect of followship is the popularity of each location, as in Location Entropy.

Private Computation. We believe the development of the building blocks of PLACE would open up new research challenges in the area of privacy. This is because to protect the privacy of the individual data holders, their location data must be transformed prior to the block computation, in order to prevent the disclosure of *location*, *statistical information* about their location history (against frequency attacks), and associated *timestamps* (against side information attacks) to the untrusted server and other parties. However, the majority of existing approaches provide interface and disclosure control for only one type of information, such as PIR-based protocols for hiding locations and differential-privacy-based approaches for sanitizing statistics. Moreover, encryption-based approaches, such as PIR and PPT protocols, incur prohibitive computation/communication overheads. We need a holistic approach that simultaneously provides strong privacy guarantees for various types of information (location, timestamps, and statistics), and achieves high accuracy and efficiency for computing all the blocks over large-scale spatiotemporal data.

Our vision is to design innovative computation methods to utilize encryption and differential privacy primitives and provide comprehensive privacy protection to the spatiotemporal data collected from individual devices. Firstly, to protect location information, we will design several practical encrypted search schemes with Deterministic Encryption [12] and Probabilistic Encryption [20] for proximity testing, by integrating efficient cryptographic primitives with novel representations of location data and queries, which provides a trade-off between efficiency and privacy. Secondly, to sanitize statistical information, differential privacy primitive is deployed in a distributed setting in addition to encryption. We design a dynamic perturbation model to add/remove individual location records prior to encryption. Furthermore, the error caused by removing location records to non-statistical blocks, e.g., Location Proximity, can be carefully

examined and noise recovery mechanisms will be designed to alleviate the problem. Thirdly, lightweight schemes similar to locality sensitive hashing [11] can be designed to prevent the disclosure of time as well as to improve the computational efficiency. In addition, indexing structures, such as Quadtree, can be incorporated with encryption primitives to enable the untrusted server to create and update the index on top of encrypted data.

4. CONCLUSION

We presented our vision for PLACE, an extensible framework which enables social relationship studies by analyzing individually generated location data. PLACE utilizes an untrusted server and performs location analytics without disclosing location information to the server and other parties. We illustrated three example social relationship studies enabled by PLACE, i.e., *Reachability*, *Social Strength*, and *Spatial Influence*, and presented four novel privacy-preserving building blocks: *Location Proximity*, *Co-Occurrence Vector*, *Location Entropy*, and *Followship* to support our use cases. We proposed to utilize encryption and differential privacy primitives to prevent the disclosure of people's location, statistical information about their location history, and associated timestamps. The private blocks can be utilized across different use cases as well as to define new blocks. The successful realization of PLACE will facilitate private location data acquisition from individual devices, thanks to the strong privacy guarantees, and will enable a wide range of applications in epidemiology, criminology, political science, and etc.

5. REFERENCES

[1] Please rob me. `http://pleaserobme.com/`.

[2] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *CCS*, 2013.

[3] R. Chen, B. C. Fung, B. C. Desai, and N. M. Sossou. Differentially private transit data publication: A case study on the montreal transportation system. In *KDD*, pages 213–221, 2012.

[4] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky. Searchable symmetric encryption: Improved definitions and efficient constructions. In *CCS*, pages 79–88, 2006.

[5] J. S. Damico, J. W. Oller, and J. A. Tetnowski. Investigating the interobserver reliability of a direct observational language assessment technique. *International Journal of Speech-Language Pathology*, 1(2):77–94, 1999.

[6] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, Mar. 2013.

[7] C. Dwork. Differential privacy. In *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. 2006.

[8] B. Gedik and L. Liu. Location privacy in mobile systems: A personalized anonymization model. In *ICDCS*, pages 620–629, June 2005.

[9] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan. Private queries in location based services: Anonymizers are not necessary. In *SIGMOD*, pages 121–132, 2008.

[10] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI*, pages 211–220, 2009.

[11] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *VLDB*, pages 518–529, 1999.

[12] J. Katz and Y. Lindell. *Introduction to Modern Cryptography: Principles and Protocols*. Chapman & Hall/CRC Cryptography and Network Security Series. 2007.

[13] K. Kreijns, P. A. Kirschner, and W. Jochems. Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior*, 19(3):335 – 353, 2003.

[14] A. Narayanan, N. Thiagarajan, M. Lakhani, M. Hamburg, and D. Boneh. Location privacy via private proximity testing. In *NDSS*, 2011.

[15] H. Pham, C. Shahabi, and Y. Liu. Ebm: An entropy-based model to infer social strength from spatiotemporal data. In *SIGMOD*, pages 265–276, 2013.

[16] P. D. Renshaw and S. R. Asher. Children's goals and strategies for social interaction. *Merrill-Palmer Quarterly*, 29(3):pp. 353–374, 1983.

[17] H. Shirani-Mehr, F. Banaei-Kashani, and C. Shahabi. Efficient reachability query evaluation in large spatiotemporal contact datasets. *VLDB*, 5(9):848–859, May 2012.

[18] D. X. Song, D. Wagner, and A. Perrig. Practical techniques for searches on encrypted data. In *IEEE Security and Privacy*, pages 44–55, 2000.

[19] L. Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, Oct. 2002.

[20] B. Wang, M. Li, H. Wang, and H. Li. Circular range search on encrypted spatial data. In *IEEE CNS*, 2015.

[21] Y. Xiao and L. Xiong. Protecting locations with differential privacy under temporal correlations. In *CCS*, 2015.

[22] Y. Xiao, L. Xiong, L. Fan, S. Goryczka, and H. Li. Dpcube: Differentially private histogram release through multidimensional partitioning. *Transactions on Data Privacy*, 7(3):195–222, 2014.

[23] M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu. Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 366–375, April 2008.

[24] Y. Zheng, M. Li, W. Lou, and Y. Hou. Sharp: Private proximity test and secure handshake with cheat-proof location tags. In *Computer Security - ESORICS 2012*, volume 7459 of *Lecture Notes in Computer Science*, pages 361–378. 2012.