Privacy-Preserving Inference of Social Relationships from Location Data

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 Computer Science
Utah State University
ming.li@usu.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. Due to the popularity of GPS-enabled mobile devices or online services, a large amount of high-resolution spatiotemporal 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 meet and some statistical properties about their meeting location, which can be computed in a private manner. In this paper, we envision a novel 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 the building blocks to support various social relationship studies, without disclosing location information to the server and other untrusted parties. We showcase PLACE with three use cases and four novel building blocks and ensure privacy for block computation 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.

1. INTRODUCTION

For decades, anthropologists and social scientists have been studying people's social behaviors by utilizing sparse datasets obtained by observations [3] and interviews [14]. 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) [8, 11]. However, due to the nature of the utilized dataset, 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 [15, 13]. For example, 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 [13].

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. [4] 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 [13], 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 statistics about the frequency of meetings and the popularity of meeting locations, for example by using *differential privacy* on location statistics.

To this end, we envision a novel 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 perform 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 showcased: *Reachability* use case answers the question whether one person can be reached by another through a sequence of pair-wise 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 novel 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. Existing encryption and differential privacy primitives will be utilized to ensure privacy for block computation, as well as innovative unified schemes for dynamic data acquisition.

The successful realization of PLACE will facilitate private location data acquisition from individual devices, thanks to the strong privacy guarantees. Our use cases enable many applications such as Reachability in epidemiology to study the spread of diseases through human contacts, Social-Strength in criminology to identify the new or unknown members of a criminal gang or a terrorist cell, and Spatial-Influence in policy to induce local influence in electing a tribal representative. New use cases can be easily developed under PLACE framework and built on top of existing and/or newly defined blocks.

2. RELATED WORKS

A plethora of works has been developed to protect location privacy. Here we briefly review related works on Location Obfuscation, Private Information Retrieval, Differential Privacy, and Private Proximity Testing.

Some location obfuscation techniques hide the actual location among a set of dummy and send redundant queries to the server [20], while others adopt the concept of k-anonymity [17] and use a Cloak Region (CR) which includes the actual location as well as $k-1$ other users [6]. 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.

Cryptographic approaches based on Private Information Retrieval (PIR) protocols, e.g., in [7] 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 query issuer's privacy, meaning other people's location data are disclosed to the server. Furthermore, they would incur prohibitive cryptographic operations and communication overheads when applied to computing

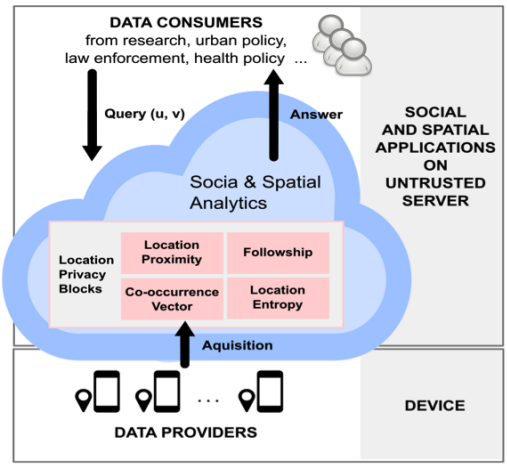

Figure 1: The Vision of PLACE

blocks on large amount of location data [16].

Differential privacy [5] 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 [19] and trajectories [2]. 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.

Private Proximity Testing (PPT) [12, 21] protocols 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 for computation, it is not straightforward to utilize PPT protocols for complex tasks.

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:

- *Reachability*: If two people have come in close contact with each other 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 [15].
- *Social Strength*: If two people have been to the same places at the same time, i.e., *co-occurred*, we can infer that they are socially connected. Depending on the places they visited, we can also infer how strong their social connection is [13].
- *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 person. Depending on the spatial and temporal properties of the "followship", we can also quantify the amount of influence one individual exerts on the other.

The three use cases above 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 its own unique applications due to its geospatial property, 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 input data to PLACE, i.e. individually generated data tuples, as $d = (u, l, t)$ where u is a user/device 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).

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

- *Location Proximity* block tests whether two locations are within close distance, as in [12]. It is fundamental to all three uses cases and is a basic procedure needed 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 represented by its latitude and longitude and r is the range threshold.
- *Co-occurrence Vector* block computes/maintains the frequency of two individual co-occurring at each place, as defined in [13]. Co-occurrence is an important block to measure the 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.
- *Location Entropy* block computes/maintains the popularity of a location based on how frequently people visit it, as in [13]. It is another important block to measure the Social Strength and Spatial Influence between two people. The input to computing the Location Entropy of a given location l is a list of frequencies: $F_l = (f_1, f_2, ...)$, where f_i represents the number of visits to l by user u_i.
- *Followship* block is a novel concept used in Spatial Influence. It computes the co-location between two people at different times, in order to quantify the concept of influence. This temporal aspect of followship is measured as the time delay or the time interval between the visits by u and v at each 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 one type of information, such as PIR-based protocols for hiding location and differential-privacy-based approaches for hiding statistics. Moreover, considering how the raw location data is transformed, current approaches fall near either ends of the spectrum, between flexibility of computation and privacy guarantees. For instance, Location obfuscation approaches can be used to build all of our blocks but provide weak privacy guarantees, while PIR protocols are highly private but incur prohibitive computation/communication overheads for our block computations. Therefore, there is a need for a unified approach which provides privacy guarantee for various types of information and is practical for computing all blocks.

Our vision is to design innovative block computation methods to utilize encryption and differential privacy primitives and provide comprehensive privacy protection to the spatiotemporal data collected from individual devices. To protect locations, encryption-based schemes can be adopted and efficient, private designs with Deterministic Encryption [10] and Probabilistic Encryption [18] can be developed. To sanitize statistical information, differential privacy primitive can be deployed in a distributed setting, in addition to encryption. Lightweight schemes similar to locality sensitive hashing [9] can be designed to prevent the disclosure of time as well as to improve the computational efficiency.

4. CONCLUSION

We presented our vision for PLACE, a novel 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 for block computation. 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. 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] 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 *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 213–221, New York, NY, USA, 2012. ACM.

[3] 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.

[4] 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.

[5] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2006.

[6] B. Gedik and L. Liu. Location privacy in mobile systems: A personalized anonymization model. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 620–629, June 2005.

[7] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan. Private queries in location based services: Anonymizers are not necessary. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 121–132, New York, NY, USA, 2008. ACM.

[8] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 211–220, New York, NY, USA, 2009. ACM.

[9] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

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

[11] 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.

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

[13] H. Pham, C. Shahabi, and Y. Liu. Ebm: An entropy-based model to infer social strength from spatiotemporal data. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 265–276, New York, NY, USA, 2013. ACM.

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

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

[16] R. Sion. On the computational practicality of private information retrieval. In *In Proceedings of the Network and Distributed Systems Security Symposium, 2007. Stony Brook Network Security and Applied Cryptography Lab Tech Report*, 2007.

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

[18] B. Wang, M. Li, H. Wang, and H. Li. Circular range search on encrypted spatial data. 2015.

[19] 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.

[20] 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.

[21] Y. Zheng, M. Li, W. Lou, and Y. Hou. Sharp: Private proximity test and secure handshake with cheat-proof location tags. In S. Foresti, M. Yung, and F. Martinelli, editors, *Computer Security ? ESORICS 2012*, volume 7459 of *Lecture Notes in Computer Science*, pages 361–378. Springer Berlin Heidelberg, 2012.