Privacy in Mobility Data Mining

Aris Gkoulalas-Divanis
Information Analytics Lab
IBM Research–Zurich
Säumerstrasse 4
Rüschlikon, Switzerland
agd@zurich.ibm.com

Yücel Saygin
Faculty of Engineering and
Natural Sciences
Sabanci University
Istanbul, Turkey
ysaygin@sabanciuniv.edu

Dino Pedreschi
Knowledge Discovery and
Delivery Laboratory
University of Pisa
Pisa, Italy
pedre@di.unipi.it

1. INTRODUCTION

As mobile telecommunications and ubiquitous computing pervade our society, the location of mobile users can be continuously sensed and recorded. This has given rise to a set of novel applications that deliver or manage content based on users' location. Examples of such applications are the well-known *location based services* and, more recently, the *location based social networks* (e.g., Foursquare, Gowalla) and the *participatory sensing systems* (e.g., CycleSense). These applications allow users to take dynamic decisions on issues like transportation, visiting of places of interest, or meeting friends in nearby locations. Furthermore, they enable the participation of users in campaigns that involve the collection of data from certain locations (through sensor-equipped mobile devices) and its transmission to remote servers to support different forms of analysis. Apart from on-line applications, the recent advances in database systems have made possible the storage and processing of large amounts of mobility data that can be used to support off-line analysis. All these emerging applications, arising from mobility data, pose unique challenges to the data mining community. Specifically, the mining of personal mobility data collected through these applications can produce reliable knowledge to aid traffic engineers, city managers and environmentalists towards decision making in a wide spectrum of tasks, such as urban planning, sustainable mobility, intelligent transportation, and environmental pollution, thereby enhancing the way we live today.

Although mobility data represents a very useful source of information, the sharing of user location or movement information raises serious privacy concerns that need to be addressed in order for data mining to be more accepted in this domain. On one side, revealing user location to untrustworthy service providers (e.g., for the offering of location based services) may risk privacy attacks, allowing the discovery of personal and sensitive information about the users. On the other side, publishing user trajectories to support analytic or other purposes, may come at a high cost to the privacy of the individuals whose movement is recorded in the data, once sensitive personal information is disclosed to malicious users. Privacy concerns are also important barriers for the adoption of new applications, whose offering depends on user location. As it becomes evident, in both on-line and off-line settings, sophisticated privacy-enhancing solutions are necessary to protect the location and mobility information of the individuals from potential privacy attacks.

This issue is dedicated to the area of privacy preservation in the context of mobility data, which is an emerging field for the KDD community. We invited three survey articles that cover the state-of-the-art on privacy models and data transformation techniques for protecting user privacy in location based services, as well as the state-of-the-art on anonymization methods for privacy-aware trajectory data publishing. In addition, we invited an article that reports on interesting privacy challenges in a new class of emerging applications, known as *participatory sensing*, and proposes a method to address some major privacy leaks in participatory systems. To make this issue of value to the largest possible audience, the invited articles are written at a level that does not require from the readers significant prior knowledge of the area, and contain several figures and examples to explain well the presented concepts.

2. SUMMARY OF THIS ISSUE

The invited papers that are published in this special issue cover different research directions in the area of privacy for mobility data. In what follows, we provide a concise summary of the content that is covered in each article.

2.1 Privacy in Snapshot Location Based Services and Social Networks

The first article, titled "Privacy preservation in the dissemination of location data", is authored by Manolis Terrovitis and provides a survey on the attack scenarios, the offered privacy guarantees, and the data transformation approaches for protecting user privacy in the context of location based services. Location based services (LBS) and location based social networks are popular applications as they enable users to take dynamic, informed decisions on issues like transportation, identification of places of interest, or the opportunity to meet a friend or a colleague in a nearby location. The article presents the three LBS architectures that have been considered for the communication of the users with the LBS providers, as well as elaborates on the different privacy models that have been proposed in the research literature and on the privacy guarantees that they offer. Attacks against user identity, user location, and user query content are discussed and analyzed, with emphasis being placed on the so-called *snapshot* LBS, in which only one (current) location of the user needs to be reported to the LBS provider to allow the offering of the service. Moreover, the article discusses privacy challenges that arise in the context of location based social networks, and reports on recent initiatives to

help users control access to their data. Some open technical issues and promising research directions on privacy preserving location data dissemination, are also given.

2.2 Privacy in Continuous Location Based Services and Trajectory Data Publication

The second article is titled "Trajectory privacy in location based services and data publication" and is authored by Chi-Yin Chow and Mohamed F. Mokbel. It surveys the state-of-the-art approaches that have been proposed for the offering of privacy in the context of *continuous* location based services, and trajectory data publication. Unlike snapshot LBS, in continuous LBS a mobile user has to report her location to the service provider in a periodic or on-demand manner, in order to be offered the requested continuous service. Protecting location privacy in continuous LBS is a very challenging problem because adversaries can use the spatiotemporal correlations in the reported location samples for the users to infer their location with higher certainty and breach their anonymity. Continuous LBS can be partitioned into two categories depending on whether they require consistent user identities to offer the service, or do not require user identities. The authors discuss five families of privacy-preserving methods for continuous LBS, namely *spatial cloaking*, *mix-zones*, *vehicular mix-zones*, *path confusion* and *fake location trajectories*, and present approaches that fall in each family. Following that, they cover two well-studied trajectory anonymization approaches, namely the *clustering-based approach* and the *generalization-based approach*, that have been recently proposed for the offering of privacy in trajectory data publishing. The paper concludes with some interesting directions for future research.

2.3 Trajectory Anonymity in the Publication of Personal Mobility Data

The third invited article is titled "Trajectory anonymity in publishing personal mobility data" and is authored by Francesco Bonchi, Laks V. S. Lakshmanan and Hui (Wendy) Wang. The authors survey the state-of-the-art developments in anonymity-preserving data publishing of personal mobility data. Unlike the case of location based services and participatory sensing systems, where privacy has to be offered in an on-line, potentially dynamic, and service-centric manner, in anonymous personal mobility data publishing the goal is to construct and publish a dataset that effectively maintains most of the utility of the original mobility data, while it effectively protects the personal and sensitive information of the users from potential attackers. The approaches that have been proposed so far assume that a trusted entity collects trajectories representing the movement for a large number of users. This data has to be shared with a set of potentially untrustworthy entities for data analysis purposes, e.g. to enable traffic optimization research. To accomplish this goal, the trusted entity has to anonymize the mobility data so that no privacy breach, based on the assumed models of adversarial attacks, may occur when the data is published. The authors categorize adversary knowledge into two types, namely *location based knowledge*, and *mobility pattern-based knowledge*. For each category, they explain the considered privacy model and elaborate on the anonymization methodologies that have been proposed to offer user privacy. Last, they discuss some interesting open research directions in this popular research area.

2.4 Privacy in Participatory Sensing Systems

The fourth invited article is titled "A privacy-aware framework for participatory sensing" and is authored by Leyla Kazemi and Cyrus Shahabi. It reports on some interesting privacy challenges that arise in the context of a new class of emerging applications, known as *participatory sensing*, in which people aim to contribute content (e.g., images, video), collected by their mobile devices, to central data servers. This content can be used at a later point to support analytic tasks or other types of data processing. Participatory sensing systems can be either *unsolicited* (e.g., Flickr or Youtube), where users participate by arbitrarily collecting data, or *campaign-based* (e.g., CycleSense), when a coordinated effort of the participants is necessary to collect the data that is needed by the data server to support some purpose (e.g., collect traffic information). The authors formally define the problem of **P**rivacy-**A**ware **P**articipatory **A**ssignment (**PAPA**) in participatory sensing systems and introduce a privacy-aware framework that enables the participation of users to these services without compromising their location privacy. Although there exist several techniques for the offering of location privacy in conventional LBS, there are certain unique characteristics of participatory campaigns that constitute these approaches inapplicable. The authors propose a method, called **P**artical-**i**nclusivity and **R**ange **i**ndependence – **PiRi**, which solves the **PAPA** problem and is experimentally verified to be efficient.

3. CONCLUSION

The invited articles in this issue cover the research in the two principal directions of privacy for mobility data: user privacy in location based services, and anonymity in the publication of personal mobility data. Furthermore, they report on privacy issues in emerging applications, such as location based social networks and participatory sensing systems. Applications arising from mobility data pose unique challenges to the data mining community. Data mining techniques, when applied on data collected by location based applications, have a lot of potential in supporting decision making in tasks such as urban planning, intelligent transportation, and environmental pollution. However, privacy-enhancing methods are necessary to ensure that the collected data are protected against privacy threats. With this special issue we wanted to bring the state-of-the-art on privacy issues in this newly emerging domain to the attention of the KDD community.

Acknowledgement

We would like to thank the authors of the four invited articles for their hard work in covering the state-of-the-art research in this area. We also wish to express our deep gratitude to the Editorial Board of SIGKDD Explorations for giving us the opportunity to organize this special issue. Last, but not least, we want to acknowledge the support of the MODAP (Mobility, Data Mining, and Privacy) project in our research. MODAP is a Coordination Action type project that is funded by the EU (2009-2012). It continues the efforts of the GeoPKDD (Geographic Privacy-aware Knowledge Discovery and Delivery, `www.geopkdd.eu`) project by coordinating and boosting the research activities in the intersection of mobility, data mining, and privacy. Detailed information about MODAP can be found at `www.modap.org`.