Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Protecting private geosocial networks against practical hybrid attacks with heterogeneous information

Yuechuan Li [a], Yidong Li [a,*], Guandong Xu [b]

[a] School of Computer and Information Technology, Beijing Jiaotog University, China
[b] Advanced Analytics Institute, University of Technology Sydney, Australia

ARTICLE INFO

Article history:
Received 14 April 2015
Received in revised form
21 July 2015
Accepted 31 August 2015

Keywords:
Geosocial network
Privacy preservation
Anonymization
Location-based attack

ABSTRACT

GeoSocial Networks (GSNs) are becoming increasingly popular due to its power in providing high-performance and flexible service capabilities. More and more Internet users have accepted this innovative service model. However, even GSNs have great business value for data analysis by integrated with location information, it may seriously compromise users' privacy in publishing the GSN data. In this paper, we study the identity disclosure problem in publishing GSN data. We first discuss the attack problem by considering both the location-based and structure-based properties, as background knowledge, and then formalize two general models, named (k, m)– *anonymity* and (k, m, l)– *anonymity* . Then we propose a complete solution to achieve (k, m)– *anonymization* and (k, m, l)– *anonymization* to prevent the released data from the above attacks above. We also take data utility into consideration by defining specific information loss metrics. It is validated by real-world data that the proposed methods can prevent GSN dataset from the attacks while retaining good utility.

1. Introduction

A GeoSocial Network (GSN) is essentially a social network with the capability of geocoding or geotagging services. It is becoming more and more popular in the era of mobile Internet, as the geographic information can be easily obtained by using mobile communication base stations, wireless hotspots and even high-precision GPS sensors. With these geographic data, service providers can offer a large number of location-based services, and users can have better and customized service experience. Therefore, from the well-known social networks (e.g. Facebook and Twitter) to the location-based services (e.g. Foursquare), more and more users have no any concerns to share their locations by checking in themselves through various geosocial mobile applications.

However, publishing such GSN datasets may seriously compromise users' privacy. A well-known breaching is identity disclosure (ID) with the so-called background knowledge attack, in which the attacker knows or collects some information related to the target, and uses such information to uniquely identify the target in the published dataset.

The ID problem has been studied extensively on graphs [2,12,3,4,6,9,7]. However, it is insufficient to preserve privacy in GSNs since users' location information is much more distinctive. For example, a typical GSN dataset can be seen in Table 1, which is composed of two text documents, one for social network and the other for check-in records. If we know that Alice has visited location f3bb9560a2532e in August 2008, we can observe that User 1 matches condition. If no other users match condition in the whole dataset, we can predicate the identity of Alice. Furthermore, the adversary can significantly increase the attack efficiency by combining structure information and location information. In our sample, if the adversary knows Alice has been to location f3bb9560a2532e in 2008, but does not know which month. Then both User 0 and User 1 match condition. If the adversary also knows that Alice's good friend Bob visited location ddaa40aaa22411 in 2010, then the identity of Alice is disclosed with high probability. Although being easy to obtain, such property of Alice is pretty specific. Few pieces of information can successfully re-identify the target.

There exists research work regarding the ID problem in GSNs. The work in [5] introduces a top location representation for geosocial network datasets and two notions of k – *anonymity* for GSN datasets. The quasi-identifier is the top m locations of each user as well as its friends. They assume that the adversary knows all the top locations of the victim and its friend. We argue that this is not a realistic assumption, because the attacker usually cannot know all of the target's whereabouts. Furthermore, they cannot collect the location information of all of the target's friends. Under such strong assumptions, while preserving users' privacy, we may make the data useless. In this paper, we propose a more realistic

* Corresponding author.
E-mail addresses: 13120412@bjtu.edu.cn (Y. Li), ydli@bjtu.edu.cn (Y. Li), Guandong.Xu@uts.edu.au (G. Xu).

http://dx.doi.org/10.1016/j.neucom.2015.08.132
0925-2312/

Table 1
An example of a GSN dataset.

(a) An example of a social network

User ID	User ID
0	1
0	2
1	0
1	2
2	0
2	1
⋮	⋮

(b) An example of check-in records

User ID	Check-in time	Latitude	Longitude	Location ID
0	2008-12-03T21:09:14Z	39.633321	−105.317215	ee8b88dea22411
0	2008-11-30T22:30:12Z	39.633321	−105.317215	ee8b88dea22411
0	2008-11-20T17:55:04Z	41.295474	−95.999814	f3bb9560a2532e
1	2008-08-14T21:23:55Z	41.257924	−95.938081	4c2af967eb5df8
1	2008-08-14T06:54:21Z	41.295474	−95.999814	f3bb9560a2532e
2	2010-04-06T06:45:19Z	46.521389	14.854444	ddaa40aaa22411
⋮	⋮	⋮	⋮	⋮

assumption that the attacker knows at most m locations of the target and at most l of its friends. Then, we introduce the concepts of location (k, m)– *anonymity* and (k, m, l)– *anonymity*, which are relaxations of $\mathcal{L}_k$ – *anonymity* and $\mathcal{L}_k^2$ – *anonymity* in [5]. Stokes [8] presents a similar definition named (k, l)– *anonymity* which is applied on tabular data and be interpreted in terms of combinatorial set systems. In GSN, data format is different from tabular data where top locations and social links are extracted as attributes of users. Since the background knowledge is heterogeneous, the anonymizing algorithms in [8] cannot be directly utilized.

The main contributions of our work are:

1. We propose a hybrid attack model considering structure-based attack and location-based attack in geosocial networks, where only partial information is known to the adversary.
2. We propose a hypergraph-based anonymization model to prevent identity disclosure.
3. We develop a complete solution based on the proposed anonymization model.
4. We consider various data utilities and propose algorithms to enhance these utilities.

The remainder of this paper is organized as follows. Section 2 presents a brief survey of identity disclosure on SN and GSN. Section 3 introduces the top location model and data preprocessing. In Section 4, we provide formal definitions of the attack model and anonymity model. We also define several practical utility metrics to measure information loss in the anonymization process. We propose a full solution to the problem in Section 5. The experimental results are shown in Section 6. Finally, Section 7 concludes this paper.

2. Related work

The identity disclosure problem has been studied extensively on graphs [2,12,3]. Zheleva and Getoor [11] considered the problem of protecting sensitive relationships among the individuals in the anonymized social networks. The work in [10] studies how anonymization algorithms that are based on randomly adding and removing edges change certain graph properties. The work in [6] provides a method that perturbs the structure of a social graph in order to provide link privacy, at the cost of slight reduction in the utility of the social graph. Liu et al. [4] first take weights as consideration for privacy preserving in social networks. The work in [9] extends the above work by formulating an abstract model based on linear programming. However, the objective of their work still focuses on maintaining certain linear property of a social network by reassigning edge weights. The authors in [7] study mechanisms for enhancing privacy while using social-network routing by obfuscating the friends lists used to inform routing decisions.

Amirreza and James [5] introduce a top location representation for geosocial network datasets and two notions of k – *anonymity* for GSN datasets, i.e., $\mathcal{L}_k$ – *anonymity* and $\mathcal{L}_k^2$ – *anonymity*. The quasi-identifier is the top m locations of each user as well as its friends. They develop two separate anonymization algorithms to achieve $\mathcal{L}_k$ – *anonymity* and $\mathcal{L}_k^2$ – *anonymity*. They assume that the adversary knows all the top locations of the victim and its friend. We argue that this is not a realistic assumption, because the attacker usually cannot know all of the target's whereabouts. Furthermore, they cannot collect the location information of all of the target's friends. Under such strong assumptions, while preserving users' privacy, we may make the data useless. A more realistic assumption is that the attacker knows at most m locations of the target and at most n of its friends. Under these assumptions, we introduce the concepts of location (k, m)– *anonymity* and (k, m, l)– *anonymity*, which are relaxations of $\mathcal{L}_k$ – *anonymity* and $\mathcal{L}_k^2$ – *anonymity*.

Compared to k – *anonymity*, the limitations on background knowledge bring about more computational complexity. In [8], the concept of (k, l)– *anonymity* on tabular data is introduced, and is interpreted in terms of combinatorial set systems. They show that (k, l)– *anonymity* can be represented in the form of hypergraph for the sake of algorithm design. However, the hypergraph representation is still suitable to use to simplify anonymity models. We observe that the top locations can be viewed as a set of attributes without metadata. Unlike the top-m model in [5], the number of top locations is not fixed for each user, since there is a large variance in the histogram of check-in location. Also, in the anonymization process, the number of top locations usually decreases during generalization, which needs special considerations.

3. Top location model

In the real-world scenarios, a location table such as Table 1 (b) in our example may have a large number of small aggregation values. That is, an entity may visit a place only once so that the location becomes meaningless for statistical analysis. To preserve privacy, one approach is to do some aggregation on the original datasets. The aggregation strategies involve counting the number of check-in at the same location and extracting the most representative locations for each user. Although they lead to information loss, the data still has research or business value. In this section, we introduce two selection metrics and discuss several location extraction strategies.

3.1. Top location extracting

Our aim is to select some locations from all the visited locations for each user. The simplest scheme is to select the most visited location as representatives. Let $c(v, l)$ be the number of reports of user v in location l, and L_{tf} be the set of top frequent locations (i.e., most visited locations). For any user v, we have $\forall\, l_1 \in L_{tf}\ \ l_2 \notin L_{tf}$, $c(v, l_1) \geq c(v, l_2)$. It is obvious that such data can be used for business such as marketing and resource allocation.

Alternatively, we can extract the most unique location for each

Please cite this article as: Y. Li, et al., Protecting private geosocial networks against practical hybrid attacks with heterogeneous information, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.08.132