A New Privacy Breach: User Trajectory Recovery From Aggregated Mobility Data

Abstract:

Human mobility data have been ubiquitously collected through cellular networks and mobile applications, and publicly released for academic research and commercial purposes for the last decade. Since releasing individual’s mobility records usually gives rise to privacy issues, data sets owners tend to only publish aggregated mobility data, such as the number of users covered by a cellular tower at a specific timestamp, which is believed to be sufficient for preserving users’ privacy. However, in this paper, we argue and prove that even publishing aggregated mobility data could lead to privacy breach in individuals’ trajectories. We develop an attack system that is able to exploit the uniqueness and regularity of human mobility to recover individual’s trajectories from the aggregated mobility data without any prior knowledge. By conducting experiments on two real-world data sets collected from both the mobile application and cellular network, we reveal that the attack system is able to recover users’ trajectories with an accuracy of about 73%*\_*91% at the scale of thousands to ten thousands of mobile users, which indicates severe privacy leakage in such data sets. Our extensive analysis also reveals that by generalization and perturbation, this kind of privacy leakage can only be mitigated. Through the investigation on aggregated mobility data, this paper recognizes a novel privacy problem in publishing statistic data, which appeals for immediate attentions from both the academy and industry. Aggregated data filtered through various method like user availability during day time, evening and throughout the day.

**EXISTING SYSTEM**

The existing system does not Hungarian algorithm which is very efficient analyzing the Aggregated data provided from the application side and cellular operator side.No individual information can be directly acquired from the datasets.Since the aggregated mobility dataonly contains general information of the population,we cannot directly distinguish each mobile user and extract personal information. It makes publishingaggregated mobility data automatically complies with*k*-*anonymity* privacy model, which prevents the*re*-*identif ication* attack

**DISADVANTAGE**

The procedure of publishing aggregated mobility datasets does not require generalizing, suppressing or permuting original records. Therefore, it preserves the truthfulness and accuracy of the original datasets at record level. Such accurate aggregated mobility data is of great importance in numerous applications, ranging from transportation scheduling to business locating. It will be very difficult separate the data from each segments provided by the operators.

**PROPOSED SYSTEM**

To overcome the limitations and disadvantages of existing system , We use two real-world mobility datasets to understand and investigate how can we recover user’s trajectory from the aggregated mobility data.Dataset Collected by a Mobile Application*,*This dataset is collected from mobile devices by a popular mobile application.It records the mobile users’ spatiotemporal points when it is activated for service interactions. The dataset traces over 15,000 mobile users from November 1st to 14th, 2015. It records fine-grained spatiotemporal information of mobile users, including anonymized user identification, accessed base stations and timestamp.Dataset Collected by a Cellular Operator*,* This dataset

is collected by a major mobile network operator in China. It is a large-scale dataset including 100,000 mobile users with the duration of one week, between April 1st and 7th, 2016. It records the spatiotemporal information of mobile subscribers when they access cellular network (i.e., making phone calls, sending texts, or consuming data plan).

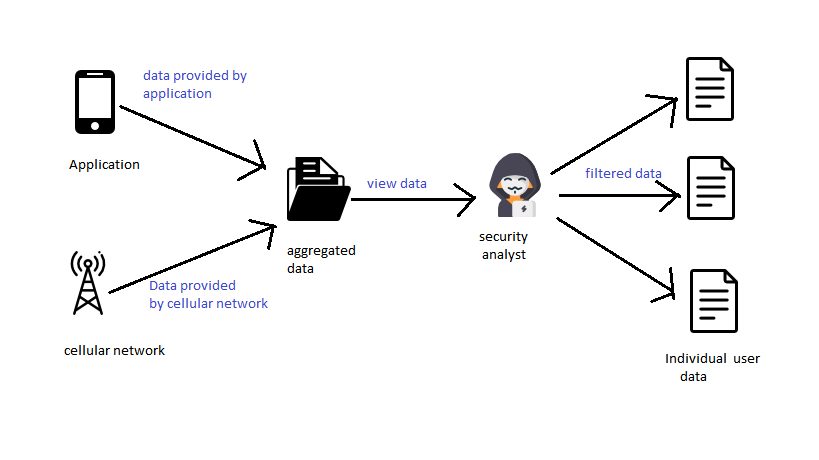
**ADVANTAGE**

The advantages that ,our attack system is effective in breaching the privacy of aggregated mobility data in terms of recovering mobile users’ trajectories with high recovery accuracy and low recovery error. In addition, the recovered trajectories have a high possibility to be linked to the victims with external information provided, such as two most frequent locations. These results suggest that the privacy leakage is surprisingly severe even in publishing aggregated mobility data, which contradicts the conventional wisdom and appeals attention to investigate privacy problem in such dataset.

**ALGORITHM**

Hungarian Method is for assigning jobs by a one-for-one matching to identify the lowest-cost solution. Each job must be assigned to only one machine. It is assumed that every machine is capable of handling every job, and that the costs or values associated with each assignment combination are known and fixed. The number of rows and columns must be the same. Trial and error works well enough for this problem, but suppose you had ten salespeople flying to ten cities? How many trials would this take? There are n! ways of assigning n resources to n tasks. That means that as n gets large, we have too many trials to consider.

**Architecture**



**MODULES:**

* Mobile application
* Cellular data
* Security analyst
* Trajectory recovery

* Mobile application:

This dataset is collected from mobile devices by a popular mobile application.It records the mobile users’ spatiotemporal points when it is activated for service interactions. The dataset traces over 15,000 mobile users from November 1st to 14th, 2015. It records fine-grained spatiotemporal information of mobile users, including anonymized user identification, accessed base stations and timestamp. the user location of each record is represented by already known GPS coordinates of the accessed base station, which usually covers its neighboring area within hundreds of meters.

* Cellular data:

This dataset is collected by a major mobile network operator in China. It is a large-scale dataset including 100,000 mobile users with the duration of one week, between April 1st and 7th, 2016. It records the spatiotemporal information of mobile subscribers when they access cellular network (i.e., making phone calls, sending texts, or consuming data plan). It also contains anonymous user identification, accessed base stations and timestamp of each access. The user locations are also represented by GPS coordinates of the accessed base stations. the empirical Cumulative Distribution Function (CDF) of interval time between two consecutive records and the CDF of the number of records

* Security analyst

The security analyst is the person who tries breach the datas got from cellular level and application level. The flow of data trajectory from the flow trajectory of night,day and whole day is calculated and list each user visiting the particular place is calculated and will be different to the each user. Data such as credit cards ,sms ,location of place visited is calculated. Security analyst process data and calculated for each and every time when it is updated.

* Trajectory recovery :

We design an unsupervised framework that leverages the universal characteristics of human mobility to recover users’ trajectories from aggregated mobility data without any prior knowledge. Our framework includes three modules: nighttime, daytime, and cross-day trajectory recovery. The key insight of nighttimetrajectory recovery is that mobile users tend to stay in fixed locations during nighttime for natural sleeping cycle. the percentage of users with different number of visited locations during nighttime, i.e., 0 am*∼*6 am. From the results.we can observe that 62% and 88% of mobile users only visit one base station during nighttime in operator and application datasets, respectively. Different from nighttime, users move frequently during daytime, which requires a different scheme to estimate the next locations. The key insight is the continuity of human mobility, which enables the next location estimation by using the current location and velocity

**Future work**

The current data is calculated as user visited locations, cellular data level, Application data level may be in future it will be calculated as users call data with others, users browsing level etc.

**Conclusions**

In this paper, we identify and evaluate the risks of trajectory recovery attack in the aggregated mobility dataset. To the best of our knowledge, we are the first to recognize and study the privacy problem of inferring individual’s information from statistic data. Our investigation reveals that there is serious privacy leakage in the aggregated mobility data since individuals’ trajectories can be recovered with high accuracy. In addition, our evaluation demonstrates that the spatiotemporal resolution and scale of datasets have notable impact on the privacy breach. We believe that this work opens a new angle of protecting the privacy in publishing and sharing statistic data. We believe it will pave the way to more advanced privacy preserving mechanisms.