Revisiting Utility Metrics for Location Privacy-Preserving Mechanisms

Virat Shejwalkar, Amir Houmansadr, Hossein Pishro-Nik, Dennis Goeckel
University of Massachusetts Amherst
Location Based Services (LBSes)

- Ride Hailing Services
  - Uber
  - Lyft
  - Waze

- Social Networks
  - Facebook
  - WeChat
  - WhatsApp

- Recommendation Systems
  - TripAdvisor
  - Zomato
  - Yelp

- Online Gaming
  - Parallel Kingdom
  - Pokémon GO
  - Zombie Run

- Goods Delivery
  - Walmart Eats
  - Uber Eats
Privacy Issues
Privacy Issues

Media

Location, location, location: ACCC sues Google over collection and use of users' location data

Lyft Sued For Violating California Privacy Law

by Wendy Davis @wendyndavis, February 12, 2014

Australian regulator files privacy suit against Google alleging location data misuse

Google agrees to pay $13 million in Street View privacy case
Privacy Issues

Media

Location, location, location: ACCC sues Google over collection and use of users’ location data

Lyft Sued For Violating California Privacy Law
by Wendy Davis @wendyndavis, February 12, 2014

Australian regulator files privacy suit against Google alleging location data misuse

Google agrees to pay $13 million in Street View privacy case

Academia

De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice

Huangdong Wang1, Chen Gao1*, Yong Li1, Gang Wang1, Depeng Jin2, and Jingbo Sun3
1Department of Electronic Engineering, Tsinghua University
2Department of Computer Science, Virginia Tech

walk2friends: Inferring Social Links from Mobility Profiles

Michael Hackes
CERI, Stanford University

Knock Knock, Who’s There? Membership Inference on Aggregate Location Data

Apostolos Pyrgitis
University College London
apostolos.pyrgitis.14@ucl.ac.uk

Carmelo Troncoso
IMDEA Software Institute
carmelo.troncoso@imdea.org

Emiliano De Cristofaro
University College London
c.d.cristofaro@ucl.ac.uk

Abstract—Aggregate location data is often used to support smart services and applications, e.g., generating live traffic maps or predicting visit to businesses. In this paper, we present the first study on the feasibility of membership inference attacks on aggregate location time-series. We introduce a game-theoretic definition of the adversarial task, and cast it as a classification problem where machine learning can be used to distinguish whether or not a target user is part of the aggregates. We empirically evaluate the power of these attacks on both raw and differentially private aggregates using two mobility datasets. We find that membership inference is a serious privacy threat, and show how its effectiveness depends on the adversary’s prior knowledge, the characteristics of the underlying location data as well as the number of users and the timeframe on which aggregation is performed. Although differentially private mechanisms can indeed reduce the extent of the attacks, they also yield a significant loss in utility. Moreover, a strategic adversary mimicking the behavior of the defense mechanism can greatly limit the protections they provide. Overall, our work presents a novel methodology geared to evaluate membership inference on aggregate location data in real-world settings and can be used by researchers, policy makers, and practitioners to better understand the consequences before data release or by regulators to detect violations.

1. INTRODUCTION

Motivation. The ability of an adversary to ascertain the presence of an individual in aggregate location time-series constitutes an obvious privacy threat if the aggregate relates to a group of users that share a sensitive characteristic. For instance, learning that an individual is part of a dataset aggregating movements of Alzheimer’s patients implies learning that she suffers from the disease. Similarly, inferring that statistics collected over a sensitive timeframe or sensitive locations include a particular user also harms the individual’s privacy.

Recent work [22, 34] also shows that an adversary with some prior knowledge about a user’s mobility profile can exploit aggregate information to improve this knowledge, or even localize her. Also, users’ “trajectories” can in some cases be extracted from aggregate mobility data, even without prior knowledge [30]. However, in order to mount these attacks, the adversary needs to know that the user is part of the aggregate dataset, which further motivates our research objectives.

Membership inference can also be leveraged by providers
Location Privacy Preserving Mechanisms
Location Privacy Preserving Mechanisms

Real location $l_r$ \rightarrow LPPM($l_r$) \rightarrow Obfuscated location $l_o$
Location Privacy Preserving Mechanisms

LPPM($l_r$)

- Geo-indistinguishability
  - Andres et al. CCS’13

- Game theoretic defenses
  - Shokri et al. CCS’12

- Cloaking, precision reduction
  - Micinski et al. arxiv’13

- Remapping based utility improvements
  - Chatzikokolakis et al. PETS’17

$l_r$ Real location

$l_o$ Obfuscated location
Location Privacy Preserving Mechanisms

LPPM($l_r$)

Geo-indistinguishability
Andres et al. CCS’13

Game theoretic defenses
Shokri et al. CCS’12

Cloaking, precision reduction
Micinski et al. arxiv’13

Remapping based utility improvements
Chatzikokolakis et al. PETS’17

$l_r$  
Real location

→

$l_o$  
Obfuscated location
Location Privacy Preserving Mechanisms

LPPM($l_r$)

- Geo-indistinguishability
  - Andres et al. CCS’13

- Game theoretic defenses
  - Shokri et al. CCS’12

- Cloaking, precision reduction
  - Micinski et al. arxiv’13

- Remapping based utility improvements
  - Chatzikokolakis et al. PETS’17

$l_r$ → LPPM($l_r$) → $l_o$

Real location

Obfuscated location

Analyzed only using generic utility metrics!
E.g. Squared/Euclidean distance
Our contributions

• **Generic utility metrics are insufficient** to analyze utility perceived in real-world LBSes

• Design and analysis of LPPMs using **generic metrics do not provide privacy-utility trade-offs perceived** in real-world LBSes

• We build and **open-source RHSE**, a RHS data synthesis tool for location privacy research
Why generic metrics are bad
Why generic metrics are bad
Why generic metrics are bad

\[
d(l_r, l_r^1) = d(l_r, l_r^2) \Rightarrow \text{In theory, generic utility metrics conclude that both locations are equally useful}
\]
Why generic metrics are bad

In theory, generic utility metrics conclude that both locations are equally useful.

But, in practice, multiple parameters of LBSes play part in constituting utility of its users.
Tailored Utility Metrics in Location Based Services

- Time to reach destination, distance travelled
- Calories burnt
- Accuracy of weather forecast

Maps application

Weather forecast application

Fitness application
Problem Statement

How do generic utility metrics perform in practice?
Problem Statement

How do generic utility metrics perform in practice?

We investigate this question using ride hailing services (RHSes) and geo-indistinguishable mechanisms.
Challenges

- Existing LBSes do not incorporate LPPMs
- Multiple possible tailored utility metrics
- Commercial LBSes do not release their location data
Challenges

Existing LBSes do not incorporate LPPMs

Multiple possible tailored utility metrics

Commercial LBSes do not release their location data
PRide – Privacy Preserving RHS

PRide ensures privacy of riders from RHS provider
PRide – Privacy Preserving RHS

PRide ensures privacy of riders from RHS provider

• Rider’s pick-up and destination locations are obfuscated
PRide – Privacy Preserving RHS

- Rider’s pick-up and destination locations are obfuscated
- Drivers’ locations used for ride matching are obfuscated
- Drivers only report pick-up and destination locations of a ride, which are already obfuscated

PRide ensures privacy of riders from RHS provider
PRide – Privacy Preserving RHS

PRide ensures privacy of riders from RHS provider

RHS provider’s view
Challenges

- Existing LBSes do not incorporate LPPMs
- **Multiple possible tailored utility metrics**
- Commercial LBSes do not release their location data
PRide: Tailored Utility Metric

Tailored utility metric for riders

Time to complete ride

Tailored utility loss for riders

Time to complete ride with $l_o$ - Time to complete ride with $l_r$
Challenges

Existing LBSes do not incorporate LPPMs

Multiple possible tailored utility metrics

Commercial LBSes do not release their location data
Ride Hailing Service Emulator (RHSE)

To emulate RHS data

- Number of riders and drivers
- Distribution of riders and drivers
- Behavior of drivers
- Geo-indistinguishability level
- LPPM
- Geographical region
Ride Hailing Service Emulator (RHSE)

RHSE can be tuned to various common RHS scenarios

- Number of riders and drivers
- Distribution of riders and drivers
  - Behavior of drivers
  - Geo-indistinguishability level
  - LPPM
  - Geographical region
Ride Hailing Service Emulator (RHSE)

RHSE can be tuned to various common RHS scenarios

- Number of riders and drivers
- Distribution of riders and drivers
- Behavior of drivers
- Geo-indistinguishability level
- LPPM
- Geographical region
Ride Hailing Service Emulator (RHSE)

RHSE can be tuned to various common RHS scenarios

- Number of riders and drivers
- Distribution of riders and drivers
- Behavior of drivers
- Geo-indistinguishability level
- LPPM
- Geographical region
Ride Hailing Service Emulator (RHSE)

RHSE can be tuned to various common RHS scenarios

- Number of riders and drivers
- Distribution of riders and drivers
- Behavior of drivers
- Geo-indistinguishability level
- LPPM
- Geographical region
LPPM Evaluation

ETA tolerance of drivers
LPPM Evaluation

ETA tolerance of drivers

Generic QL fails to account for variations in dynamic parameters of RHS
LPPM Evaluation

ETA tolerance of drivers

Generic QL fails to account for variations in dynamic parameters of RHS

Tailored QL varies with changes in such parameters of RHSes
LPPM Comparison: Planar Laplace versus Exponential Mechanism

Generic and tailored metrics do not agree on the superiority of one mechanism over the other.
LBS-aware LPPM Design Choices Using Tailored Utility Metrics

- Privacy budget $\epsilon$ is an important design parameter
- Consider a rider with
  - $QL_{\text{generic}}^+ = 2000$ meters
  - $QL_{\text{tailored}}^+ = 1000$ seconds
- Consider two RHS settings:
  - Drivers’ ETA$_t$ is 900 seconds
  - Drivers’ ETA$_t$ is 400 seconds
Implications to Utility Improvement Techniques

• Consider non-uniformly distributed drivers

• **Greedy remapping** remaps obfuscated locations to the area with high driver density
Implications to Utility Improvement Techniques

• Consider non-uniformly distributed drivers

• **Greedy remapping** remaps obfuscated locations to the area with high driver density

![Planar Laplace mechanism](image)

- Tailored QL (seconds), Generic QL (meters)
- **with remap** vs. **without remap**
- Privacy radius $r$ (Km)
Implications to Utility Improvement Techniques

- Consider non-uniformly distributed drivers
- **Greedy remapping** remaps obfuscated locations to the area with high driver density

![Planar Laplace mechanism](image)

Greedy remapping may not improve generic loss
Implications to Utility Improvement Techniques

• Consider non-uniformly distributed drivers

• **Greedy remapping** remaps obfuscated locations to the area with high driver density

---

**Planar Laplace mechanism**

Greedy remapping may not improve generic loss

Greedy remapping always improves tailored loss
• **Conclusion**
  - Generic utility metrics do not represent the utility perceived by LBS users
  - This leads to misleading analysis and design of LPPMs

• **Recommendations**
  - Devise utility metrics tailored to specific LBSes
  - Design application aware LPPMs to truly optimize for privacy-utility tradeoffs

Please try our RHSE ([Link here](#)) and give feedback!!