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

NDSS 2018

Apostolos Pyrgelis [1], Carmela Troncoso [2] and Emiliano De Cristofaro [1]

[1]UCL, [2]EPFL

February 20, 2018
San Diego, CA, USA

Introduction

Introduction

- Location data enable mobility analytics in the context of smart cities

Introduction

- Location data enable mobility analytics in the context of smart cities
- But, they are very **privacy sensitive**

Introduction

- Location data enable mobility analytics in the context of smart cities
- But, they are very **privacy sensitive**
- Analysts use *aggregate* location statistics
    - e.g., Uber Movement or Telefonica Smart Steps

Introduction

- Location data enable mobility analytics in the context of smart cities
- But, they are very **privacy sensitive**
- Analysts use *aggregate* location statistics
  - e.g., Uber Movement or Telefonica Smart Steps
- Recent works (**PETS'17, WWW'17**) show that aggregate location statistics might violate the privacy of individuals that are part of the aggregates

Introduction

- Location data enable mobility analytics in the context of smart cities
- But, they are very **privacy sensitive**
- Analysts use *aggregate* location statistics
  - e.g., Uber Movement or Telefonica Smart Steps
- Recent works (**PETS'17, WWW'17**) show that aggregate location statistics might violate the privacy of individuals that are part of the aggregates
- We focus on **membership inference** attacks
  - i.e., an adversary attempts to determine whether or not location data of a target user is part of the aggregates

Motivation

Motivation

- Membership inference is a first step to other types of attacks on location aggregates, e.g., **profiling** or **localization**

Motivation

- Membership inference is a first step to other types of attacks on location aggregates, e.g., **profiling** or **localization**
- Aggregates might be collected over sensitive locations / time-frame, or might relate to a group of users that share a sensitive characteristic

Motivation

- Membership inference is a first step to other types of attacks on location aggregates, e.g., **profiling** or **localization**
- Aggregates might be collected over sensitive locations / time-frame, or might relate to a group of users that share a sensitive characteristic
- Regulators can verify possible misuse of the data, e.g., when aggregate location data has been released without permission

In this work...

In this work...

- We reason about membership inference in the context of location data

In this work...

- We reason about membership inference in the context of location data
- We model the problem as a *game* in which an adversary aims at distinguishing location aggregates that include data of a target user from those that do not

In this work...

- We reason about membership inference in the context of location data
- We model the problem as a *game* in which an adversary aims at distinguishing location aggregates that include data of a target user from those that do not
- We instantiate the distinguishing task with a machine learning classifier trained on the *adversarial prior knowledge* and use it to infer membership in *unseen* aggregate statistics

Main Findings

Main Findings

- We deploy membership inference attacks on two real-world mobility datasets and find that releasing **raw** aggregates poses a significant privacy threat

Main Findings

- We deploy membership inference attacks on two real-world mobility datasets and find that releasing **raw** aggregates poses a significant privacy threat
- We evaluate the privacy protection of defense mechanisms that guarantee **differential privacy** and show how they are effective at preventing inference at the cost of utility

Distinguishability Game

Distinguishability Game

Distinguishability Game

Adversary	Challenger
Pick u	

Distinguishability Game

Adversary | **Challenger**

Pick u

$$\xrightarrow{\quad u \quad}$$

Distinguishability Game

Adversary		**Challenger**
Pick u		
	$\xrightarrow{\quad u \quad}$	
		$b \leftarrow_{\$} \{0, 1\}$

Distinguishability Game

Adversary		Challenger
Pick u		
	$\xrightarrow{u}$	
		$b \leftarrow_{\$} \{0, 1\}$
		If b == 0:
		$loc_u \in A$
		Else:
		$loc_u \notin A$

Distinguishability Game

Adversary		**Challenger**
Pick u		
	$\xrightarrow{\quad u \quad}$	
		$b \leftarrow_{\$} \{0, 1\}$
		If b == 0:
		$loc_u \in A$
		Else:
		$loc_u \notin A$
	$\xleftarrow{\quad A \quad}$	

Distinguishability Game

Adversary		Challenger
Pick u		
	$\xrightarrow{\quad u \quad}$	
		$b \leftarrow_{\$} \{0,1\}$
		If b $==$ 0:
		$loc_u \in A$
		Else:
		$loc_u \notin A$
	$\xleftarrow{\quad A \quad}$	
$b' \in \{0,1\}$		

Distinguishing Function

Distinguishing Function

- **Intuition :** Membership inference can be modeled as a *binary classification* task
  - i.e., was the target's data used to calculate the aggregate location time-series under examination?

Distinguishing Function

- **Intuition :** Membership inference can be modeled as a *binary classification* task
  - i.e., was the target's data used to calculate the aggregate location time-series under examination?
- We utilize a *supervised* machine learning classifier trained on data that is included in the **adversarial prior knowledge**

Adversarial Prior Knowledge

Adversarial Prior Knowledge

- **Subset of Locations :** The adversary knows the real locations for a subset of users that includes her target
  - e.g., a telecommunications provider

Adversarial Prior Knowledge

- **Subset of Locations :** The adversary knows the real locations for a subset of users that includes her target
  - e.g., a telecommunications provider
- **Participation in Past Groups :** The adversary knows the target's participation for location aggregate time-series observed in the past

Adversarial Prior Knowledge

- **Subset of Locations :** The adversary knows the real locations for a subset of users that includes her target
  - e.g., a telecommunications provider
- **Participation in Past Groups :** The adversary knows the target's participation for location aggregate time-series observed in the past
  - **Same Groups as Released :** continuous data release over stable groups

Adversarial Prior Knowledge

- **Subset of Locations :** The adversary knows the real locations for a subset of users that includes her target
  - e.g., a telecommunications provider
- **Participation in Past Groups :** The adversary knows the target's participation for location aggregate time-series observed in the past
  - **Same Groups as Released :** continuous data release over stable groups
  - **Different Groups than Released :** continuous data release over dynamic user groups

Privacy Loss

Privacy Loss

- For a target, we play the distinguishability game multiple times

Privacy Loss

- For a target, we play the distinguishability game multiple times
- **Privacy Loss :** The adversary's advantage in winning it over a random guess

Privacy Loss

- For a target, we play the distinguishability game multiple times
- **Privacy Loss :** The adversary's advantage in winning it over a random guess
- We utilize the Area Under Curve (AUC) score to evaluate the classifier's performance

Datasets

Tranport For London (TFL):

- 60M trips - 4M unique oyster cards - 582 stations (regions of interest - ROIs)
- Monday, March 1 - Sunday, March 28, 2010
- Sample the top 10K oyster ids per total # of trips, being active for 115 ± 21 out of the 672 timeslots and reporting 171 ± 26 ROIs in total (sparse, regular)

Datasets

Tranport For London (TFL):

- 60M trips - 4M unique oyster cards - 582 stations (regions of interest - ROIs)
- Monday, March 1 - Sunday, March 28, 2010
- Sample the top 10K oyster ids per total # of trips, being active for 115 ± 21 out of the 672 timeslots and reporting 171 ± 26 ROIs in total (sparse, regular)

San Francisco Cabs (SFC):

- 11M GPS coordinates - 534 cabs in SF - May 19 to June 8, 2008
- Grid $10 \times 10 = 100$ ROIs of 0.5×0.37 mi^2
- Taxis are active for 340 ± 94 out of the 504 timeslots and report $3,663 \pm 1,116$ ROIs in total (dense, irregular)

Experimental Setup

Experimental Setup

- **Target Users :** For each dataset, we *randomly* pick 50 users from 3 mobility groups (highly, mildly, somewhat) and run membership inference attacks

Experimental Setup

- **Target Users :** For each dataset, we *randomly* pick 50 users from 3 mobility groups (highly, mildly, somewhat) and run membership inference attacks
- **Sample & Aggregate :** Sample groups that include and exclude the target user to create a *balanced* dataset of *labeled* aggregate location time-series

Experimental Setup

- **Target Users :** For each dataset, we *randomly* pick 50 users from 3 mobility groups (highly, mildly, somewhat) and run membership inference attacks
- **Sample & Aggregate :** Sample groups that include and exclude the target user to create a *balanced* dataset of *labeled* aggregate location time-series
- **Feature Extraction :** Extract various statistics from the time-series of each ROI
    - i.e., mean, variance, std, median, min, max, sum

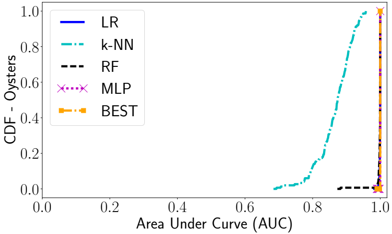
Experimental Setup

- **Target Users :** For each dataset, we *randomly* pick 50 users from 3 mobility groups (highly, mildly, somewhat) and run membership inference attacks
- **Sample & Aggregate :** Sample groups that include and exclude the target user to create a *balanced* dataset of *labeled* aggregate location time-series
- **Feature Extraction :** Extract various statistics from the time-series of each ROI
  - i.e., mean, variance, std, median, min, max, sum
- **Classification :** Train and test the classifier

Evaluating Raw Aggregates

Evaluating Raw Aggregates

TFL



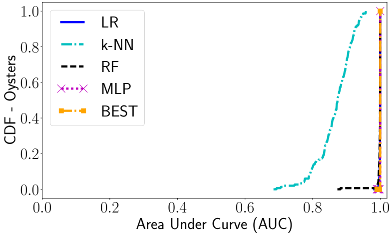
Prior : Same Groups As Released

Group Size : 1,000

Inference Period : 1 Week

Evaluating Raw Aggregates

TFL

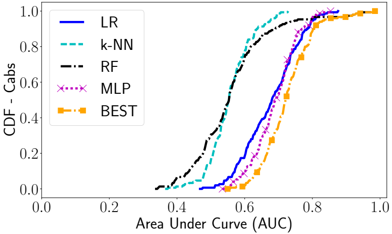


Prior : Same Groups As Released

Group Size : 1,000

Inference Period : 1 Week

SFC



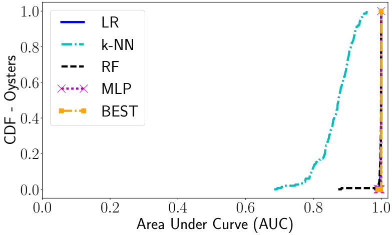
Prior : Subset of Locations

Group Size : 100

Inference Period : 1 Week

Evaluating Raw Aggregates

TFL

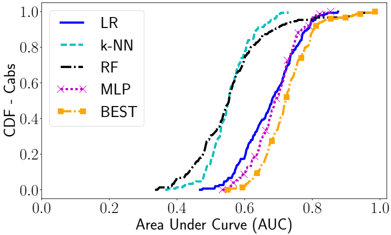


Prior : Same Groups As Released

Group Size : 1,000

Inference Period : 1 Week

SFC



Prior : Subset of Locations

Group Size : 100

Inference Period : 1 Week

* More experimental results in the paper

Take Aways

Take Aways

- Membership inference is **successful** when the adversary knows the locations of a subset of users or the past aggregates for the same groups on which she performs inference

Take Aways

- Membership inference is **successful** when the adversary knows the locations of a subset of users or the past aggregates for the same groups on which she performs inference
- Privacy leakage on the commuter dataset (TFL) is higher compared to the cab one (SFC)

Take Aways

- Membership inference is **successful** when the adversary knows the locations of a subset of users or the past aggregates for the same groups on which she performs inference
- Privacy leakage on the commuter dataset (TFL) is higher compared to the cab one (SFC)
- Users enjoy more privacy on larger groups

Take Aways

- Membership inference is **successful** when the adversary knows the locations of a subset of users or the past aggregates for the same groups on which she performs inference
- Privacy leakage on the commuter dataset (TFL) is higher compared to the cab one (SFC)
- Users enjoy more privacy on larger groups
- Inference is easier if aggregates of longer periods are released and at times when mobility patterns are more regular

Evaluating Differentially Private (DP) Mechanisms

Evaluating Differentially Private (DP) Mechanisms

- We choose a worst-case adversary that obtains *perfect* prior knowledge for the users
  - i.e., given *raw* aggregates she can train a classifier that achieves AUC score of 1.0

Evaluating Differentially Private (DP) Mechanisms

- We choose a worst-case adversary that obtains *perfect* prior knowledge for the users
  - i.e., given *raw* aggregates she can train a classifier that achieves AUC score of 1.0
- We modify the game, so that the challenger applies a DP mechanism before sending her challenge to the adversary
  - LPA, GSM, FPA, EFPAG

Evaluating Differentially Private (DP) Mechanisms

- We choose a worst-case adversary that obtains *perfect* prior knowledge for the users
    - i.e., given *raw* aggregates she can train a classifier that achieves AUC score of 1.0
- We modify the game, so that the challenger applies a DP mechanism before sending her challenge to the adversary
    - LPA, GSM, FPA, EFPAG
- We evaluate the privacy protection offered by DP mechanisms against an adversary that trains the classifier on:

Evaluating Differentially Private (DP) Mechanisms

- We choose a worst-case adversary that obtains *perfect* prior knowledge for the users
  - i.e., given *raw* aggregates she can train a classifier that achieves AUC score of 1.0
- We modify the game, so that the challenger applies a DP mechanism before sending her challenge to the adversary
  - LPA, GSM, FPA, EFPAG
- We evaluate the privacy protection offered by DP mechanisms against an adversary that trains the classifier on:
  - **raw** aggregates

- We choose a worst-case adversary that obtains *perfect* prior knowledge for the users
  - i.e., given *raw* aggregates she can train a classifier that achieves AUC score of 1.0
- We modify the game, so that the challenger applies a DP mechanism before sending her challenge to the adversary
  - LPA, GSM, FPA, EFPAG
- We evaluate the privacy protection offered by DP mechanisms against an adversary that trains the classifier on:
  - **raw** aggregates
  - **noisy** aggregates using the defense mechanism under examination

Privacy vs. Utility

Privacy vs. Utility

- **Privacy Gain :** The relative decrease in the adversary's performance when challenged on *perturbed* aggregates vs. *raw* aggregates

Privacy vs. Utility

- **Privacy Gain :** The relative decrease in the adversary's performance when challenged on *perturbed* aggregates vs. *raw* aggregates
- **Utility :** Mean Relative Error (MRE)

Experimental Results - TFL - Group Size: 9,500

Utility (MRE):

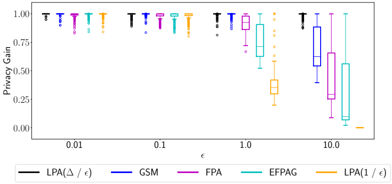
ϵ	**0.01**	**0.1**	**1.0**	**10**
LPA(Δ/ϵ)	3056.1	812.6	81.7	8.2
GSM	753.2	75.8	7.4	0.75
FPA	67.2	6.1	0.7	0.03
EFPAG	36.8	3.6	0.4	0.03
LPA(1 / ϵ)	38.5	3.7	0.3	0.002

Utility (MRE):

ϵ	0.01	0.1	1.0	10
LPA(Δ / ϵ)	3056.1	812.6	81.7	8.2
GSM	753.2	75.8	7.4	0.75
FPA	67.2	6.1	0.7	0.03
EFPAG	36.8	3.6	0.4	0.03
LPA($1 / \epsilon$)	38.5	3.7	0.3	0.002

Privacy Gain :

Train on *Raw* Aggregates

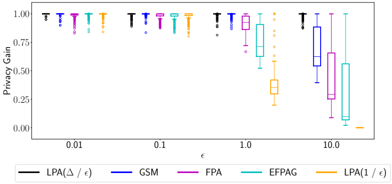


Utility (MRE):

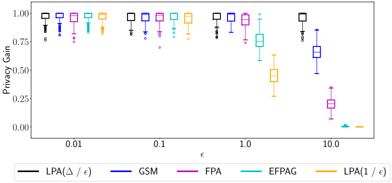
ϵ	0.01	0.1	1.0	10
LPA(Δ/ϵ)	3056.1	812.6	81.7	8.2
GSM	753.2	75.8	7.4	0.75
FPA	67.2	6.1	0.7	0.03
EFPAG	36.8	3.6	0.4	0.03
LPA($1/\epsilon$)	38.5	3.7	0.3	0.002

Privacy Gain :

Train on *Raw* Aggregates



Train on *Noisy* Aggregates



Take Aways

Take Aways

- DP mechanisms are overall **successful** at preventing membership inference

Take Aways

- DP mechanisms are overall **successful** at preventing membership inference
- But, with significant reduction in the utility of the aggregates

Take Aways

- DP mechanisms are overall **successful** at preventing membership inference
- But, with significant reduction in the utility of the aggregates
- A strategic adversary that *mimics* the behavior of the defender can reduce the privacy gain offered by a mechanism

Take Aways

- DP mechanisms are overall **successful** at preventing membership inference
- But, with significant reduction in the utility of the aggregates
- A strategic adversary that *mimics* the behavior of the defender can reduce the privacy gain offered by a mechanism
- Mechanisms specifically designed for time-series settings (e.g., FPA) achieve better utility

Take Aways

- DP mechanisms are overall **successful** at preventing membership inference
- But, with significant reduction in the utility of the aggregates
- A strategic adversary that *mimics* the behavior of the defender can reduce the privacy gain offered by a mechanism
- Mechanisms specifically designed for time-series settings (e.g., FPA) achieve better utility
- Our methods can be used to evaluate defense mechanisms!

In Conclusion

- We propose a **methodology** geared to evaluate membership inference on aggregate location data
- We define the adversarial task as a *distinguishability game* and use machine learning classification to achieve it
- We quantify the inference power with different kinds of prior knowledge and on datasets with different characteristics and show that **raw** aggregates leak information about user membership
- We utilize our methods to evaluate the privacy protection provided by mechanisms that guarantee **differential privacy** and find that they prevent membership inference but with significant cost in utility

Future Work

- Evaluate membership inference attacks on other location (and not only) datasets
- Examine the mobility characteristics of users that are affected by the attack more than others
- Obtain insights about the design of defenses with better utility

The end...

The end...

Thanks for your attention! Any questions?

The end...

Thanks for your attention! Any questions?

Contact Details: **apostolos.pyrgelis.14@ucl.ac.uk**