Location Privacy Preservation of Vehicle Data in Internet of Vehicles

Ying Ying Liu, Austin Cooke, Parimala Thulasiraman
Correspondence: umliu369@myumanitoba.ca

The Twelfth International Conference on Evolving Internet
INTERNET 2020
October 18, 2020 to October 22, 2020 - Porto, Portugal

Presenters' Bios

Ying Ying Liu is a PhD student from the InterDisciplinary Evolving Algorithmic Sicence (IDEAS) lab of University of Manitoba, Canada. She joined the IDEAS lab as an undergraduate honors student in 2013. Her strengths are in Computational Intelligence, Internet of Things, High Performance Computing, and Distributed Algorithms.

Austin Cooke received his BSc Hons. Computer Science from the University of Manitoba in 2019. His background was in Algorithms, Privacy, Security and Graph Theory. Since completing his degree, he has been working as a software development consultant with Online Business Systems in Winnipeg, Manitoba, Canada.

Agenda

- Introduction
- Motivation
- Background: Privacy Techniques
- Overview of privacy techniques and Proposed attacks, with solutions
- Experiment
- Evaluation and Analysis of Results
- Conclusion and Future Work

Agenda

Introduction

- Background
  - A branch from the Internet of Things (IoT) network.
  - Evolution of traditional Vehicular Ad Hoc Networks (VANETs) with new enabling technologies such as Cloud and 5G.
  - Data exchange is involved in IoV.
- A simple form of IoV data may include an ID, a timestamp, and the location of a vehicle.
- **In this paper, we focus on protecting the identity of individuals being revealed from sharing location data in IoV applications**.

Challenges for IoV Location Privacy

- Location data is unevenly distributed. Applying state-of-the-art privacy protection such as differential privacy protection to each single point will overwhelm sparse locations with noise.
- Preservation of location patterns and traceability is important for IoV applications.
- Design an efficient data structure to represent IoV location data due to high velocity and volume.

Related Work

- Traditional location privacy research in VANETs: replace the true vehicle ID with a pseudonym.
  - Pro's: anonymity and trace-ability
  - Con's: storage burden for preloading the digital certificates to vehicles [Raya and Hubaux, 2007], relies on Tamper-Proof Device [Wang et al., 2016], or Trace Authority (TRA) [Zhong et al., 2019]
- "Geo-indistinguishability" of fitness tracking social network
  - Pro's: protects large locations [Bates et al., 2018] and popular locations [Yin et al., 2018].
  - **Con's: does not protect sensitive locations accessed by less unique users**.
- Lack of consensus on the definition of location privacy.
- There are few holistic views of location privacy breaches and mitigation at different stages of an IoV application.

Our Contribution

1. We examine potential attacks of location privacy for IoV traffic condition service.
2. We provide a novel birds eye view of existing location privacy preserving techniques and provide a scheme of evaluating these techniques for IoV traffic condition service.
3. We show that instead of locations that are accessed frequently, the locations with **less unique visitors** are extremely sensitive. We use k-d tree data structure to aggregate locations into groups and apply Differential Privacy (DP) to protect sensitive locations. We show that our strategy produces differentially private data, good preservation of utility by achieving similar regression accuracy to the original dataset on an Long Term Short Term Memory (LSTM) neural network traffic predictor.

Motivation

Model Setting and Problem Statement

- Setting: Storage of IoV data in Cloud.
- Data composition: ID, timestamp, location.
- Cloud application: Provide traffic condition services with three stages:
  1. **Traffic update** from vehicle to cloud.
  2. **Traffic data storage** in cloud.
  3. **Traffic query** from vehicles or third party to cloud.

Attack 1: Simple UserID **background attack**.

Assumption: Cloud uses the same ID for each user. Adversary does not know a user's ID but has background information about the user.
Scenario: Each day Officer Tom checks in at a military base that only he has access to. The adversary happens to know this, as well as the location of the military base. The adversary finds out Tom's user ID by this query: "SELECT * FROM DB WHERE UserID = (SELECT UserID FROM DB WHERE location = X)", where X is the location of the military base. Now the adversary can learn the location of Officer Tom even when he is not at the military base.

Potential Attacks of Location Privacy for IoV Traffic Condition Service

Attack 2: Dynamic UserID **background attack**.

Assumption: Cloud uses different ID's for each user. Adversary still has background information about the user as in Attack 1.
Scenario: The adversary runs the query: "SELECT count(*) FROM DB WHERE location = X", where X is the location of Officer Tom's military base. The Cloud will return a value, 0 or 1 indicating whether Officer Tom is there or not at the current time.

Potential Attacks of Location Privacy for IoV Traffic Condition Service

Attack 3: **Untrusted** Cloud attack.

Assumption: Cloud is the adversary and users are innocent. The Cloud contains traffic conditions of various locations that a user may be interested in but does not have this particular user's location.
Scenario: If Tom queries the Cloud regarding a particular location, then the Cloud can infer that Tom may be interested in this location. If the Cloud's method for identifying individual users are unclear, the Cloud can still determine which locations are popular and which are not based on the number of queries about a particular location.
Note: This attack is less likely to happen in practice, but it is important to be considered.

Agenda

Background: Privacy Techniques

Differential Privacy [Dwork, 2006]

- Formal definition of privacy: A randomized function K gives ϵ - differential privacy if for all datasets D and D' differing on at most one row, and all $S \subseteq Range(K), Pr[K(D) \in S] \leq exp(\epsilon) \times Pr[K(D') \in S]$.
- Offers a framework to develop privacy solutions:
  1. A is an algorithm used to compute traffic flow at locations
  2. D is the database with Tom's record
  3. D' is the database without Tom's record
  4. Adds a random noise to the answer of A
  5. Make D indistinguishable from D' by a factor of ϵ
- Constrained to aggregate data analysis

Private Information Retrieval [Chor et al., 1997]

Technique to protect the users that query the database.

1. User's query is encrypted and given to the database
2. Database runs computation on encrypted query
3. An encrypted result is returned to the user
4. Decryption is done on the user's side

Garbled Circuit [Yao, 1986]

Garbled circuit provides an environment for secure (and therefore private) computation between two parties, where the receiving party (evaluator) is only able to perform computation on the encrypted result of the sending party's (garbler's) message.

1. The garbler takes input values to a gate and encrypt them
2. The garbler performs the gate operation on the input values prior to encryption to obtain the output value
3. Each encrypted input is paired with the corresponding output and the value is stored together in the re-arranged truth table
4. The evaluator receives the garbled gate from the garbler
5. The evaluator completes decryption of exactly one ciphertext from the garbled truth table through a process called "oblivious transfer"

Agenda

Overview of privacy techniques and Proposed attacks, with solutions

Location Privacy Metrics in IoV for Traffic Condition Service

Privacy Concerns	*Dynamic Pseudonym*	*Differential Privacy*	*Private Information Retrival*	*Trusted Agency + Garbled Circuit*
Location Privacy at Traffic Update	✓			✓
Location Privacy at Traffic Storage	✓	✓✓		✓✓
Location Privacy at Traffic Query	✓		✓✓	✓✓

The number of ticks represents the effectiveness of a technique for a particular privacy concern.

Location Privacy Parties in IoV for Traffic Condition Service

Parties	*Dynamic Pseudonym*	*Differential Privacy*	*Private Information Retrival*	*Trusted Agency + Garbled Circuit*
Third Party Agency	Trusted	N/A	N/A	Trusted
Cloud	Not Trusted	Trusted	Not Trusted	Not Trusted
Vehicle	Trusted	Not Trusted	Trusted	Trusted

Dynamic Pseudonyms

- The Cloud is the adversary and the vehicle and Trusted Agency (TA) are trusted.
- A user's real identifier is mapped to a list of pseudonyms that change at a predetermined time.
- Only the TA is able to determine the real identity of the mapped pseudonyms.
- Susceptible to Attack 2: Dynamic UserID background attack.

Differential Privacy

- The Cloud is trusted but the vehicle is not.
- The user will attempt to gain information about other users using seemingly harmless queries.
- Strong privacy at traffic storage: Noise is added to the database rows to maintain data privacy and utility at the same time.
- No location privacy protection for traffic update and query:
  1. Vehicles are constantly checking in their locations to the Cloud with updates. This breaks Differential Privacy if the Cloud is not dynamically updating its records.
  2. If an adversarial user queries the Cloud regarding traffic information in various locations they may be able to obtain a picture of what the general traffic concentration appears to be.

Private Information Retrieval

- The Cloud is the adversary and the vehicle is trusted.
- The Cloud has no way of knowing what the vehicles are querying because the queries are encrypted.
- If a vehicle chooses to update the Cloud at any point, its exact location will be revealed to the adversary.

Garbled Circuit With Trusted Agency

- This technique attempts to satisfy each metric that we are evaluating with but it is also most complex.
- The Cloud is the adversary, and the vehicles and TA are trusted.
- Location privacy at update and storage: the Cloud only receives encrypted data to store and cannot directly decrypt this without some assistance from the TA, which does not expose the location of the vehicle without the vehicle's permission.
- Location privacy at query: on a query about traffic related to a certain location, the Cloud is not aware of the value that is being requested for.

Experiment

Experiment Objective

We investigate Differential Privacy to centrally stored location data using the Gowalla location-based social network check-in data.

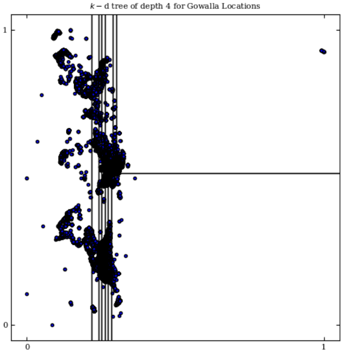
UserID	Timestamp	Lat	Long	LocationID
0	2010-10-19T23:55:27Z	30.2359091	-97.79514	22847
0	2010-10-18T22:17:43Z	30.269103	-97.749395	420315
0	2010-10-17T23:42:03Z	30.255731	-97.763386	316637
0	2010-10-17T19:26:05Z	30.2634181	-97.757597	16516
0	2010-10-16T18:50:42Z	30.2742919	-97.740523	5535878
0	2010-10-12T23:58:03Z	30.2615994	-97.758581	15372
0	2010-10-12T22:02:11Z	30.2679096	-97.749312	21714
0	2010-10-12T19:44:40Z	30.269103	-97.749395	420315
0	2010-10-12T15:57:20Z	30.2811204	-97.745211	153505
0	2010-10-12T15:19:03Z	30.269103	-97.749395	420315
0	2010-10-12T00:21:28Z	40.6438845	-73.782806	23261
0	2010-10-11T20:21:20Z	40.7413743	-73.988105	16907
0	2010-10-11T20:20:42Z	40.7413882	-73.989455	12973
0	2010-10-11T00:06:30Z	40.7249103	-73.994621	341255
0	2010-10-10T22:00:37Z	40.7297683	-73.998535	260957
0	2010-10-10T21:17:14Z	40.7285271	-73.996868	1933724
0	2010-10-10T17:47:04Z	40.7417467	-73.993421	105068
0	2010-10-09T23:51:10Z	40.7341934	-74.004164	34817
0	2010-10-09T22:27:07Z	40.7425116	-74.006031	27836
0	2010-10-09T21:39:26Z	40.7423962	-74.007543	15079
0	2010-10-09T21:36:05Z	40.7423962	-74.007543	15079
0	2010-10-09T21:05:23Z	40.7358847	-74.004968	22806
0	2010-10-09T20:55:47Z	40.7275254	-73.985399	1365909
0	2010-10-09T01:37:03Z	40.75688	-73.986225	11844
0	2010-10-08T21:48:37Z	40.7074172	-74.011363	11742
0	2010-10-08T21:45:48Z	40.7071727	-74.010545	19822
0	2010-10-08T21:43:52Z	40.7070708	-74.011953	15169
0	2010-10-08T21:43:02Z	40.7058231	-73.996696	11794
0	2010-10-08T19:28:36Z	40.769378	-73.963083	1567837
0	2010-10-08T17:24:27Z	40.7808055	-73.976473	35513

Although the dataset is not strictly IoV data, it shares similarity with IoV data by having location, timestamp, and ID in each row. The data is preprocessed to be more suitable for the experiment.

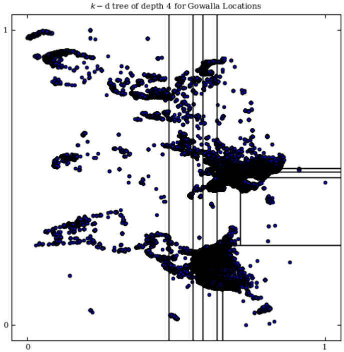
Experiment Objective

We generalize individual locations to location groups by splitting the geometric plane using $k - d$ tree such that each group has roughly the same amount of locations. Each row of the aggregated data includes timestamp, location group, and unique count of users. We then apply Laplace noise to the user count to achieve $\epsilon-$Differential Privacy for the location data.

Data Cleaning



Original Gowalla Locations Including Outliers

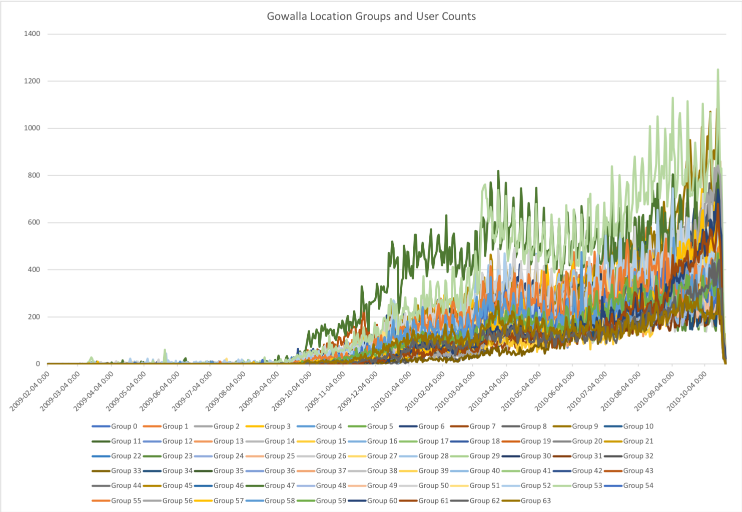


Normalized Gowalla Locations Without Outliers

Building Contingency Table

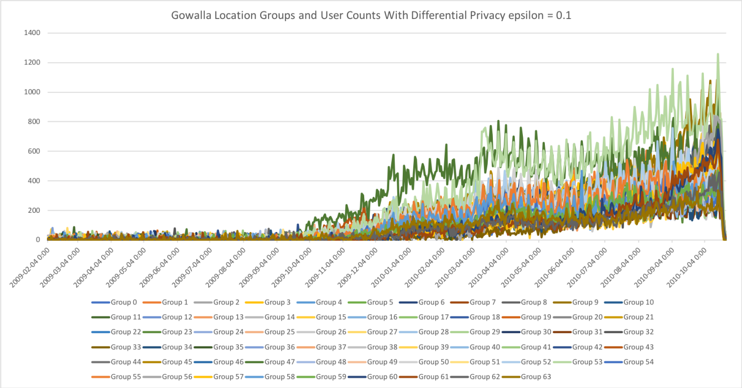
- In order to prepare a differentially private dataset for sharing and publishing, it is important to make sure a contingency table is built on top of the original generalized data and before Differential Privacy is applied.
- For our data, building a contingency table means to create continuous dates for each location group and unique user combination.
- We calculate the minimum and maximum dates in the dataset, and add missing dates to all location groups with user count set to 0.

Generalization: Gowalla Location Groups and User Counts ($k-d$ tree depth $=6$



Differential Privacy: Gowalla Location Groups and User Counts with Differential Privacy ($k-d$ tree depth $=6$, $\epsilon=0.1$)

From the first glance, this dataset shares similar distributions as the original dataset. At a closer look, we can notice the noise added to each location group.



Evaluation and Analysis of Results

Data Utility Evaluation

- In order to quantify the utility of our differentially private dataset, we measure and compare the regression accuracy of a traffic predictor when it is trained by the original dataset and the differentially private dataset.
- We use an LSTM traffic predictor utilized in [Fu et al., 2016] and train two models using 2009-02-04 to 2010-08-31 of the original and differentially private datasets as training data respectively, and then we use the 2010-09-01 to 2010-10-23 of the original dataset as test data. The model is trained with a sliding window of 7 (representing one week) and iteration of 600.

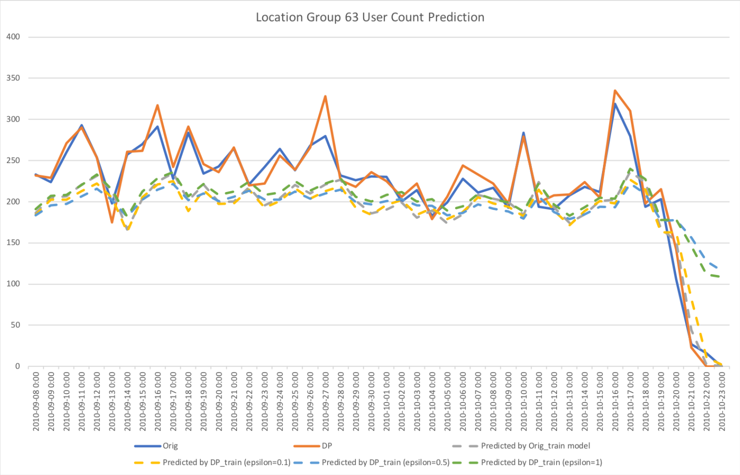
Location Group 63 Prediction Comparison of Different Training Models

After successfully training our predictors, we measure the regression accuracy of the predictors in terms of explained variance score, Root Mean Squared Error (RMSE) and R2 score using the metrics package of Python scikit-learn. The following table shows the comparison of predictions made by models trained by different versions of location data for Gowalla location group 63. We observe that the predictor trained with $0.1-$differentially private data has very close accuracy to the model trained with original data.

Measurement	*Orig model*	*DP* $\epsilon = 0.1$	*DP* $\epsilon = 0.5$	*DP* $\epsilon = 1.0$
Explained variance score	**0.713**	**0.670**	0.390	0.469
RMSE	**45.676**	**48.893**	56.583	50.343
R2 score	**0.513**	**0.442**	0.253	0.409

Location Group 63 Real Data vs. Prediction

In general, the predicted data by all DP-data-trained models are reasonable compared to the real data.



Conclusion and Future Work

Conclusion

- We conduct a thorough study of location privacy in IoV traffic condition service through investigation of potential attacks and mitigation.
- Based on this knowledge, we develop a novel overview of location privacy preservation scheme.
- For experiment, we develop a Differential Privacy strategy to centrally store location data and demonstrate the preservation of data utility quantitatively.

Future Work

- Many avenues are open for research on the techniques we have proposed here.
- Private Information Retrieval (PIR) can be studied much more extensively to determine its overall effectiveness and to examine whether there is another variant of PIR or some existing technique coupled with PIR to satisfy location privacy using the three metrics designed in this section.
- Conducting some experiments on the Trusted Agency (TA) and Garbled Circuit technique could also be an important step to implement a robust location privacy preserving technique as it provides the most utility and the most privacy of all models observed in this paper.

Thank You!