Location Privacy Preservation of Vehicle Data in Internet of Vehicles

Ying Ying Liu
Department of Computer Science
University of Manitoba
Winnipeg, Canada
Email: umliu369@myumanitoba.ca

Austin Cooke
Online Business Systems
Winnipeg, Canada
Email: austin.cooke12@gmail.com

Parimala Thulasiraman
Department of Computer Science
University of Manitoba
Winnipeg, Canada
Email: thulasir@cs.umanitoba.ca

Abstract—**Internet of Things (IoT) has attracted a recent spark in research on Internet of Vehicles (IoV). In this paper, we focus on one research area in IoV: preserving location privacy of vehicle data. We discuss existing location privacy preserving techniques and provide a scheme for evaluating these techniques under IoV traffic condition. We propose a different strategy in applying Differential Privacy using k-d tree data structure to preserve location privacy and experiment on real world Gowalla data set. 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.**

Keywords–Internet of Things; Internet of Vehicles; Location Privacy; Differential Privacy; Privacy Preservation Scheme.

I. INTRODUCTION

In recent years, a new networking concept has emerged. From the growing number of devices that are connected to each other by various means, researchers have coined a term for this network: the Internet of Things. The Internet of Things (IoT) has exploded in the last decade, facilitating the arrival of other novel ideas such as “Big Data”, and many derivatives have spawned from IoT’s core philosophy which involves a globally connected society. One of these derivatives involves facilitating the arrival of automated vehicles. This network specifically deals with vehicles communicating with each other, their infrastructure and other connected devices to form a cohesive, safe environment for automated vehicles to thrive in. This derivative is appropriately named the Internet of Vehicles (IoV). IoV is an evolution of traditional Vehicular Ad Hoc Networks (VANETs) with new enabling technologies such as Cloud and 5G [1]. Of course, in order to provide the necessary support for a network of automated vehicles, some data needs to be exchanged. Such data may include the location of a vehicle, an ID, and a timestamp. Unfortunately, however, the integrity of the data could be threatened by malicious individuals or companies.

In this paper, we focus on protecting the identity of individuals being revealed from sharing location data in IoV applications. There are several reasons why location privacy is challenging in IoT and IoV:

- Compared to relational data, location data imposes additional challenges in adding privacy protection with a balance between privacy and utility. For example, location data from wearable sensors that record an individual’s trajectory has an uneven geometric distribution. Downtown areas may have dense trajectory, whereas, suburbs may have sparse trajectory. Applying state-of-the-art privacy protection, such as differential privacy protection to each single point will greatly affect data utility because the sparsely distributed location data will be overwhelmed with noise. This challenge remains true in IoV, where certain areas have heavier traffic and certain areas have lighter traffic.
- The utility of location data is very important for many IoT applications. For example, in location-based social networks, the preservation of location patterns (combinations of locations) are important for the analysis of protected data. In IoV, traceability poses an even higher standard on the utility of location data.
- Due to the high velocity and volume of location data from sensors, it is very challenging to design an efficient data structure to represent location data in both IoT and IoV.

II. RELATED WORK

In the traditional location privacy research in VANETs, a large amount of work have concentrated on the use of pseudonyms to achieve anonymity and trace-ability at the same time. Raya and Hubaux [2] propose a Privacy-Preserving Authentication (PPA) scheme based on traditional Public Key Infrastructure (PKI) that uses traditional digital signature techniques to authenticate messages. However, this scheme is not scalable as it adds both a huge storage burden to the vehicles for preloading the digital certificates and a burden to communication bandwidth by including the digital certificates in the message. Wang et al. [3] introduce a Two-Factor LIghtweight Privacy-preserving (2FLIP) authentication scheme by using Message-Authentication-Code (MAC) and hash operations. 2FLIP is the first authentication scheme that achieves both strong privacy preservation and DoS resilience, however, it relies on the assumption of additional available devices. Each vehicle is bonded to a telematics device with biometric technology to verify the identities of multiple drivers and to provide evidence to trace each driver. A Tamper-Proof Device (TPD) is embedded in an On-Board-Unit (OBU) to store the system key and to sign and verify messages. Zhong et al. [4] propose a privacy-preserving scheme using a certificate-less aggregate signature to achieve secure Vehicle to Infrastructure (V2I) communications. The authors use a Trace Authority (TRA) to generate pseudonyms and track the real identity during the communication to achieve trace-ability. The computation cost is reduced through pre-calculation at the Road Side Unit (RSU).

ISBN: 978-1-61208-796-2

With regards to location data in IoT, Bates et al. [5] explore some ideas regarding privacy protection in a fitness tracking social network using location fuzzing to introduce "geo-indistinguishability". However, this only protects large locations and not the single location scenario that we consider here. In a recent paper [6], the authors propose an algorithm LPT-DP-k for location privacy protection of location access count data. The algorithm first constructs a Location Privacy Tree (LPT) to preserve relationships among location patterns (i.e., trajectories of locations). It then selects k location patterns with probabilities based on access frequency for data sampling. In the last step, Laplace mechanism for differential privacy protection is applied to the selected sample patterns. The authors show that their algorithm achieves high utility and effectiveness of protecting location access data. However, there are a few drawbacks of this work, which we will address in this paper. First, the protection of frequent accessed location patterns does not protect individual privacy at less popular locations in ID based IoV data. Second, the LPT data structure grows exponentially when the number of locations grow, making the algorithm impractical for large amount of data.

Throughout our research, we find that almost every researcher has had different ideas about what location privacy should be and how to protect it. Despite efforts in exploring different techniques in achieving location privacy of IoT data, there is a lack of consensus on the definition of location privacy. Furthermore, there are few holistic views of location privacy breaches and mitigation at different stages of an IoV application.

The contributions of this paper are as follows:

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 investigate a different strategy of applying Differential Privacy (DP) to the real world Gowalla dataset than the one proposed by [6]. We show that instead of locations that are accessed frequently, the locations with less *unique visitors* are extremely sensitive. Instead of applying DP to frequencies of location patterns, we apply DP to aggregated location groups based on their geometric positions. We use a k-d tree data structure which is a natural choice for generalizing the locations so that differential privacy can be more appropriately applied to protect sensitive locations. We show that our strategy produces for 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 the location groups.

The paper is organized as follows: Section III and Section IV discuss the necessary motivation, problem and background knowledge to understand the concepts discussed in this paper. Section V includes the bird's eye view of the location privacy preservation scheme and an overview of the metrics by which we evaluate each method. Section VI and section VII include explanation of our experiment in Differential Privacy and analysis of the results. Finally, we conclude with section VIII where our contributions are summarized and we propose some future work for this topic.

III. Motivation

The phrase "data is the new oil" refers to the priceless value that data has. We have been experiencing the early stages of the "information age" since the wide adoption of the world wide web. Only in recent years has the general public slowly realized the value of the data that they generate when interacting with internet capable devices. It has become common to hear about data privacy breaches in various companies. Facebook, Mariott, and even United States Postal Service have all been victims of data privacy breaches in the millions of records within the past year [7]–[9]. These companies all stored their records in plain text. However, if these companies had employed some privacy preserving techniques within their data, this would have prevented attackers from being able to derive any value from the data. This is one of the reasons privacy preservation in general is important.

A general IoV model involves communication between vehicles, infrastructure and a number of other entities. The data stored in or exchanged between any of these entities may contain all sorts of sensitive data. Even though it would be unwise to store a direct universal identifier (e.g., license plate number for an ID), these systems will need some way to identify each of the vehicles that are on the network. It has been shown that even by storing seemingly harmless qualities of vehicles or individuals, or using weak privacy protection techniques, it is trivial to re-identify an individual. Qualities such as age, salary, and geographic location can all lead to re-identification through a background information attack [10]. These qualities are referred to as "Quasi-identifiers", and can be surprisingly elusive if one is not aware of general privacy attacks and techniques.

A. Model Setting and Problem Statement

For the purposes of this paper, we are concerned with vehicle's data, specifically the storage of this data in the IoV Cloud. Each data record includes some ID, a Timestamp, and Location. The Cloud stores such information to perform operations on it in order to provide services to automated vehicles such as traffic condition services which we will focus on here. Traffic condition services are a classification of services that provide solutions to the problem of avoiding traffic related issues or gaining traffic related information. The Traffic condition service model consists of three main stages.

- The first stage is concerned with vehicles updating the Cloud with its ID, location, and timestamp. This is essential for being able to support basic traffic related services as it provides information about where vehicles are at a particular time.
- The second stage regards the vehicle querying the Cloud about traffic information around a particular location. An example query may look like: "How many vehicles are at location X?" The Cloud will compute the query and return the answer to the vehicle that queried the Cloud via the third stage of this model. It is clear that the data about a vehicle or group of vehicles contain extremely sensitive information, and

should be stored with the utmost privacy in the data storage unit (the Cloud in this case).

- Although the data are stored in the Cloud, they may be requested for a number of reasons that do not fall under the use case of providing the traffic condition service. The data may be published to allow for the research community to experiment ideas on real data sets in order to fine tune, improve and innovate new services. Additionally, and unique to traffic data, the data here may be audited by an insurance company in the event of a vehicle insurance claim. The data may also be requested by a police department or other law enforcement agency to aid in the investigation or search for a criminal. These requirements alone cast a wide and complicated net when considering how best to store the data so that it maintains realistic utility, and preserves the privacy of the individuals using the services.

In our paper, we analyze location privacy preservation in the three stages of data handling for IoV traffic condition service. We take this opportunity to outline some of the potential attacks that can be carried out on this model. We assume for the first two following attacks menitoned below, the Cloud uses some ID for each user, that the adversary does not initially know. However, the Cloud does not utilize any other privacy protection techniques. In the last attack, we assume that the Cloud is the adversary and would like to track a user through the user's queries.

B. Attack 1: Simple UserID background attack

The first attack involves an adversary querying the Cloud to gain an individual's location based on their user ID. The adversary does not know the user's ID in the Cloud. However, with a small amount of background information, the adversary can easily obtain this ID as we will demonstrate. Assume our victim is Officer Tom. Each day Tom checks in at a military base that only he has access to. Therefore he is the only person that is ever at this location. The adversary happens to know this, as well as the location of the military base. The adversary decides they would like to find out Tom's user ID but cannot query this directly. So the clever adversary decides to query the Cloud with the following instead: "SELECT * FROM DB WHERE UserID = (SELECT UserID FROM DB WHERE location = X)", where X is the location of the military base. Since Officer Tom is the only person ever at this base, the Cloud will return a single row from the database that contains Tom's UserID. Now the adversary can learn the location of Officer Tom even when he is not at the military base.

C. Attack 2: Dynamic UserID background attack

This attack is similar to Attack 1, however in this case, the Cloud employs the use of dynamic UserIDs, where the ID for any user is mapped to a unique list of values that change from time to time, so that even if an adversary obtains one of their UserIDs, they cannot successfully track the location of that particular user. However, we show that this approach is still not effective to ensure location privacy. Consider the situation from Attack 1, where Officer Tom is still the only resident at a military base that the adversary knows the location of. The adversary can determine whether Officer Tom is there at a given time or not. Suppose the adversary runs the query: "SELECT count(*) FROM DB WHERE location = X", again 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.

D. Attack 3: Untrusted Cloud attack

For this final attack, the Cloud is untrusted and is the adversary. We assume that the users are innocent and trusted. The Cloud contains traffic conditions of various locations that a user may be interested in but does not have this particular user's location. The Cloud would like to find the location of the user, say user Tom. If Tom queries the Cloud regarding a particular location, then the Cloud can infer that Tom may be interested in this location and may either be heading there at some time in the future, or Tom may already be in that 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. Although, this attack is less likely to happen in practice, it is important to be considered.

In the following section, we describe work that has been conducted on data privacy in general that is relevant to the techniques we explore in this paper.

IV. Background: Privacy Techniques

This section details the core concepts of the privacy techniques that we have chosen to consider in our paper, as well as some pros and cons of these techniques as a whole.

A. Differential Privacy

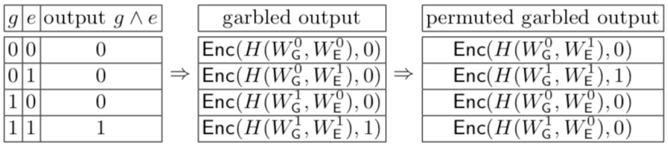
Differential Privacy is first presented in 2006 by Dwork [11]. Differential privacy is a technique used in long-term data storage or data publishing. The core idea here is to eliminate the risk of an individual joining a statistical data set [12] (i.e., the risk is the same as if you had not joined the set). Differential Privacy involves comparing two databases that differ by at most one row [11]. Differential Privacy is achieved if the probability of selecting any two rows from the databases is the same or worse than a coin flip [12]. This effectively removes the possibility of a background knowledge attack since the likelihood of picking any two rows is the same, regardless of what an individual knows about the data set. Specifically, we explore ϵ-differential privacy. To achieve ϵ-differential privacy, noise is added among the rows of a database using a Laplace distribution, according to the value of ϵ [12]. As ϵ is increased, the utility of the data is increased and privacy is decreased. As ϵ is reduced, the opposite happens and we achieve better privacy at the cost of losing utility [12]. Generally, Differential Privacy is superior to many other data privacy techniques such as k-anonymity [10], t-closeness [13], l-diversity [14] and their variants since Differential Privacy provides privacy and removes the possibility of a background knowledge attack, which all of these other techniques are susceptible to [15]. ϵ-Differential Privacy can also be extended to group privacy or individuals that contribute more than one row to the data set. Although Differential Privacy is a valuable concept and an admirable goal, it is not perfectly private. To be perfectly private would mean not releasing any data at all, ever. However in order to be productive, as a society we need to agree that the benefits of sharing some data outweigh the risks [12].

B. Private Information Retrieval

Private Information Retrieval (PIR) is a concept that is proposed in 1997 by Chor et al. [16]. The authors of this original paper realize that although there are many techniques developed to protect the privacy of data stored in a database, there are no techniques to protect the users that query the database. For example, if a user queried the database about some points of interest at some location, this implies that the user has some interest and may be heading to this location, or is already at this location. The authors achieved this private information retrieval by encrypting the user's query and giving the database the encrypted query. Then, the database will run some computation on the encrypted query and return an encrypted result. Here the database has no idea what has been queried or returned. A recent improvement on PIR for vehicles is known as PIR in Vehicle Location-Based Services (VLBS) proposed by Tan et al. [17] in 2018. This technique is designed to work well in the vehicular setting and is much faster than standard PIR. PIR in VLBS allows the user to filter the queried data set such that privacy is maintained. This is achieved by partitioning the queryable area into segments and assigning Points of Interest (POI) to certain areas based on their distance from a road segment. The size of the groups of POI are always the same, and since the query and response are encrypted, there is no way for the adversary to know what data have been requested or returned.

C. Garbled Circuit

Garbled Circuits are a relatively old concept presented by Yao in 1986 [18]. This concept 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. In circuit logic, a set of gates can be mapped to a simple truth table, where the gates represent logical operations. In a garbled circuit however, the mapping to the truth table is rearranged by the garbler. The garbler will take input values to a gate and encrypt them, so that the other communicating party does not know the input. The garbler will perform the gate operation on the input values prior to encryption to obtain the output value. Then, each encrypted input is paired with the corresponding output and the value is stored together in the re-arranged truth table. Figure 1 follows from [19], here W_X^Y is mapped from $X = Y$, so $W_G^0 = (g = 0)$:



g	e	output $g \wedge e$		garbled output		permuted garbled output
0	0	0		$\mathsf{Enc}(H(W_{\mathsf{G}}^0, W_{\mathsf{E}}^0), 0)$		$\mathsf{Enc}(H(W_{\mathsf{G}}^0, W_{\mathsf{E}}^1), 0)$
0	1	0	⇒	$\mathsf{Enc}(H(W_{\mathsf{G}}^0, W_{\mathsf{E}}^1), 0)$	⇒	$\mathsf{Enc}(H(W_{\mathsf{G}}^1, W_{\mathsf{E}}^1), 1)$
1	0	0		$\mathsf{Enc}(H(W_{\mathsf{G}}^1, W_{\mathsf{E}}^0), 0)$		$\mathsf{Enc}(H(W_{\mathsf{G}}^0, W_{\mathsf{E}}^0), 0)$
1	1	1		$\mathsf{Enc}(H(W_{\mathsf{G}}^1, W_{\mathsf{E}}^1), 1)$		$\mathsf{Enc}(H(W_{\mathsf{G}}^1, W_{\mathsf{E}}^0), 0)$

Figure 1. The garbling of an AND gate [19]

Now the evaluator would like to decrypt exactly one ciphertext from the garbled truth table, to revel the values of g and e that correspond to W_G^g and W_E^e that the garbler has sent [19]. The evaluator also receives the garbled gate from the garbler. But there are some restrictions on this decryption. The evaluator cannot be sent both W_E^0 and W_E^1 because then the evaluator can decrypt two ciphertexts [19]. The evaluator can not ask for which specific value they want either since they do not want the garbler to know which specific value they are after. This is called oblivious transfer and allows the evaluator to find out only W_E^e without revealing e to the garbler. The evaluator also needs to know when decryption succeeds and when it does not in order for this technique to succeed [19].

The next section will describe techniques that we consider will achieve vehicle location privacy, as well as some attacks and mitigation that can be imposed on these techniques.

V. Overview of privacy techniques and Proposed attacks, with solutions

We evaluate the techniques using three metrics. As shown in Table I, each metric is concerned with location privacy at a different stage in our model. The three metrics are: location privacy at traffic update, location privacy at traffic storage (or trajectory privacy) and location privacy at traffic query. The number of ticks represents the effectiveness of a technique for a particular privacy concern. Table II shows that each technique uses a slightly different model in terms of which communicating party is identified as the adversary. In certain techniques, the Cloud is the adversary, and the vehicle is an innocent user. In other techniques, the Cloud is trusted and the vehicle is not trusted.Some models may also involve a trusted third party, in addition to the Cloud and the vehicle. This third party is commonly referred to as a Trusted Authority (TA) in IoV literature.

TABLE I. 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	✓		✓✓	✓✓

TABLE II. 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

A. Dynamic Pseudonyms

The first technique we examine involves the use of pseudonyms. The Cloud is the adversary and the vehicle and TA are trusted. This technique is centered around the idea of protecting a user's location by mapping their real identifier to a constant pseudonym that is generated by the TA. Here, the TA is used as an intermediary between the vehicle and the Cloud. However, this technique is susceptible to Attack 1. So location privacy is not adequately preserved here.

An alternative pseudonym technique that is also considered is the dynamic pseudonym, where, a user's real identifier is mapped to a list of pseudonyms that change at a predetermined time, and therefore appear different to the Cloud. Only the

TA is able to determine the real identity of the mapped pseudonyms. This technique is, however, susceptible to Attack 2. Therefore the pseudonym approach does not achieve location privacy for an individual.

B. Differential Privacy

The second technique we consider involves adding Differential Privacy to the Cloud that stores our data. In this model, we have only the Cloud and the vehicle involved in communications, where the Cloud is trusted however the vehicle/ user is not. Here the user will attempt to gain information about other users using seemingly harmless queries. As is standard in Differential Privacy, noise is added to the database rows to add privacy. However 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. Another weakness of this technique is the fact that 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. Making an attack similar to Attack 1 or Attack 2 viable.

C. Private Information Retrieval

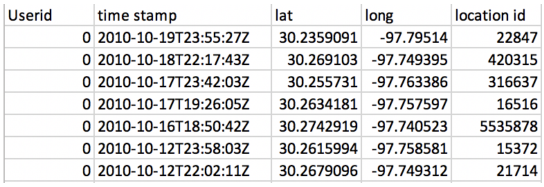
For our third technique, we explore a special case of Private Information Retrieval. PIR in VLBS can be utilized to provide privacy at query time. Here the Cloud is the adversary and the vehicle is trusted. Since the queries to the Cloud are encrypted, the Cloud has no way of knowing what the vehicle's are querying. However, the Cloud attempts to figure this out based on which rows are returned after a query. But according to PIR, a group of the same number of rows are returned each time a query is asked making it impossible to pin point exactly what the vehicle was querying about. However, if a vehicle chooses to update the Cloud at any point, its exact location will be revealed to the adversary. As a consequence of this, the vehicle's location privacy at storage is not maintained since the vehicle checks in to update its location over time.

D. Garbled Circuit

Finally, our fourth technique involves using a garbled circuit in conjunction with a TA. This technique attempts to satisfy each metric that we are evaluating with. In this model, the Cloud is our adversary and is untrusted once again, and the vehicle/ user is trusted, along with the TA. Consider the situation where the vehicle updates the Cloud with its location. 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. On this same note, location privacy of a vehicle is preserved over the long term as the data stored is encrypted. On a query about traffic related to a certain location, the Cloud is not aware of the value that is being requested for. Therefore location privacy is preserved once again. This technique seems to be the most private, however it is also the most complex.

VI. Experiment

As shown in section 2, row based location data is susceptible to attacks that may personalize sensitive location data. In our experiment section, we investigate Differential Privacy to centrally stored location data in the same real dataset used by Yin et al. [6]. The location check in data Gowalla was a location-based social network that was active between 2007 and 2012. The dataset includes a total of 6,442,890 check-ins of these users over the period of Feb 2009 and Oct 2010. Figure 2 shows a snapshot of the Gowalla dataset. Although the dataset is not strictly IoV data, it shares similairty with IoV data by having location, timestamp, and ID in each row. It should be mentioned that since this dataset is not strictly traffic data and does not necessarily have continuous timestamps for each user by minutes or hours, this issue can be rectified by generalizing timestamps to dates and then building a contingency table with missing dates, as we will discuss in a later subsection.



Userid	time stamp	lat	long	location id
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

Figure 2. Snapshot of Gowalla Dataset

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. The programs are written in Python, and the experiments are run on a MacBook Pro with 2.3 GHz Intel Core i5 and 8 GB 2133 MHz LPDDR3.

A. Data Cleaning

After plotting the normalized Gowalla locations, we notice some outliers that affect the generalization of geometric distribution. As shown in Figure 3, a few outliers at the topright corner greatly affects the performance of $k-d$ tree.

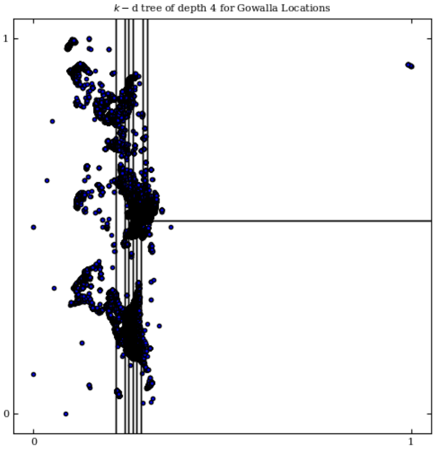


Figure 3. Gowalla Locations With Outliers

We remove these outliers by removing 37 locations with large z scores, a statistical metric of a value relative to the sample mean and standard deviation. Figure 4 shows the plot of normalized Gowalla locations after the outliers are removed.

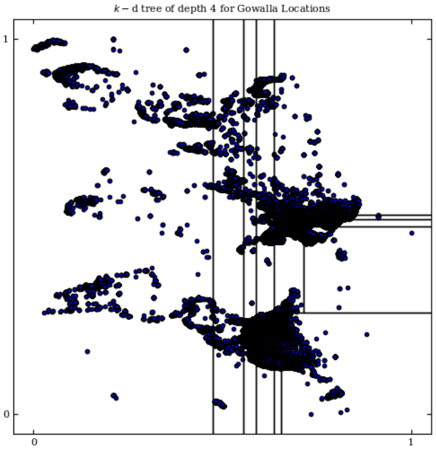


Figure 4. Normalized Gowalla Locations Without Outliers

B. 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 [20]. For our data, building a contingency table means to create continuous dates for each location group and unique user combination. To do this, we calculate the minimum and maximum dates in the dataset, and add missing dates to all location groups with user count set to 0. Note that the original dataset has timestamps based on hours and minutes, however it is less common for a user to visit a location on an hourly basis and more common for the user to visit the location at different times of different dates. Therefore, we generalize the timestamp to dates to avoid excessive numbers of rows being added to the contingency table, which affects the data utility.

C. Generalization

We experimented different depths of 4, 5, 6 of the $k-d$ tree for the generalization of locations. Through evaluation we determine that depth 6 is proper for the group generalization as it provides more granularity. After each location is assigned a group ID, the original dataset is aggregated to a dataset with dates, location groups, and count of unique users at the date/location group combination. The data cleaning and generalization shrinks 6,442,890 checkins to 40,128 aggregated records. Figure 5 shows the generalized location data of the Gowalla dataset.

D. Differential Privacy

For each generalized data point, Lap($1/\epsilon$) is added to the user counts. In our experiments, we tried $\epsilon = 0.1, 0.5, 1.0$. Figure 6 shows $0.1-$differentially private Gowalla dataset. 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.

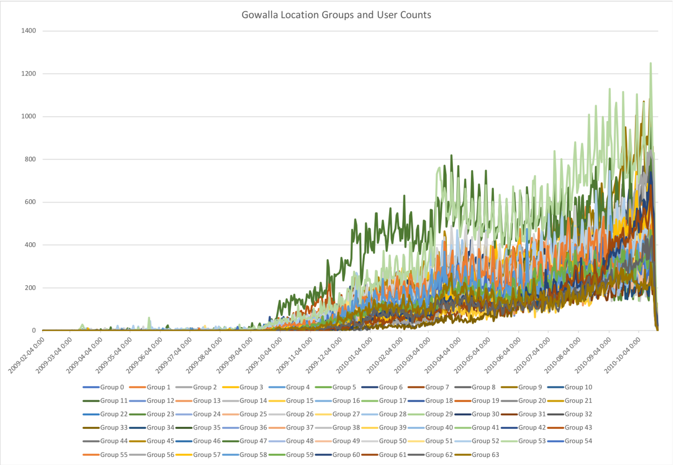


Figure 5. Gowalla Location Groups and User Counts ($k-d$ tree depth = 6)

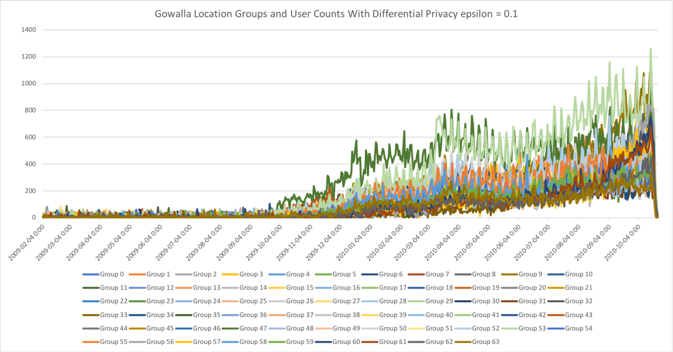


Figure 6. Gowalla Location Groups and User Counts with Differential Privacy ($k-d$ tree depth = 6)

VII. Evaluation and Analysis of Results

A. Data Quality

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. This approach is similar to the evaluation of classification quality in Mohammed et al. [20]. We use an LSTM traffic predictor utilized in Fu et al. [21] 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. 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 [22].

TABLE III. Location Group 63 Prediction Comparison of Different Training Models

Measurement	*Orig model*	*DP ϵ = 0.1*	*DP ϵ = 0.5*	*DP ϵ = 1.0*
Explained variance score	**0.713**	**0.670**	0.390	0.469
RMSE (root mean squared error)	**45.676**	**48.893**	56.583	50.343
R2 score	**0.513**	**0.442**	0.253	0.409

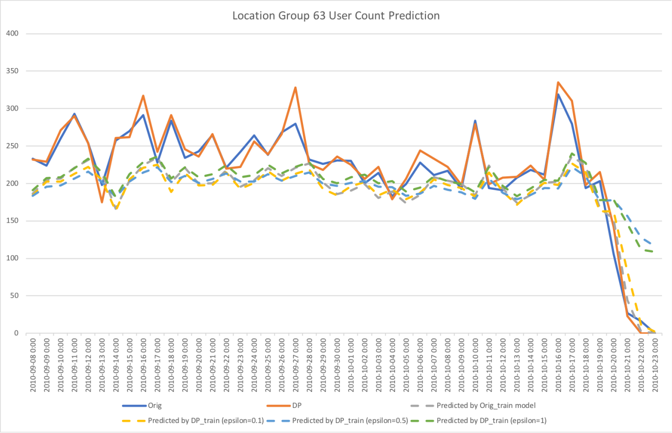


Figure 7. Location Group 63 Real Data vs. Prediction

Table III 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. Figure 7 shows that in general, the predicted data by all DP-data-trained models are reasonable compared to the real data.

VIII. Conclusion and Future Work

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

There is a lot of potential for future work. Section V leaves many avenues open for pursuing research on the techniques we have proposed here. Private Information Retrieval 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 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.

References

[1] E. Borcoci, From Vehicular Ad-hoc Networks to Internet of Vehicles, 2017, URL: https://www.iaria.org/conferences2017/ [accessed: 2020-06-01].

[2] M. Raya and J.-P. Hubaux, “Securing vehicular ad hoc networks,” Journal of computer security, vol. 15, no. 1, 2007, pp. 39–68.

[3] F. Wang, Y. Xu, H. Zhang, Y. Zhang, and L. Zhu, “2flip: a two-factor lightweight privacy-preserving authentication scheme for vanet,” IEEE Transactions on Vehicular Technology, vol. 65, no. 2, 2016, pp. 896–911.

[4] H. Zhong, S. Han, J. Cui, J. Zhang, and Y. Xu, “Privacy-preserving authentication scheme with full aggregation in vanet,” Information Sciences, vol. 476, 2019, pp. 211–221.

[5] W. U. Hassan, S. Hussain, and A. Bates, “Analysis of privacy protections in fitness tracking social networks -or- you can run, but can you hide?” in Proceedings of the 27th USENIX Security Symposium. USENIX, 2018, pp. 497–512.

[6] C. Yin, J. Xi, R. Sun, and J. Wang, “Location privacy protection based on differential privacy strategy for big data in industrial internet of things,” IEEE Transactions on Industrial Informatics, vol. 14, no. 8, 2018, pp. 3628–3636.

[7] “Zuckerberg says facebook working with fbi to investigate security breach,” URL: https://www.cnbc.com/video/2018/09/28/zuckerberg-says-facebook-working-with-fbi-to-investigate-security-breach.html [accessed: 2020-06-01].

[8] “Marriott data breach is traced to chinese hackers as u.s. readies crackdown on beijing,” URL: https://www.nytimes.com/2018/12/11/us/politics/trump-china-trade.html [accessed: 2020-06-01].

[9] “Usps site exposed data on 60 million users,” URL: https://krebsonsecurity.com/2018/11/usps-site-exposed-data-on-60-million-users [accessed: 2020-06-01].

[10] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, 2002, pp. 557–570.

[11] C. Dwork, “Differential privacy,” Automata, languages and programming, 2006, pp. 1–12.

[12] ——, “Differential privacy: A survey of results,” in International conference on theory and applications of models of computation. Springer, 2008, pp. 1–19.

[13] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007, pp. 106–115.

[14] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” in 22nd International Conference on Data Engineering (ICDE’06). IEEE, 2006, pp. 24–24.

[15] D. Kifer, “Attacks on privacy and definetti’s theorem,” in Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009, pp. 127–138.

[16] B. Chor, O. Goldreich, E. Kushilevitz, and S. Madhu, “Private information retrieval,” 1997, pp. 0–20.

[17] Z. Tan, C. Wang, M. Zhou, and L. Zhang, “Private information retrieval in vehicular location-based services,” IEEE, 2018, pp. 56–61.

[18] A. C.-C. Yao, “How to generate and exchange secrets,” in 27th Annual Symposium on Foundations of Computer Science (sfcs 1986). IEEE, 1986, pp. 162–167.

[19] S. Yakoubov, “A gentle introduction to yao’s garbled circuits,” 2017, pp. 1–12, URL: http://web.mit.edu/sonka89/www/papers/2017ygc.pdf [accessed: 2020-06-01].

[20] N. Mohammed, R. Chen, B. Fung, and P. S. Yu, “Differentially private data release for data mining,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 493–501.

[21] R. Fu, Z. Zhang, and L. Li, “Using lstm and gru neural network methods for traffic flow prediction,” in 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC). IEEE, 2016, pp. 324–328.

[22] “scikit-learn metrics,” URL: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics [accessed: 2020-06-01].