2016 IEEE International Conference on Data Science and Advanced Analytics

Anonymizing NYC Taxi Data: Does It Matter?

Marie Douriez*, Harish Doraiswamy†, Juliana Freire†, and Cláudio T. Silva†
*Ecole Polytechnique †New York University
Email: marie.douriez@polytechnique.edu, {harishd, juliana.freire, csilva}@nyu.edu

***Abstract*—The widespread use of location-based services has led to an increasing availability of trajectory data from urban environments. These data carry rich information that are useful for improving cities through traffic management and city planning. Yet, it also contains information about individuals which can jeopardize their privacy. In this study, we work with the New York City (NYC) taxi trips data set publicly released by the Taxi and Limousine Commission (TLC). This data set contains information about every taxi cab ride that happened in NYC. A bad hashing of the medallion numbers (the ID corresponding to a taxi) allowed the recovery of all the medallion numbers and led to a privacy breach for the drivers, whose income could be easily extracted. In this work, we initiate a study to evaluate whether "perfect" anonymity is possible and if such an identity disclosure can be avoided given the availability of diverse sets of external data sets through which the hidden information can be recovered. This is accomplished through a spatio-temporal join based attack which matches the taxi data with an external medallion data that can be easily gathered by an adversary. Using a simulation of the medallion data, we show that our attack can re-identify over 91% of the taxis that ply in NYC even when using a perfect pseudonymization of medallion numbers. We also explore the effectiveness of trajectory anonymization strategies and demonstrate that our attack can still identify a significant fraction of the taxis in NYC. Given the restrictions in publishing the taxi data by TLC, our results indicate that unless the utility of the data set is significantly compromised, it will not be possible to maintain the privacy of taxi medallion owners and drivers.**

***Keywords*-privacy attacks, trajectory privacy, taxi data, spatio-temporal data**

I. Introduction

The easy availability of low cost sensors has enabled the collection of enormous amounts of data pertaining to cities. In particular, the widespread use of GPS-enabled smart phones and other location-based devices coupled together with social media has led to an explosion in the volume of spatio-temporal data sets. Cities all over the world are not only collecting these data, but they are also making them available (see e.g., [1], [2], [3]). Analysis of these urban data can greatly help plan and improve policies of cities, thus improving the lives of their residents [4], [5].

Since many of these data sets relate to organizations as well as individuals of the city, their publication can severely threaten the privacy of the concerned parties. Even though each of these data sets might be safe individually, the privacy risks permeating from them still remains due to two main reasons – 1) many of the data sets are released by multiple unrelated sources, thus increasing the possibility of recovering an individual's identity through careful collation of the appropriate data sets; and 2) it is also possible for an adversary to legally gather additional data which can ease the said collation process. Restrictions due to government regulations makes maintaining the privacy of published data further difficult.

Privacy breach using the New York City taxi data. The taxi data set gathered by the New York City's (NYC) Taxi and Limousine Commission (TLC) is composed of historical data of the trips taken by yellow cabs in NYC. Each trip consists of pickup and drop-off locations and times, along with other relevant data such as the distance, duration, fare, and tip. There are on average 500 thousand trips each day amounting to approximately 170 million trips per year. When the data set was published, there were several breaches of privacy on the customers that were discovered by combining the taxi data with data gathered from other sources. For example, using the time and place obtained from pictures found on the internet of celebrities getting on or off a taxi, and matching it with the released trip data, it was possible to retrieve the probable trip done by the celebrity and thus identify not just where they might reside, but also gather their tipping patterns [6], [7].

Another type of breach, which is the focus of this work, was on the identities of the taxi drivers and medallion owners. In particular, TLC is interested in hiding the income of the drivers. In order to do this, they used an MD5 hash to pseudonymize the medallion and license numbers. However, making use of the fixed patterns these IDs have, it was possible to recover the actual IDs, thus allowing an adversary to easily obtain each driver's income as well as driving patterns [8].

The taxi data was initially released under the Freedom of Information Law of NYC (FOIL) wherein people can request data from TLC and gain access to it. Legal restrictions do not allow TLC to publish incorrect data. Further, if any data is modified, then the extent of such modification should also be published. For example, if the time of a trip is approximated to the nearest half-hour, then the actual time range for that trip should also be mentioned when publishing the data. A naive solution under the above restrictions to avoid revealing the drivers' income is to simply not publish

978-1-5090-5206-6/16 $31.00 © 2016 IEEE
DOI 10.1109/DSAA.2016.21

the fare (and the tip) for each trip. However, the fare can still be recomputed from the data itself using the start and end locations and times, or using the duration and distance attributes. Thus, these attributes should also not be published to avoid revealing driver incomes. But, these are important attributes and are used to gain insights into not only patterns of taxi movement [9], but also traffic mobility patterns [10]. Thus not revealing them, or only partially revealing the data, would lead to a great a loss of information, making the data useless.

Contributions. Working together with experts from TLC, the goal of this work is to evaluate whether "perfect" anonymity is possible and if *identity disclosure* can be avoided given the availability of diverse sets of external data sets through which the hidden information can be recovered. To do so we first formalize the problem as that of trajectory anonymization. Then, assuming an ideal pseudonymization of medallion IDs, we devise an attack to recover these IDs. This is accomplished under the assumption of availability of an external data set that provides the actual medallion IDs of a subset of taxis at a given location and time. We also show how such a data set can be obtained relatively easily by a determined adversary.

Our attack is based on performing a spatio-temporal join of the published taxi data with the above external data. Given that a spatio-temporal join is costly due to the large size of the data, we propose an equivalent and simpler intersection-based algorithm to perform this join using the properties of the external data set. We show that using this attack, even with an ideal pseudonymization, it is possible to recover more than 90% of the total number of medallion IDs.

We then explore different anonymization strategies (on top of pseudonymization) that can possibly be used within the legal constraints to prevent such attacks. Even using a coarse generalization strategy, we show that it is still possible to recover a significant fraction of the true medallion IDs. Finally, we discuss various properties of our attack as well as the taxi data which makes maintaining anonymity for this data set difficult.

II. Related Work

Privacy preserving techniques have been extensively studied with respect to relational data (see e.g., [11], [12], [13]). In this section, we restrict the discussion to those related to spatio-temporal data, which is the focus of this work.

Privacy preserving strategies. As more and more location information about individuals becomes available, through smart phones, GPS or other devices and applications, several systems that record and process this location data, called location-based systems (LBS) are developed. These systems raise a lot of privacy issues and several techniques have thus been proposed in the literature to protect it. These techniques primarily aim at protecting the user's *location* (also called *l-diversity* in this context), since a location in itself and especially its semantic can be a sensitive attribute (for example if the location is an hospital). To this end, the definition of *k-anonymity* used in relational databases has been extended to location databases, where one point should be indistinguishable from at least k-1 points. To achieve this property, a method using *position dummies* [14] has been proposed. Another idea was to define *mix-zones* [15] in which all the users' locations are hidden. More recently, the notion of *differential privacy* [16] has gained popularity: it requires that modifying a single user's data should have a negligible effect on the query outcome. The concept has for example been extended to *geo-indistinguishability* by Andres et al. [17] and is achieved by adding controlled random noise to the users location. The above mentioned techniques are typically used for maintaining online privacy, i.e., in real time when using LBS.

When historical data with respect to an entity or person is used, it results in a trajectory, i.e. a set of locations at different times. Abul et al. [18] were the first to address this perspective and proposed the notion of (k, δ)-*anonymity*, where δ represents the possible location imprecision. To achieve this, they proposed a method called *Never Walk Alone* which is based on trajectory clustering and space translation. A similar clustering and aggregation method is used by Chow et al. [19]. However, the notion of (k, δ)-*anonymity* was called into question by Trujillo-Rasua et al. [20] who proved that methods based on it, like Never Walk Alone, could offer trajectory k-anonymity only when $\delta = 0$.

Terrovitis et al. [21] assumed that an attacker has partial trajectory knowledge which works as a quasi-identifier, and links points of the trajectory to a person identity. The only difference of this quasi-identifier from the one defined in relational databases is that this one may have variable lengths. The authors then proposed a method to anonymize the trajectories by deleting carefully chosen points. Such methods however cannot be used by TLC to publish the taxi data because of the legal constraints that should be respected during its publication. Bettini et al. [22] proposed the notion of historical k-anonymity and defined location-based quasi-identifiers. A similar definition was later used by Sui et al. [23] to show that parking point information of Beijing taxi cabs enabled them to re-identify anonymized trajectories of drivers. Monreale et al. [24] presented an approach based on spatial generalization in order to achieve k-anonymity. However, these methods involve deleting points thus distorting not only the data but also the information derived from it. Differential privacy offers stronger anonymity guarantees and is also applied to trajectory data sets [25], and involves adding random noise to the data. Our goal is different from the above works – we are interested in protecting the user's *identity*, namely the driver or medallion owner. Therefore the semantics of the locations do not

matter much since they mostly depend on the customers. In order to maintain privacy of user identity, the notion of k-anonymity was extended to require that the location for one user be indistinguishable from at least k-1 other users' locations [26], [27], [28]. Gruteser et al. [26] proposed an online method using *spatial and temporal cloaking* that generalizes the user's location to a greater area and the time to a greater interval that contain at least k-1 other users.

More recently, Poulis et al. [29], [30] proposed an anonymization technique based on space generalization which does not involve deleting points. Here the generalization is done in a discrete sense, where the actual location is one among a set of locations.

Attack strategies. There has been much work on identifying attack strategies for anonymization techniques of relational data [31], [32], [33], [34]. However there has not been much work on attack strategies for location and trajectory-based privacy techniques. While multiple works hypothesize that it is possible to breach the privacy using external data sets [18], [21], [24], [29], [30], none of them actually show how such an attack can be accomplished.

To the best of our knowledge, the work by Sui et al. [23] is the only work that attempts an attack to test the privacy of a published spatio-temporal data set. They use the Beijing taxi data itself to identify parking points or stay points and use it to re-identify taxis based on the patterns of these parking points. However, they do not show the effectiveness of such an attack when using anonymization strategies. Unlike this work, our goal is to use an external data set to identify taxis in NYC. We also show the effectiveness of our strategy with respect to allowed anonymization techniques. Our strategy can be considered as an extension of the composition-based attack used for relational data [32]. Additionally, we also show how such external data set can be obtained by an adversary.

III. Problem Setup

The NYC taxi data provided by the TLC is composed of historical data of the yellow taxi trips in NYC. Each trip consists of pickup and drop-off locations and times, along with other relevant data such as the fare and tip. There are around 13,000 yellow taxis making on an average 500,000 trips each day. The TLC were interested in publishing this data in a manner that prevents recovering of driver incomes. To test the feasibility of existing anonymization techniques, the data has to be transformed into the appropriate format. Given the spatial and temporal nature of the taxi data, trajectory anonymization techniques are most suited for this purpose.

In this section, we describe the data setup used by our attack strategy to test the effectiveness of the anonymization techniques. Since the provided taxi data does not satisfy the format of a trajectory that is used by the various techniques, we first define a transformation of the taxi trips into a set of trajectories. We then describe the format of an external medallion data set that can be gathered by an adversary and used for the purpose of recovering the identities of the taxis. We finally describe how we simulate this medallion data set for our experiments.

A. Moving Object Database

A *trajectory* is defined as a sequence of couples $T = (l_1,t_1) \rightarrow (l_2,t_2) \rightarrow \ldots \rightarrow (l_n,t_n)$ where $\{l_i \in \mathbb{R}^2\}$ is a set of geospatial coordinates and $\{t_i \in \mathbb{R} | t_i < t_j \text{ iff } i < j\}$ is a set of time stamps corresponding to the locations. A trajectory $T' = (l'_1,t'_1) \rightarrow (l'_2,t'_2) \rightarrow \ldots \rightarrow (l'_{n'},t'_{n'})$ is a *sub-trajectory* of T ($T' \preceq T$) if there exist integers $1 < i_1 < \ldots < i_{n'} < n$ such that $\forall 1 \leq j \leq n'$, $(l'_j,t'_j) = (l_{i_j},t_{i_j})$. A *moving object database* is a set of trajectories $\mathscr{D} = \{T_1,T_2,\ldots,T_m\}$.

The NYC taxi data set can be easily transformed into a moving object database. To do so, we make use of the pickup (p) and drop-off (d) information associated with each trip as follows. We first sort the trips on the time of the trip. For each taxi, its trajectory is then defined as the set of locations of that taxi over time. This is provided by the pickup and drop-off locations of the ordered set of trips. This process is illustrated in Figure 1, where each taxi is associated with a unique ID. The sensitive attribute for each taxi corresponds to the total income which is the sum of the fare and tip over all trips. For the rest of this paper, we assume that the taxi IDs are *perfectly pseudonymized*, and therefore cannot be directly recovered.

B. Taxi Medallion Data Set

The key idea behind our attack strategy is to combine (or join) the taxi data with another data set. Typically, such an external data set is also "safe" in the sense that it does not reveal any sensitive information by itself. In particular, since the goal is to recover the ID's of taxis, we are interested in a data set that provides us with the actual medallion ID of taxis that are present in various locations over different time steps.

One way for an adversary to obtain such data is to collect it by themselves through the use of cameras. However, for such a data to be effective, it has to include data corresponding to almost all the taxis. A naive way to ensure this is to place cameras at every road intersection. However, such a strategy is expensive and inefficient. The cameras should therefore be placed at strategic locations to reduce the cost of this data gathering process. One way to reduce this cost is to look at only hot spots corresponding to locations having a high density of taxis [9]. At these locations, there are a lot of distinct taxicabs that pass every day and a majority of them tend to pass through these locations several times in a single day itself.

For this work, we focus only on the following four locations, shown in Figure 2 – NY Penn Station, Columbus Circle, Port Authority Terminal and the intersection between

id	Trajectory	Sensitive Attribute
tid_1	$(l_{1,1},t_{1,1},p) \rightarrow (l_{1,2},t_{1,2},d) \rightarrow \ldots \rightarrow (l_{1,n_1},t_{1,n_1},d)$	$\frac{1}{2}\sum_{i=1}^{n_1}(fare_i+tip_i)$
tid_2	$(l_{2,1},t_{2,1},p) \rightarrow (l_{2,2},t_{2,2},d) \rightarrow \ldots \rightarrow (l_{2,n_2},t_{2,n_2},d)$	$\frac{1}{2}\sum_{i=1}^{n_2}(fare_i+tip_i)$
$\ldots$	$\ldots$	$\ldots$
tid_m	$(l_{m,1},t_{m,1},p) \rightarrow (l_{m,2},t_{m,2},d) \rightarrow \ldots \rightarrow (l_{m,n_m},t_{m,n_m},d)$	$\frac{1}{2}\sum_{i=1}^{n_m}(fare_i+tip_i)$

Figure 1. Transforming the taxi data into a moving object database.

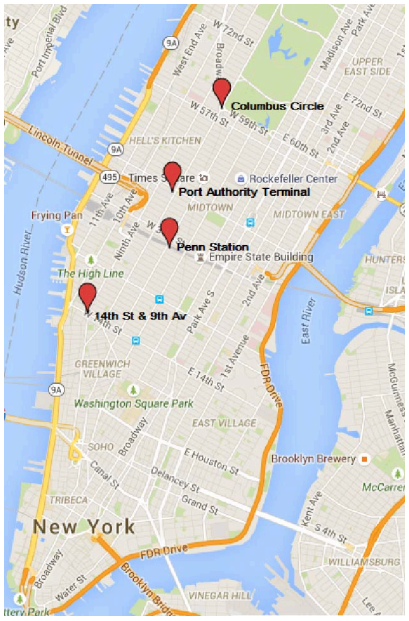


Figure 2. The four locations chosen to build our external medallion data set.

14th street and 9th avenue. Even if we consider the taxis at these locations only from 1 pm to 6 pm every day of a single month, we can still observe 12,248 distinct taxis, which is more than 91% of the total number (13,427) of yellow cabs plying in NYC [35]. As we show later, the information obtained from these locations is sufficient to effectively identify the taxis from the pseudonymized data. Thus, a determined adversary could alternatively also allocate observers at each of these locations to manually make note of the medallion IDs and times of the taxis that pass through these locations.

C. Simulating the Medallion Data Set

To simulate the medallion data set, we make use of the taxi data set itself as follows. Since the provided data contains exact GPS locations and times of the pickups and drop-offs, we approximate the medallion data set to include only taxis whose pickup or drop-off locations are within a small distance (approximately 100 ft) of the selected locations. We assume that the precision of the time stamps of the observations of the medallions is upto a minute. For each taxi ID, we assign a unique medallion ID to simulate the actual medallion ID.

id	trajectory
id_1	$(l'_{1,1},t'_{1,1}) \rightarrow \ldots \rightarrow (l'_{1,n'_1},t'_{1,n'_1})$
id_2	$(l'_{2,1},t'_{2,1}) \rightarrow \ldots \rightarrow (l'_{2,n'_2},t'_{2,n'_2})$
$\ldots$	$\ldots$
id_m	$(l'_{m,1},t'_{m,1}) \rightarrow \ldots \rightarrow (l'_{m,n'_m},t'_{m,n'_m})$

Figure 3. Medallion data set transformed into a moving object database.

The above medallion data set can also be transformed into a set of trajectories. In particular, it is a set of sub-trajectories of the taxi trajectories obtained from the original data set as shown in Figure 3.

IV. Spatio-Temporal Join Based Attack

Given the taxi data set together with the external medallion data set, the idea behind our attacking strategy is to match the sub-trajectories of the medallion data to the taxi trajectories. The set of unique matches would then provide the required taxi IDs. This matching can be accomplished using a spato-temporal join between the location and time periods of the two data sets.

A spatio-temporal join is a costly operation given the large size of the data sets. However, due to the way we have designed the medallion data, we know that the spatial attributes corresponding to the medallion data is restricted to a handful of locations. We use this observation to design a simpler equivalent algorithm to perform the required matching. We first describe this algorithm in Section IV-A. Next, in Section IV-B we show that even when using a perfect hashing has to pseudonymize the taxi IDs, our attack strategy can uniquely identify almost all the hashed IDs.

A. Attack Technique

The idea behind the algorithm is to perform a set of select operations followed by a set of intersection operations. Recall that the precision with respect to time of the medallion data is 1 minute. Given this, we first group the taxi data into 1-minute intervals. Next we perform a set of select operations on both the taxi data as well as the medallion data for each of the locations. For each location, there is one query corresponding to every 1-minute interval for which the medallion data is available. Note that by appropriately pre-processing the data (e.g., sorting on time and location), this operation can be efficiently accomplished. Alternatively, an efficient spatio-temporal index [36] can also be used to perform these queries. Let each select query, q_i, results in a list (set) of taxi trips R_i.

Real time	5 min	15 min	30 min
3:07 pm	3:05-3:10 pm	3:00-3:15 pm	3:00-3:30 pm

Table I
EXAMPLE OF TEMPORAL CLOAKING.

Time interval	1 min	5 min	15 min	30 min
# Taxis identified	12,224	12,131	11,893	11,639
Percentage /13,427	91.0%	90.3%	88.6%	86.7%

Table II
DE-ANONYMIZING TEMPORALLY CLOAKED TAXI DATA USING OUR ATTACK STRATEGY.

Given a taxi from the external medallion data, we first identify the time intervals this taxi appears in the selected locations. Next, we compute the intersection of all the query results R_i, where the query q_i corresponds to these time intervals. This intersection represents the set of taxis that appear at the selected locations during all the identified time intervals. For instance, if a taxi appears at three different times in a particular location in the medallion data, then we intersect the three different lists of the hashed IDs corresponding to these three times. If this intersection contains only one taxi, then that taxi's medallion ID corresponds to the remaining hashed ID. By removing the identified taxis from both data, and iteratively repeating the above procedure until no more taxis can be identified will result in a set of mappings between the real and pseudonymized IDs.

B. Effectiveness of the Attack

Pseudonymizing the IDs is a common tactic used to preserve user privacy [37]. We first test the effectiveness of our attack when such a strategy is used. As mentioned in the previous section, when using a hash, we assume that the IDs in the taxi data by itself cannot be recovered. The medallion data gathered using the four strategic locations in Manhattan contains sub-trajectories corresponding to 12,248 taxis. This data was gathered for a period of 1 month (March 2011) from 1 pm to 6 pm every day. Using the above algorithm to recover the taxi data using this medallion data, we were able to uniquely identify 12,224 out of the 12,248 medallion numbers, i.e., more than 99.8% of the observed taxis and more than 91% of the total number of taxis. Note that by gathering the medallion data for a longer time period will only improve the effectiveness of the attack.

V. USING TRAJECTORY ANONYMITY STRATEGIES

As shown in the previous section, even using a perfect pseudonymization is not sufficient since it is possible to recover the identity of almost all the taxis. It is therefore necessary to add further preventive measures to maintain the privacy of the taxi drivers. However, due to the imposed legal restrictions, trajectory anonymization techniques that add fake data or delete data cannot be used for this purpose. Also, during publication of the data, the TLC should publish the extent of any modification that is done to the data.

Real location	-73.991585, 40.749516
Census tract	101
ZIP code	10001
Neighborhood	Midtown-Midtown South

Table III
EXAMPLE OF SPATIAL CLOAKING.

Generalization-based techniques (or cloaking) can therefore be used, provided that the extent of the modification is specified during the publication of the data. These strategies consist of aggregating the data and thus decrease the precision of the spatial and / or temporal components of the data. In this section, we test the effectiveness of different types of generalization strategies with respect to our attack and report results from our experiments.

A. Temporal cloaking

In the first experiment, we consider data that has a decreased resolution for the pickup and drop-off times. In particular, we consider temporal resolutions of 5, 15 and 30 minutes. For a given resolution, we assume that the published data specifies the intervals in which a trip falls. Table I shows an example of the reported time for an example trip that happens at 3:07 pm.

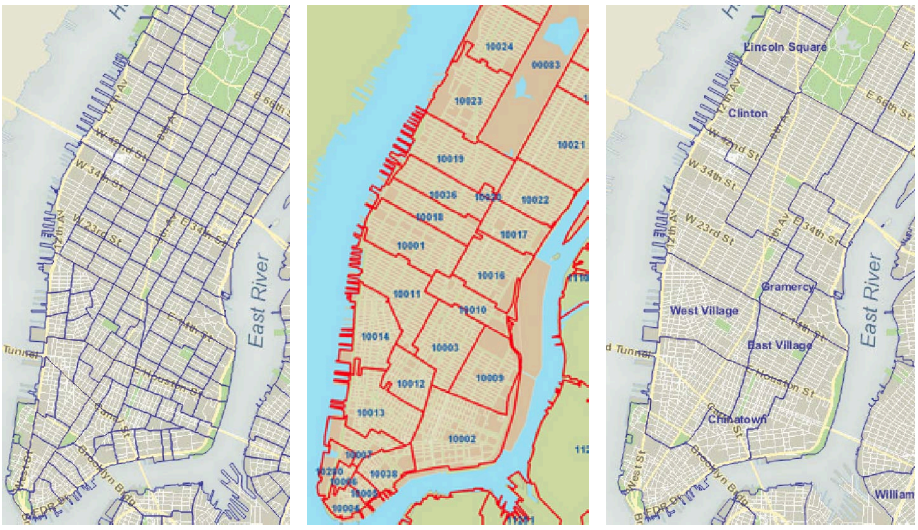
Using our attack strategy with the same medallion data set, we are still able to identify a significant number of taxis as shown in Table II. Given that over 11,500 taxis can be identified even with a coarse temporal resolution of 30 minutes, we can safely conclude that a temporal cloaking based strategy cannot prevent an attack aimed at recovering the identity of the taxis.

B. Spatial cloaking

In the next experiment, we consider the case when the resolution of the published pickup and drop-off locations is decreased. While there are different scales that are possible for this purpose, we consider three resolutions that are common in the NYC open data [2]. These three resolutions, in increasing order of size of the regions are census tract, zip code and neighborhood. The partitioning of Manhattan under these resolutions is illustrated in Figure 4. Table III shows an example of how the locations would be indicated in the different resolutions.

It happens that, in some cases, a selected location might be situated on the border between two (or more) regions (as defined by the resolution). For instance, Port Authority Terminal lies between two neighborhoods, "Clinton" and "Midtown South". In such a case, we use the trips from and to all these regions during the intersection process of the attack. The lists of encrypted IDs are therefore larger as they cover a larger area. Table IV shows the results obtained using the same medallion data observed at the four locations during March 2011.

Thus, even when locations are rounded off to large regions when using the neighborhood resolution, it is still possible to identify more than 11,000 taxis that ply in NYC.



(a) Census Tract (b) Zip Code (c) Neighborhood

Figure 4. Spatial resolutions that are used for cloaking the locations of the taxi trips.

Real location	12,224	91.0%
Census tract	11,794	87.8%
ZIP code	11,299	84.1%
Neighborhood	11,083	82.5%

Table IV
DE-ANONYMIZING SPATIALLY CLOAKED TAXI DATA USING OUR ATTACK STRATEGY.

C. Spatio-Temporal Cloaking

Given that both temporal and spatial cloaking techniques by themselves are not sufficient to prevent the proposed attack, we now experiment by applying a combination of both to the published data. Table V shows the results from our attack for different combinations of spatial and temporal resolutions.

The results indicate that even when with the coarsest resolution with respect to both time and space, almost 4000 taxis can be uniquely identified. While this is indeed an improvement over higher resolutions, this is still a significant fraction of the total number of taxis present in NYC, thus illustrating the power of our attack.

	1 min	5 min	15 min	30 min
Real location	12,224 91.0%	12,131 90.3%	11,893 88.6%	11,639 86.7%
Census tract	11,794 87.8%	11,217 83.5%	10,927 81.4%	10,134 75.5%
ZIP code	11,299 84.1%	10,527 78.4%	9,141 68.0%	7,376 54.9%
Neighborhood	11,083 82.5%	9,409 70.0%	6,784 50.5%	3,980 29.6%

Table V
RESULTS WHEN AGGREGATING OVER TIME AND SPACE. TOTAL NUMBER OF TAXIS IN 2014: 13,427

VI. DISCUSSIONS

Choice of locations for capturing the medallion data. For our study, we chose to use the medallion data set built on observations made every day from 1 pm to 6 pm during one month. It enabled us to observe 12,248 distinct medallions, out of the 13,000+ taxis in service in NYC. Our goal when building this data was to observe as many distinct taxi cabs as possible while spending the least possible amount of time doing it. The choice of the locations had two advantages: 1) it enabled us to observe a lot of taxis; and 2) these taxis tend pass several times at these locations, so that their sub-trajectories consisted of several points for a majority of the taxis. On average, during our selected time period, a taxi appears 2.28 times at Penn Station, 1.96 at Port Authority Terminal, 2.03 at 9th avenue & 14th street and 2.64 times at Columbus Circle. Figure 5 shows a histogram of the occurrences of the taxis observed during the data capture times at the four locations.

We believe that due to the nature of taxi trips, the location where this data is captured affects the quality of the attack. For example, if we gathered data only at Penn Station during the same time period, then we could still observe 8,659 taxis in the medallion data, and could identify 7,269 of them (without any generalization). By increasing the time period of the data collection to be from 8 am to 6 pm, we could improve the attack and were able to identify 10,560 taxis. Similarly, increasing the data gathering time to over one month will also improve the performance of the attack. This implies that a determined adversary could also manually take notes at just one location to recover the taxi IDs.

But we expect this is due to the fact that a lot of taxi trips begins or ends at Penn Station. To test the effectiveness of the attack if data is captured at other locations, we performed

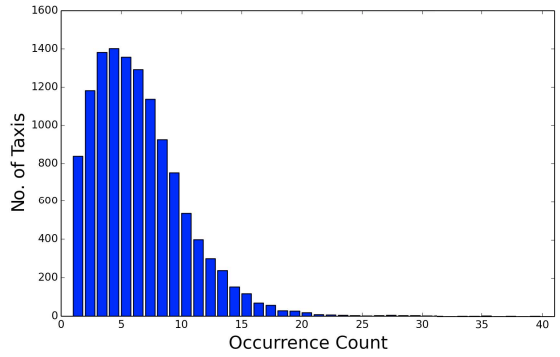


Figure 5. Plot of the frequency of occurrence of taxis at the four locations used for gathering the medallion data.

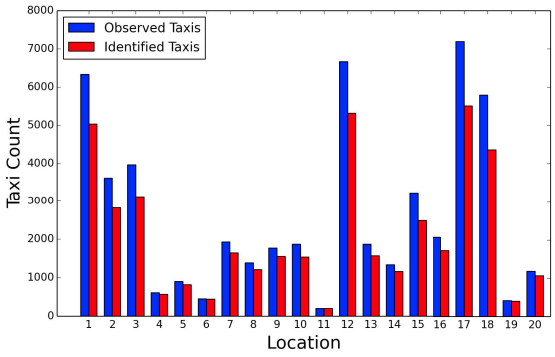


Figure 6. Effectiveness of our attack when the medallion data is gathered from different locations. Note the number of unique taxis present in the medallion data is small in many locations, thus reducing the count of the taxis that can be identified using the attack.

the same experiment after gathering the medallion data from different locations within the same four neighborhoods our initial choice of locations were part of. As shown in Figure 6, the quality of the attack deteriorates if a good location is not chosen.

Table VI compares the average number of taxis identified over the 20 locations against the number of taxis identified when using Penn Station and when different generalization strategies are used. It further reiterates the above observation on the dependence of the effectiveness of our attack on the choice of locations. Again, note that increasing the time period for gathering data will improve the attack.

Other trajectory privacy techniques. Almost all trajectory privacy preserving techniques satisfying legal requirements, including differential privacy [25] and k-anonymity [20], distort the data in some form usually on space. In these techniques, the distortion is usually different for each trip. However, our attack can still be used on such data since the extent of this distortion will also be published. In such a case, we can represent each location as a circle with radius equal to the distortion. Then, when querying for the taxi IDs with respect to a location, we have to query for all circles that intersect with the query location. The rest of the algorithm remains the same. A similar modification to the query step of the attack can again be used when a discrete generalization [29] is used to anonymize the trajectories.

The distortion used in these techniques is usually much smaller than the size of the neighborhoods used in this work. Hence, we expect our attack to be effective when a distortion based technique is used. In case such a distortion becomes large, while it might help reduce the effectiveness of the attack, it results in a significant reduction in the utility of the data set as discussed below.

Affect of anonymity on analytics. One of the main reasons for releasing urban data is to help with the analytics that allows to improve policies and planning in cities. While either distortion of the data through generalization, or hiding information might help with anonymity, these transformations to the data could adversely affect such analytics of the data. For example, Poco et al. [10] used the actual trip locations and times to derive traffic information and study taxi as well as traffic mobility patterns in Manhattan. A decrease in the resolution due to generalization would not allow for such analysis. However, not providing taxi IDs for the trips would not affect such a technique, but it would prevent economists from studying patterns of individual taxis. This is a difficult problem to be able to satisfy requirements of different parties. In the future, we plan to investigate alternative service-based (or query based) models of publishing data which might be able to maintain privacy while also allowing for seamless and accurate analytics.

	1 min	5 min	15 min	30 min
Real location	2126.4 7269	1207.1 5316	778.5 4755	666.1 4537
ZIP code	518.1 4386	150.4 2165	41.8 1011	11.4 425
Neighborhood	318.2 2600	79.3 734	20.6 84	0.25 6

Table VI
AVERAGE NUMBER OF TAXIS IDENTIFIED WHEN USING 20 RANDOM LOCATIONS (TOP) VERSUS THE NUMBER OF TAXIS IDENTIFIED WHEN USING PENN STATION (BOTTOM).

VII. IS ANONYMITY POSSIBLE?

In this paper, we have studied the problem of maintaining privacy of medallion owners and drivers in the NYC taxi data. We have showed that in practice even with a perfect pseudonymization of the taxi IDs, an external data set containing sub-trajectories is indeed sufficient to identify almost all the taxis that ply in NYC. Besides, the TLC which releases the data, must respect several constraints that prevent it from altering or faking the data. We show that, even when using privacy preserving techniques satisfying these constraints to publish the taxi data, our attack can still identify a significant fraction of the taxis, thus demonstrating the difficulty in obtaining a good anonymization of the taxi

data. In order to avoid the above problem, TLC has now removed the medallion and driver license IDs from the newly released data sets. Unfortunately this would adversely impact certain types of analysis on the data.

One of the main reasons we believe our attack was successful is due to the way the set of taxi trips are distributed in NYC. Such privacy issues are common even in other location-based services. For example, people could be tracked using Wifi networks they connect to, or customers could be tracked using their credit card transactions and information from shop owners [21]. However, the spatial distribution in these situations might be different. An interesting future work would be to analyze the properties of spatio-temporal data sets where an external data-based attack will not work.

ACKNOWLEDGMENTS

The authors thank the NYC Taxi and Limousine Commission for providing the data used in this paper and feedback on our results. This work was supported in part by NSF awards CNS-1229185, CI-EN- 1405927 and CNS-1544753, and by the Moore-Sloan Data Science Environment at NYU. Juliana Freire is partially supported by the DARPA Memex program.

REFERENCES

[1] "Chicago Open Data," https://data.cityofchicago.org/.

[2] "NYC Open Data," http://data.ny.gov.

[3] "Seattle Open Data," http://data.seattle.gov.

[4] A. Feuer, "The mayor's geek squad," http://www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html?pagewanted=all\&_r=0, March 2013.

[5] B. Goldstein and L. Dyson, *Beyond Transparency: Open Data and the Future of Civic Innovation*. San Francisco, USA: Code for America Press, 2013.

[6] A. Tockar, "Riding with the stars: Passenger privacy in the nyc taxicab dataset," http://research.neustar.biz/author/atockar/, September 2014.

[7] J. Trotter, "Public nyc taxicab database lets you see how celebrities tip," http://gawker.com/the-public-nyc-taxicab-\\database-that-accidentally-track-1646724546, October 2014.

[8] V. Pandurangan, "On taxis and rainbows lessons from nycs improperly anonymized taxi logs," https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1, June 2014.

[9] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. Silva, "Using topological analysis to support event-guided exploration in urban data," *IEEE TVCG*, vol. 20, no. 12, pp. 2634–2643, 2014.

[10] J. Poco, H. Doraiswamy, H. Vo, J. Comba, J. Freire, and C. Silva, "Exploring traffic dynamics in urban environments using vector-valued functions," *Eurographics Conference on Visualization (EuroVis)*, vol. 34, no. 3, 2015.

[11] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty*, vol. 10, no. 5, pp. 557–570, 2002.

[12] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 3, 2007.

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," *ICDE*, 2007.

[14] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," *Proceedings of the international conference on pervasive services*, p. 8897, 2005.

[15] B. Palanisamy and L. Liu, "Mobimix: protecting location privacy with mix-zones over road networks," *Proceedings of the 27th IEEE international conference on data engineering*, p. 494505, 2011.

[16] C. Dwork, "Differential privacy," *Proc. of ICALP*, vol. 4052 of LNCS, pp. 1–12, 2006.

[17] M. E. Andres, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," *ArXiv e-prints*, 2012.

[18] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," *Proceedings of the 24th International Conference on Data Engineering*, pp. 376–385, 2008.

[19] C.-Y. Chow and M. F. Mokbel, "Trajectory privacy in location-based services and data publication," *SIGKDD Explorations*, vol. 13, no. 1, pp. 19–29, 2011.

[20] R. Trujillo-Rasua and J. Domingo-Ferrer, "On the privacy offered by (k,d)-anonymity," *Inform. Syst.*, vol. 38, no. 4, p. 491494, 2013.

[21] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," *Proceedings of the 9th International Conference on Mobile Data Management*, pp. 65–72, 2008.

[22] C. Bettini, X. S. Wang, and S. Jajodia, "Protecting privacy against location-based personal identification," *Proc. of VLDB Workshop on Secure Data Management (SDM)*, p. 185199, 2005.

[23] P. Sui, T. Wo, Z. Wen, and X. Li, "Privacy risks in publication of taxi gps data," *Proceedings of the IEEE 6th International Symposium on Cyberspace Safety and Security*, pp. 1189 – 1196, 2014.

[24] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel, "Movement data anonymity through generalization," *Transactions on Data Privacy*, vol. 3, pp. 91–121, 2010.

[25] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: a case study on the montreal transportation system," *KDDM, KDD 12*, p. 213221, 2012.

[26] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," *Proc. of USENIX MobiSys*, pp. 31–42, 2003.

[27] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The new casper: Query processing for location services without compromising privacy," *Proc. of VLDB*, 2006.

[28] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," *Proc. of ICDCS. IEEE Computer Society*, 2005.

[29] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-Divanis, "Distance-based km-anonymization of trajectory data," *MDM*, 2013.

[30] ——, "Select-organize-anonymize: A framework for trajectory data anonymization," *13th IEEE International Conference on Data Mining Workshops*, p. 867874, 2013.

[31] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, 2008, pp. 111–125.

[32] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08, 2008, pp. 265–273.

[33] D. Kifer, "Attacks on privacy and definetti's theorem," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '09, 2009, pp. 127–138.

[34] R. C.-W. Wong, A. W.-C. Fu, K. Wang, P. S. Yu, and J. Pei, "Can the utility of anonymized data be used for privacy breaches?" *ACM Trans. Knowl. Discov. Data*, vol. 5, no. 3, pp. 16:1–16:24, Aug. 2011.

[35] NYC Taxi & Limousine Commission, "2014 taxicab fact book," http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf, 2014.

[36] N. Ferreira, J. Poco, H. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, p. 21492158, 2013.

[37] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Personal and Ubiquitous Computing*, vol. 18, no. 1, pp. 163–175, 2014.