Research Article

Privacy Protection Method for Vehicle Trajectory Based on VLPR Data

Hua Chen,1,2 Chen Xiong,1,2 Jia-meng Xie,3 and Ming Cai1,2

1School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510275, China
2Guangdong Provincial Key Laboratory of Intelligent Transportation System, School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510275, China
3Traffic Administration Bureau of Guangdong Province, Guangzhou 510440, China

Correspondence should be addressed to Chen Xiong; xiongch8@mail.sysu.edu.cn and Ming Cai; caiming@mail.sysu.edu.cn

Received 23 July 2019; Revised 27 May 2020; Accepted 6 June 2020; Published 23 June 2020

Copyright © 2020 Hua Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of data acquisition technology, data acquisition departments can collect increasingly more data. Various data from government agencies are gradually becoming available to the public, including license plate recognition (VLPR) data. As a result, privacy protection is becoming increasingly significant. In this paper, an adversary model based on passing time, color, type, and brand of VLPR data is proposed. Through experimental analysis, the tracking probability of a vehicle’s trajectory can be more than 94% if utilizing the original data. To decrease the tracking probability, a novel approach called the \((m, n)\)-bucket model based on time series is proposed since previous works, such as those using generalization and bucketization models, cannot deal with data with multiple sensitive attributes (SAs) or data with time correlations. Meanwhile, a mathematical model is established to expound the privacy protection principle of the \((m, n)\)-bucket model. By comparing the average calculated linking probability of all individuals and the actual linking probability, it is shown that the mathematical model that is proposed can well expound the privacy protection principle of the \((m, n)\)-bucket model. Extensive experiments confirm that our technique can effectively prevent trajectory privacy disclosures.

1. Introduction

With the coming of the big data era, big data are increasingly connected with our lives. The development of location-aware technology, such as mobile communications and sensing, digitizes the geographical locations of people and objects. Generalized data have the characteristics of being large scale, fast updating, and high value, which has brought great changes to people’s lives, business operations, and scientific research. Trajectory data are the sequence of location information from a moving object that is recorded in chronological order.

There are two main applications for trajectory data: one for online applications and the other for offline applications. The former generally refers to location-based services (LBSs) in which location service providers can provide customized services by obtaining the real-time location information of mobile objects. For example, when a customer searches for a destination using a map application, the service provider can provide the customer with a reasonable path based on the customer’s current trajectory. The latter refers to the location service provider’s desire to discover the intrinsic value of a user’s trajectory so that the trajectory data can be directly acquired and analyzed or made available to third parties. Traditional car insurance companies, for example, set insurance prices based on the average risk of vehicles, but as user customization needs increase, companies have begun to rely on travel time, public locations, travel processes, and other information in data organizations to provide customized services to users [1, 2]. However, while location data facilitates people, it also increases the risk of personal privacy disclosure because the data contain not only the user’s location information but also the user’s home address, lifestyle, health status, social relations, behavioral patterns, and
other implicit information. For example, Gambs et al. use GPS data to successfully infer a person’s home address and company information [3]. Matsuo et al. use position sensor data to build user feature matrices, classify data using machine learning techniques, and accurately infer individuals’ work, age, and smoking habits [4]. Song et al. use personal trajectory data to explore people’s past, present, and future positions with 93% accuracy [5]. Using telephone call data from a European country and analyzing it based on random space-time points, Montjoye et al. found that 95% of data from a European country and analyzing it based on future positions with 93% accuracy [5]. Using telephone call data from a European country and analyzing it based on random space-time points, Montjoye et al. found that 95% of users could be accurately identified by randomly selecting 4 space-time points [6]. These examples show that privacy protection is a critical issue.

Trajectory privacy is defined as not being traceable as moving objects at a certain distance or time. Trajectory privacy preservation mainly focuses on LBS-based privacy protection and the privacy protection related to released data; the related work is described below. Olurin et al. believe that the privacy preference platform (P3P) and P3P-based technologies have great developmental potential; therefore, they summarize the research on P3Ps and explain that the current challenge is to design a formalized privacy policy language [7]. For smartphone query privacy, Shokri et al. propose a user collaboration privacy protection method based on LBSs, which saves user query information in a buffer and passes it to other people who search for such information. Simulation results show that this method is lightweight and has low collaboration costs [8]. Khoshgozaran et al. desired strong privacy and proposed a PIR-based framework that processes server answered client queries without inferring any information. Experiments that were conducted for real-time networks have validated the usefulness of the scheme, and the existing challenge is how to reduce the space and time costs [9]. Terrovitis and Mamoulis show that there is a high risk of privacy breaches by adversaries if the trajectories are published during processing that only hides the identities of users. They develop a data suppression technique that iteratively suppresses selected locations from the original trajectory until privacy constraints are met to prevent this type of breach [10]. Based on the k-anonymity model, Nergiz et al. propose a method based on generalization to solve the problem of individual recognition in static trajectory datasets. They redefine the notion of k-anonymity for sequences of space-time points and further enhance privacy protection by releasing only a randomly generated set of representative trajectories. Experiments on real and synthetic data show that the method is effective for publishing trajectory datasets [11]. Sun et al. deem that a privacy mechanism should be consciously and consistently integrated in system designs, and, on this basis, the virtual travel line model (VTL) is proposed. The VTL can be used to standardize data when and where mobile data are collected to meet the needs of traffic modeling and privacy protection. Traffic-knowledge-based adversary models are developed to evaluate the performance of the VTL zone method and filtering algorithms. The algorithm is evaluated in terms of its privacy protection and data demands to meet traffic applications [12, 13]. The above studies show that trajectory privacy has become a key issue in trajectory data publishing. VLPR data, as a kind of trajectory data, face the same problem.

With the development of intelligent transportation, increasingly more urban service organizations hope to be able to obtain VLPR data from traffic management departments to provide better urban service solutions [14]. However, even if the vehicle license plate numbers have been replaced by qualified identifier (QI) before the data are released, the trajectories are still unique relative to their QIs. Previous works [15] have shown that adversaries can still connect trajectories to users using background knowledge. Gao et al. find that even when the temporal granularity is set to be half a day, five spatiotemporal records are sufficient to uniquely identify approximately 90% of individuals using VLPR data. Besides, in a dataset, an adversary can easily get a unique trajectory in a time interval \([t_1, t_2]\) when the point of interest (POI) information for the trajectory is open to the public; thus, if the adversary has the background knowledge that Mike once visited the POI in the same time interval \([t_1, t_2]\), then the adversary has a reason to infer that the trajectory comes from Mike. In this situation, traffic management departments decide to disclose the VLPR data without the vehicle license plate numbers in case the trajectory is linked and thus the privacy information is violated. However, in our research, it is found that the privacy adversary can still use different kinds of methods to acquire the unique vehicle trajectories from incomplete data. If the VLPR data are released without processing, more than 90% of the vehicle trajectories between the upstream and downstream VLPRs can be successfully linked. To reduce the probability of the successful linkages of vehicle trajectories, the \((m, n)\)-bucket model based on time series is proposed, and the results can provide theoretical suggestions for relevant institutions when releasing VLPR data.

The rest of this paper is organized as follows. The notations used throughout this paper are summarized in Section 2. Section 3 offers a brief introduction of the VLPR system. An adversary model based on VLPR data is proposed in Section 4. A VLPR data publishing method, namely, the \((m, n)\)-bucket model based on time series, is proposed in Section 5. The criteria that are used to evaluate the performances of the \((m, n)\)-bucket model and address the data requirements for privacy preservation are discussed in Section 6. Some case studies are carried out to evaluate the effectiveness of the \((m, n)\)-bucket model by making privacy attacks in Section 7. Some conclusions are summarized in Section 8.

2. Notations

We use the notations listed in Table 1 throughout this paper.

3. VLPR System

The VLPR is a system that adopts advanced photoelectric technology, image processing technology, and pattern recognition technology to take an image of every vehicle passing by and automatically identify the license plate of the
vehicle. Today, VLPR systems are ubiquitous in cities, and they are usually placed a few meters away from the road. When a vehicle is detected at a distance in front of the device, the system begins to capture the vehicle’s image. When a vehicle is detected at a distance in front of the device, the system begins to capture the vehicle’s image. Today, VLPR systems are ubiquitous in cities, and they are usually placed a few meters away from the road. When a vehicle is detected at a distance in front of the device, the system begins to capture the vehicle’s image. vehicle. When a vehicle is detected at a distance in front of the device, the system begins to capture the vehicle’s image. vehicle. Today, VLPR systems are ubiquitous in cities, and they are usually placed a few meters away from the road. When a vehicle is detected at a distance in front of the device, the system begins to capture the vehicle’s image. To link the traces from two VLPRs, an adversary model based on the passing time, color, type, and brand of vehicles is proposed. Even though this adversary model only considers three QI attributes of the vehicle other than passing time, the successful linking rate is impressive (see Section 6). In this part, the adversary model is illustrated in detail. Two cases are given.

As shown in Table 2, the vehicle license plate number is directly associated with user privacy, which is classified as SA information. The other information consists of additional descriptions of the vehicle passing through the device, which are called QI attributes. It is a common way to encrypt or encode SA attributes before releasing data so that adversaries do not have direct access to owner information, but in this situation, the vehicle trajectory can be easily obtained. Therefore, the data management department will choose to suppress the SA attributes; however, based on our research, there are still ways to breach trajectory privacy by using only QI attributes, which increases the likelihood for an adversary to breach a vehicle owner’s privacy. To the best of our knowledge, related work on this topic has not yet been reported.

### Table 1: Summary of notations.

| Notations | Description |
|-----------|-------------|
| QI<sub>i</sub> | The ith quasi-identifier of vehicle |
| SA<sub>j</sub> | The jth sensitive attribute of vehicle |
| T<sub>n</sub> | Departure time of vehicle n |
| T<sub>n</sub> | Estimated arrival time of vehicle n |
| E(x) | Expected travel time |
| S<sub>n</sub> | A vehicle set of VLPR 2 |
| T<sub>n</sub> | Time for vehicle m in S<sub>n</sub> |
| T<sub>i</sub>, T<sub>j</sub> | Time that vehicle i is at VLPR 1 and VLPR 2, respectively |
| t<sub>i</sub>, t<sub>j</sub> | Vehicle i at VLPR 1 and VLPR 2, respectively |
| S<sub>n</sub> | Suspect list |
| MB | Vehicle k in the suspect list S<sub>n</sub> |
| T | Predicted vehicles |
| T | VLPR data table |
| T | Vehicle i in T |
| t<sub>i</sub> (i = 1, . . . e) | Bucket i in buckets |
| B<sub>i</sub> (i = 1, . . . b) | Vehicle t with QI, value |
| t<sub>i</sub> [SA<sub>j</sub>] | Bucket b with QI, value |
| B[SA<sub>j</sub>] | Vehicle t with SA<sub>j</sub>, value |
| p(t, B) | Bucket b with SA<sub>j</sub>, value |
| SA<sub>j</sub> (v) | Probability that tuple t is in bucket B |
| p(t, SA<sub>j</sub>>(v)) | Represents the value of the specified column SA<sub>j</sub> being shuffled |
| p(t, SA<sub>j</sub>>(v)) | The probability that t contains SA<sub>j</sub> (v) |
| p(t, SA<sub>j</sub>>(v)) | Represent the probability that t contains SA<sub>j</sub> (v) given that t is in bucket B |
| [MB[SA<sub>j</sub>>(v)] | The number of values v that are contained in the SA<sub>j</sub> in MB |
| s_num | The number of tuples with the same QI<sub>i</sub> |
| p_max_onemb | The maximum probability of one matching bucket |
| P_one | The data structure that stores all the p_max_onemb of one tuple |
| p_all_onemb | The sum of all the p_all_onemb in P_one |
| Pr | Actual linking probability |
| Pc | Total calculated probability |

### 4. Adversary Model

To link the traces from two VLPRs, an adversary model based on the passing time, color, type, and brand of vehicles is proposed. Even though this adversary model only considers three QI attributes of the vehicle other than passing time, the successful linking rate is impressive (see Section 6). In this part, the adversary model is illustrated in detail. Two cases are given.

#### 4.1. Trajectory Linking between Two Neighboring VLPRs without an Intersection (Case 1)

Consider two neighboring VLPRs where there is no entrance or exit between them, which is the simplest case, as shown in Figure 1. Suppose that the travel time from VLPR 1 to VLPR 2 is relatively stable.

Define the time when vehicles are at VLPR 1 as the departure time (denoted as Ω<sub>1</sub>) and the time when vehicles are at VLPR 2 as the arrival time (denoted as Ω<sub>2</sub>). By observing the departure and arrival times from the VLPR system of Shenzhen highway in Foshan, China, it is found that they almost satisfy a log-normal distribution, as shown in Figure 2. In most studies [13, 16, 17], the estimated travel time is estimated using the length of the road divided by the limited speed on the road. However, the traffic flow is different at different times, resulting in different average...
Table 2: VLPR data.

| Index | Passing time       | Vehicle license plate number | Vehicle type | Vehicle color | Vehicle brand | VLRP longitude |
|-------|--------------------|-----------------------------|--------------|---------------|---------------|---------------|
| 1     | 2018-07-30 13:08:01| adheddd324dadf5             | Small        | Red           | Nissan        | 114.71        |
| 2     | 2018-07-30 13:09:01| adekdd433da6               | Middle       | Blue          | Toyota        | 113.62        |

In this paper, if the vehicle \( n \in \Omega_1 \) and the vehicle \( m \in \Omega_2 \) are released, \( T_m^{a,p} \) is used to represent the estimated arrival time of vehicle \( n \). This is shown in the following equation, where \( E(x) \) is the expected travel time and \( T_1^a \) is the departure time of vehicle \( n \):

\[
T_m^{a,p} = T_1^a + E(x) = T_1^a + \exp \left[ \mu + \frac{\delta^2}{2} \right]. \tag{2}
\]

However, it is impossible for each vehicle to travel exactly as predicted. Therefore, the estimated time should be predicted by using a threshold \( T_i \). While \( T_i = \text{rate} \times E(x) \), rate is a constant. That is, when a trace \( n \in \Omega_i \), the estimated arrival time can be predicted within the time period \([T_1^{a,p} - T_i, T_1^{a,p} + T_i]\). In other words, a suspect list (denoted as \( S_n \)) can be obtained for VLPR 2 in the time period. If \( S_n \) is not empty, when \( S_n \) only contains one vehicle \( m \), it can be said that \( m \) is the same vehicle as \( n \). However, \( S_n \) usually contains a number of vehicles, which indicates that certain rules should be employed to confirm the target. To determine which vehicle in \( S_n \) is \( n \), the other attributes of the VLPR data are used, which have been illustrated in detail in Section 2. In this paper, in addition to time, only three QI attributes (type, color, and brand) of the VLPR data are considered to recognize which vehicle at VLPR 2 is the same vehicle as that at VLPR 1. In this adversary model, the QI attributes from the VLPR data are utilized to reduce the number of suspected vehicles. Obviously, the more the QI attributes that are utilized, the fewer the suspicious objects that will remain. In the case that the filtered suspect list still contains more than one vehicle, the closest time rule is used to find the predicted vehicle, which satisfies:

\[
m = \arg \min_{m \in S^n} \left| T_m^{a,p} - T_2^{a,p} \right|, \quad m \in S^n,
\]

where \( T_m^{a,p} \) is the time for vehicle \( m \) in \( S_m \) and \( S^n \) is a vehicle set of VLPR 2. The previous equation indicates that vehicles \( m \) and \( n \) are the same vehicle since the arrival time of \( m \) is the closest to \( n \). That is, the vehicle trajectories at the two VLPRs are successfully linked, which will lead to a privacy disclosure.

Algorithm 1 describes the process of trajectory linking in detail.

**Definition 1.** \( T_i^a \) and \( T_1^a \) represent the time that vehicle \( i \) is at VLPR 1 and VLPR 2, respectively. In addition, we also have the color (referred to as \( C_1^i \) and \( C_2^i \)), type (referred to as \( T_y^i \) and \( T_y^{a,p} \)), and brand (referred to as \( B_{ri}^i \) and \( B_{ri}^{a,p} \)) of the vehicle. \( t_i^1 \) and \( t_i^2 \) denote vehicle \( i \) at VLPR 1 and VLPR 2, respectively. \( x_i^k \) represents vehicle \( k \) in the suspect list \( S_m \). \( S_m \) is the output of the algorithm, namely, the predicted vehicles.

4.2. Trajectory Linking with an Intersection Situation (Case 2). The road topology is complicated in real situations. Vehicles that arrive at one VLPR usually come from different VLPRs (e.g., VLPR 1, VLPR 2, VLPR 3, and VLPR 4). The vehicles
that arrive at VLPR 1 may come from VLPR 2, VLPR 3, and VLPR 4 with the same probability, and the adversary wants to link the vehicle traces from VLPR 2, VLPR 3, and VLPR 4 to those in VLPR 1, as shown in Figure 3. That is, the adversary wants to know what upstream VLPR each vehicle in VLPR 1 comes from. The travel time for each VLPR is estimated using Algorithm 1. Therefore, the records in the same equivalence group are increasingly easier for adversaries to link traces between multiple VLPRs. In the near future. If all vehicle traces are released without processing, as shown in Section 6, a majority of the traces can be linked by using specific adversary models. Therefore, a certain data processing approach is needed to process the data that are collected in VLPR systems to ensure the privacy of travelers.

5.2. (m, n)-Bucket Model Based on Time Series

5.2.1. Related Work. The application of generalization to protect privacy disclosures has been very well studied in the literature [19–21]. This approach consists of generalizing a dataset to a certain extent by transforming the QI values into more general forms and dividing tuples into equivalent groups, where the values of each QI attribute are the same. Therefore, the records in the same equivalence group are indistinguishable, which meets k-anonymity [22–24] and l-diversity [25, 26]. Although generalization is a common method for privacy preservation, it makes the data lose much information and fails with high-dimensional data, which is not conducive to the effective analysis of data. To overcome the drawbacks of generalization, a bucketization method called anatomy, which protects both the privacy and correlation in the microdata, is proposed [27]. Bucketization first divides the tuples in the table into buckets and then randomly separates the quasi-identifiers (QIs) with sensitive properties by traversing the SA values in each bucket. Although bucketization can be used to deal with high-dimensional data, it also tends to disclose more information because all the QIs are released. To overcome the limitations

---

### Algorithm 1: Trajectory linking process in case 1.

**Input:** A set of released traces $(\Omega_1)$ and traces $(\Omega_2)$ that pass through VLPR 1 and VLPR 2, respectively

**Output:** Predicted vehicles $(S_m)$

1. for $t_1 \in \Omega_1$
2. for $t_2 \in \Omega_2$
3. if $T_2 \in [T_1 - T_r, T_1 + T_r]$
4. $S_n \leftarrow T_1 \cup T_2$
5. end if
6. end for
7. for $t_h \in S_n$
8. if $C_i = C_i^h, T_i = T_i^h, Br_i = Br_i^h$ and $T_i^h$ is closest to $T_1$
9. $S_m \leftarrow S_m \cup T_i^h$
10. end if
11. end for
12. end for

5.1. Publishing the Original Data. Publishing the original data refers to releasing all the data that are collected in a VLPR, which provides the least privacy protection but the most usable data. If all data that are collected in VLPR systems are released, then when VLPR systems’ layouts are dense enough and the data quality is good enough, it is easy for an adversary to link the traces between the upstream and downstream VLPRs. In fact, due to the complexity of the road network, it is not a simple process for adversaries to keep track of the same vehicle between several intersections. However, with the development of smart cities, road network designs have become increasingly more standardized, VLPR systems’ layouts have become increasingly more intensive, and the data that are collected have become increasingly more accurate; therefore, it will become increasingly easier for adversaries to link traces between multiple VLPRs in the near future. If all vehicle traces are released without processing, as shown in Section 6, a majority of the traces can be linked by using specific adversary models. Therefore, a certain data processing approach is needed to process the data that are collected in VLPR systems to ensure the privacy of travelers.
of generalization and bucketization to get better privacy preservation and data utility, the slicing algorithm was proposed, which can prevent attribute disclosure and membership disclosure while releasing data [28]. However, slicing has been found to risk the disclosure of some absolute facts, which would help adversaries find invalid records in released datasets, resulting in personal privacy violations [29]. To improve the confidentiality of the published data in a slice table, a method called value swapping is proposed. Its purpose is to decrease the attribute disclosure risk for absolute facts and ensure l-diverse slicing. To meet the different requirements for the protection of identities and sensitive values, an approach called cross-bucket generalization is presented [30]. This approach can meet the requirements by providing higher attribute protection than identity protection and is proven to be useful for privacy preservation.
Besides, there are other methods to replace sensitive attributes with anonymous characters independently at random [31] or add randomized records to the original database [32]. These methods usually can achieve good privacy protection. However, the former causes a great deal of information loss, and the latter adds a lot of information. This is unacceptable in the field of big data. In order to achieve the effect of privacy protection and keep the original data as much as possible, we proposed the algorithm that shuffles the sensitive attribute values in the same bucket. This method neither adds data nor decreases data but just shuffles sensitive data in the same bucket.

All the related works that are mentioned above are not suitable for vehicle trajectory privacy preservation because existing privacy protection methods are applicable to data without time correlation. Therefore, in order to meet the k-anonymity or l-diversity requirement, the order of the data in the table can be adjusted at will. However, the VLPR data have strong time correlations, and so the positions of the data in the table cannot be adjusted at will. In conclusion, none of the above privacy protection methods are applicable to vehicle trajectory privacy protection. Therefore, based on the above research, the (m, n)-bucket model based on time series is proposed in this paper.

5.2.2. Formalization of (m, n)-Bucket Model. Define T as the VLPR data that is to be published. T contains d quasi-identifier (QI) attributes QI = {QI1, QI2, . . . , QId} and e tuples T = {t1, t2, . . . , te}. Each QIi (1 ≤ i ≤ d) can be categorical in the VLPR dataset.

Definition 1 (Bucket). A tuple partition consists of several subsets of T such that each tuple belongs to only one subset. These subsets are referred to as buckets, and we denote them as B1, B2, . . . , Bn. Then, ∪ni=1 Bi = T and, for any 1 ≤ i1 ≠ i2 ≤ b, B1 ∩ B2 = Ø.

Definition 2 ((m, n)-bucket). T is sorted according to time series and is divided into several buckets where each bucket contains m tuples. After dividing T into several buckets, the values of the specified columns are shuffled inside the bucket, and n is the number of columns being shuffled.

Definition 3 (Matching Bucket). Let {QI1, QI2, . . . , QId} be the d columns of a VLPR table. Let t be a tuple and t [QI] be the QIi value of t. B is denoted as a bucket in the VLPR table, and B[QIi] is denoted as the multiset of QIi values in B. Let {SA1, SA2, . . . , SAN} be the sensitive columns that need to be shuffled, where n represents the number of sensitive columns. Let {QI11, QI12, . . . , QId} be the nonsensitive QI columns that do not need to be shuffled. Identifying which column is the sensitive column is not the focus of the research of this paper. For any tuple t in the VLPR 1 table, if for 1 ≤ j ≤ n we can get t[SAj] ∈ B[SAj], and for n + 1 ≤ i ≤ d we can get t[QIi] ∈ B[QIi], and the value in different nonsensitive QI columns are bound together; then, B is denoted as a matching bucket (MB).

For example, in Table 3, when t2 belongs to VLPR 1, t2 = (purple, large, Toyota). When the QI brand column is the only sensitive column that needs to be shuffled, the set of MB for t2 is {B1, B2}. If the QI brand and type are the sensitive columns, the set of MB for t2 is {B1, B2, B3}.

5.2.3. Privacy Preservation. In this section, the protection principle of the (m, n)-bucket model for vehicle identities is analyzed in detail. Generally, a tuple t may have multiple MBs. Consider an adversary who knows all the QI values of t in VLPR 1 and attempts to infer t’s SAj (v) in VLPR 2 to link the trajectories according to time inference, where SAj (v) represents the value of the specified column SAj being shuffled. The adversary needs to determine where buckets t might reside. Tuple t may be in any one of its MBs. Let p(t, B) be the probability that tuple t is in bucket B. It is worth mentioning that the calculation of p(t, B) only takes the nonsensitive QIs into account. For instance, in Table 3, if the brand column is the only sensitive column, then B1 and B2 are the MBs of t2, and B1 is not because the nonsensitive type and color columns must be bound together.

In the second step, the adversary computes p(t, SAj (v)), which is the probability that t contains SAj (v). In other words, it is the probability that t is successfully identified. Specifically, let p(SAj (v) | t, B) represent the probability that t contains SAj (v) given that t is in bucket B. Then, the probability p(t, SAj (v)) can be computed as

\[
p(t, SAj (v)) = \sum_B p(t, B) p(SAj (v) | t, B).
\]

In a VLPR table T, when the (m, n)-bucket model is used for any individual t, \(p(t, SAj_1(v), SAj_2(v), . . . , SAj_n(v))\) meets the following formula:

\[
p(t, SA_j (v), SA_2 (v), . . . , SA_n (v)) \leq \sum_{\text{MB}} p(t, \text{MB}) \cdot \frac{\text{MB}[SA_1(v)] \cdot \text{MB}[SA_2(v)] \cdots \text{MB}[SA_n(v)]}{m^n}
\]

Here, \(v^{'}\) denotes the most frequent value in SAj, and \(|\text{MB}[SA_j(v')]|\) is the number of values v that are contained in the SAj in MB. For example, in Table 3 bucket 1, \(v^{'}\) is Toyota and \(|\text{MB}[SA_j(v')]|\) is 2. In formula (5), m is the number of t’s in every bucket, and n is the number of sensitive columns.

Proof. In formula (4), if \(|SA_j| \notin B[SA_j] (1 \leq i \leq d - n)\), then
In the previous formula, \( d \) and \( n \) represent the number of QI attributes and sensitive attributes, respectively. Therefore, formula (4) is converted into the following formula:

\[
p(t, B) = 0.\tag{6}
\]

For all buckets, \( v \) is always less than \( v' \); then,

\[
|MB[SA_i(v)]| \leq |MB[SA_i(v')]|.\tag{7}
\]

Hence,

\[
p(SA_i(v)|t, MB) = \frac{|MB[SA_i(v)]|}{m} \leq \frac{|MB[SA_i(v')]|}{m},
\]

\[
p(t, SA_i(v)) \leq \sum_{MB} p(t, MB)p(SA_i(v)|t, MB),\tag{8}
\]

\[
|MB[SA_i(v)]| \leq |MB[SA_i(v')]|.\tag{9}
\]

If \( n = 1 \) and \( SA_1 \) is the sensitive column, then you obtain

\[
p(t, SA_1(v)) = \sum_{MB} p(t, MB)p(SA_i(v)|t, MB),\tag{10}
\]

\[
|MB[SA_1(v)]| \leq |MB[SA_1(v')]|.\tag{11}
\]

\[
p(t, SA_1(v), SA_2(v)) = \sum_{MB} p(t, MB)p(SA_1(v)|t, MB)p(SA_2(v)|t, MB, SA_1(v)),
\]

\[
\left\{ \begin{array}{l}
|MB[SA_1(v)]| \leq |MB[SA_1(v')]|, \\
|MB[SA_2(v)]| \leq |MB[SA_2(v')]|.
\end{array} \right. \tag{12}
\]

\[
Hence,
\]

\[
\left\{ \begin{array}{l}
p(SA_1(v)|t, MB) = \frac{|MB[SA_1(v)]|}{m} \leq \frac{|MB[SA_1(v')]|}{m}, \\
p(SA_2(v)|t, MB, SA_1(v)) = \frac{|MB[SA_2(v)]|}{m} \leq \frac{|MB[SA_2(v')]|}{m}.
\end{array} \right. \tag{13}
\]

\[
p(t, SA_1(v), SA_2(v)) = \sum_{MB} p(t, MB)\frac{|MB[SA_1(v)]|}{m} \frac{|MB[SA_2(v)]|}{m} \leq \sum_{MB} p(t, MB)\frac{|MB[SA_1(v)]|}{m^2} \cdot \frac{|MB[SA_2(v)]|}{m},
\]

\[
\left\{ \begin{array}{l}
p(SA_1(v)|t, MB) = \frac{|MB[SA_1(v)]|}{m} \leq \frac{|MB[SA_1(v')]|}{m}, \\
p(SA_2(v)|t, MB, SA_1(v)) = \frac{|MB[SA_2(v)]|}{m} \leq \frac{|MB[SA_2(v')]|}{m}.
\end{array} \right. \tag{14}
\]
Therefore, when \( n = 2 \), formula (5) is true. Suppose that when \( n = k \) and \( S_{A_1}, S_{A_2}, \ldots, S_{A_k} \) are the sensitive columns, the following formula (13) is true.

\[
p(t, S_{A_1}(v), S_{A_2}(v), \ldots, S_{A_k}(v)) = \sum_{MB} p(t, MB) \left[ MB[S_{A_1}(v)] \right] \left| MB[S_{A_2}(v)] \right| \ldots \left| MB[S_{A_k}(v)] \right| m_k
\]

(13)

When \( n = k + 1 \) and \( S_{A_1}, S_{A_2}, \ldots, S_{A_k}, S_{A_{k+1}} \) are the sensitive columns, you obtain

\[
p(t, S_{A_1}(v), S_{A_2}(v), \ldots, S_{A_k}(v), S_{A_{k+1}}(v)) = \sum_{MB} p(t, MB) \left[ MB[S_{A_1}(v')] \right] \left| MB[S_{A_2}(v')] \right| \ldots \left| MB[S_{A_k}(v')] \right| \left| MB[S_{A_{k+1}}(v')] \right| m_k
\]

(14)

Because \((|MB[S_{A_{k+1}}(v')]|/m) \leq (|MB[S_{A_{k+1}}(v')]|)/m\),
you obtain

\[
p(t, S_{A_1}(v), S_{A_2}(v), \ldots, S_{A_k}(v), S_{A_{k+1}}(v)) \leq \sum_{MB} p(t, MB) \left[ MB[S_{A_1}(v')] \right] \left| MB[S_{A_2}(v')] \right| \ldots \left| MB[S_{A_k}(v')] \right| \left| MB[S_{A_{k+1}}(v')] \right| m_k^{k+1}
\]

(15)

In summary, formula (5) is proven to be correct. \( \Box \)

5.2.4. Analysis. In the \((m, n)\)-bucket model, under the condition of a constant \( m \), let \( p' = (|MB[S_{A_i}(v')]|/m) \leq 1 \). This means that the greater the value of \( n \), the lower the probability of \( t \) being successfully identified. Under the condition of a constant \( n \), in general, the greater the value of \( m \), the greater the probability of \( p' \) decreases. When \( m \) increases by 1, only if the added value is equal to \( S_{A_i}(v')p' \) increases; otherwise, it decreases. Since \( S_{A_i} \) usually contains many different values, the probability of it being not equal to \( S_{A_i}(v') \) is always greater than the probability that the added value is equal to \( S_{A_i}(v') \).

5.2.5. \((m, n)\)-Bucket Algorithm. This section presents an algorithm to implement the \((m, n)\)-bucket model. It calculates the maximum linking probability of each individual. The procedure is presented in Algorithm 2.

The variable \( t_j \) denotes the nonsensitive columns of vehicle \( t_j \) (line 5). The data structures \( T_{suspect} \) and Buckets store the suspect tuples and suspect buckets, respectively (lines 6 and 7). \( S_{A_n} \) denotes the sensitive columns (line 9).

MB stores the matching buckets (line 10). The variable \( s_{num} \) is the number of tuples with the same QI as \( t_i \) (line 12). The variable \( p_{max\_onemb} \) is the maximum probability of one matching bucket (line 16). The data structure \( P_{one} \) stores all the \( p_{max\_onemb} \) of one tuple, and the variable \( p_{all\_one} \) is the sum of all the \( p_{all\_onemb} \) in \( P_{one} \). The data structure \( P_{all\_stall} \) stores all the \( p_{all\_onemb} \).

For example, in Table 3, suppose that the times of all tuples meet the requirements. That is, for \( t_j \) in Table 3, \( t_j \) is at \( [t_i, t_i + \Delta t] \) \( E(x) - T_s, t_i + \Delta t + E(x) + T_s \). The adversary gets one tuple \( t_i = \) (purple, large, Toyota) from VLP 1. The maximum probability of \( t_i \) that the adversary can guess from Table 3 is calculated as follows. Suppose that the brand and type columns are the sensitive columns, where \( S_{A_{n_b}} \) is the brand and type column and QI\(_n\) is the color column. The tuple \( t_i \) denotes \( t_{2}, t_{11}, \) and \( t_{11} \). In other words, the tuples \( t_{2}, t_{11}, t_{11}, \) and \( t_{11} \) are stored in \( T_{suspect} \). The tuples in buckets 1, 2, 3, and 4 are added into Buckets. We can get that bucket 4 is not the matching bucket because the brand column does not contain “Toyota.” Therefore, MB stores buckets 1, 2, and 3. The variable \( s_{num} \) is the number of “purple” responses in MB, namely, \( s_{num} \) = 4. In bucket 1, the most frequent value in the brand column is “Toyota,” and
To evaluate the performance of the (m, n)-bucket model, data managers can release data with different ms and ns by using the (m, n)-bucket model that is proposed in this paper. In other words, high-priority users can access data with smaller ms and ns while low-priority users can access data with larger ms and ns.

7. Experiment

7.1. Experiments for Case 1 and Case 2. In this section, the proposed adversary is used to evaluate the effectiveness of the (m, n)-bucket model by making privacy attacks. The evaluation is done using VLPR data that were collected at Shenluo highway (Case 1) and the intersection of Songcheng road (Case 2), in Guangdong Province, China. The datasets have 3 discrete attributes as summarized in Table 4. The road network is located in Foshan, China. Shenluo highway (Case 1) and the intersection of Songcheng road (Case 2) are used for traffic modeling or other research purposes, such as pollution and people’s preferences for vehicle colors and brands. According to the needs or priorities of different users, data managers can release data with different ms and ns by using the (m, n)-bucket model that is proposed in this paper. In other words, high-priority users can access data with smaller ms and ns while low-priority users can access data with larger ms and ns.

Algorithm 2: (m, n)-Bucket algorithm.
Suppose that the original VLPR data are released and the proposed adversary model is used to track the vehicle trajectories between the upstream and downstream VLPRs with different thresholds. Figure 6 illustrates the change of $p$ with the different rates, where the $x$-coordinate is the rate, and the $y$-coordinate is the tracking probability. Figure 6(a) shows the performance in Case 1 from 10:00 a.m. to 12:00 p.m., where 10h represents the probability that the vehicle trajectories are successfully linked between 10 and 11 o’clock and so on. Figure 6(b) shows the performance in Case 2 from July 9th to 11th, where 2-1 July 9th denotes the probability that the vehicle trajectories are successfully linked between VLPR 2 and VLPR 1 on July 9th and so on.

Figure 6 illustrates that releasing all of the original data within each VLPR is useful to preserving privacy. However, if $T_t$ is more than 0.7 times the estimated travel time, this is not sufficient since there is still a large proportion of vehicle traces (approximately 94% in Case 1 and more than 90% in Case 2) that can be successfully tracked. However, the general performance of the adversary model varies from one case to another, depending on the road topology between the upstream and downstream VLPRs. For example, if vehicles get off the road or end their trips in the middle of the VLPR pair, the estimated travel time would be inaccurate, eventually leading to a decline in the successful linking rate. In conclusion, in order to obtain the maximum linking rate, the value range of $T_t$ is suggested to be no less than 0.7.

We assume that the adversary is always able to track vehicle trajectories with the maximum probability. In Case 1 and Case 2, the privacy performances with different $m$s and $n$s are illustrated in Figure 7, where the $x$-coordinate is $m$ and the $y$-coordinate is the tracking probability. In Figure 7, type means that only the column type is the sensitive column; type & color means that the type and color columns are the sensitive columns; type & color & brand means that the type, color, and brand columns are the sensitive columns; and so on. Figure 7(a) represents the results of Case 1. Figure 7(b) denotes the results between VLPR 1 and VLPR 2 in Case 2, Figure 7(c) represents the results between VLPR 1 and VLPR 3 in Case 2, and Figure 7(d) denotes the results between VLPR 1 and VLPR 4 in Case 2.

As seen from Figure 7, if only one column is the sensitive column, the probability converges at a certain value and does not continue to decline. For the same curve, $p$ decreases at first as $m$ increases. When $m$ increases to a certain value, $p$ converges within an interval range. The convergence rate is different for different sensitive columns. For different $n$s, $p$ declines at different rates. The larger $n$ is, the faster $p$ drops. If $n$ and $m$ are big enough, the probability can even reach close to zero. For Case 1 and Case 2, $p$ changes similarly, which

| Attribute | Type       | Number of distinct values |
|-----------|------------|---------------------------|
| Brand     | Categorical| 18                        |
| Color     | Categorical| 11                        |
| Type      | Categorical| 4                         |

| Experiment | Time/date       | Number of vehicles (both VLPR 1 and VLPR 2) | Number of vehicles (both VLPR 1 and VLPR 3) | Number of vehicles (both VLPR 1 and VLPR 4) |
|------------|-----------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|
| Case 1     | 2018/07/30 10:00-10:59 | 235                                         |                                             |                                             |
|            | 2018/07/30 11:00-11:59 | 274                                         |                                             |                                             |
|            | 2018/07/30 12:00-12:59 | 265                                         |                                             |                                             |
| Case 2     | 2018/07/09       | 115                                         | 4597                                        | 178                                         |
|            | 2018/07/10       | 113                                         | 5354                                        | 204                                         |
|            | 2018/07/11       | 78                                          | 4587                                        | 201                                         |

**Table 4: Description of attributes.**

**Table 5: The numbers of vehicles in Case 1 and Case 2.**

Figure 5: Network geometry. (a) Shenluo highway. (b) Intersection of Songcheng road.
Figure 6: The change curves of $p$ with different rates. (a) Shenluo highway. (b) Intersection of Songcheng road.

Figure 7: Continued.
Figure 7: The change curve of $p$ with different $m_s$ and $n_s$. (a) The results of Case 1. (b) The results of Case 2 (VLPR 1 and VLPR 2). (c) The results of Case 2 (VLPR 1 and VLPR 3). (d) The results of Case 2 (VLPR 1 and VLPR 4).

Figure 8: Continued.
proves that the adversary model that is proposed in this paper is not only useful to the simple situation (Case 1) but also useful to the intersection situation (Case 2).

7.2. Comparison. To verify the effectiveness of the \((m, n)\)-bucket model that is proposed in this paper, the following experiment is performed using the 2018/07/30 12:00–12:59 data in Case 1. Since the calculated probability is the maximum probability (denoted as \(p(t, SA_i(v))\) in Section 4.2.3) of each individual being linked while the actual linking probability (denoted as \(pr\)) is for all individuals, in order to compare \(p(t, SA_i(v))\) and \(pr\), the average value of \(p(t, SA_i(v))\) for all individual \(t\)'s is used as the total calculated probability (denoted as \(pc\)). The results are shown in Figure 8, where \(pc\)-type represents the calculated probability with type as the sensitive column, \(pr\)-type denotes the actual linking probability with type as the sensitive column, and so on. In general, \(pc\) is greater than \(pr\), but there are special circumstances. For example, in Figure 8(b), when \(m\) is equal to 2, the \(pr\)-type-brand is greater than the \(pc\)-type-brand. This is because a small probability does not necessarily mean that the link fails. The results show that the derived mathematical model can well expound the privacy protection principle of the \((m, n)\)-bucket model.

8. Conclusion

In this paper, an adversary model based on the passing time, color, type, and brand of VLPR data is developed to evaluate the performance of the proposed \((m, n)\)-bucket model based on time series. When tracking vehicles between two VLPRs, releasing original data within VLPRs helps protect privacy but not at a satisfactory level; that is, this method of publishing data can lead to a high risk of privacy disclosure. Therefore, a data publishing model that can protect privacy is needed.

The proposed \((m, n)\)-bucket model is useful for privacy preservation. It overcomes the limitations of generalization and bucketization that can only handle single-column SA and data without time correlation requirements. The proposed model is a promising technique for handling high-dimensional data because \(ns\) can be any value in the model. People can set different \(ms\) and \(ns\) as needed. Meanwhile, the mathematical principle of the model for preventing privacy disclosures is expounded. Our experiments demonstrate that the proposed model has great significance in privacy preservation.

It is worth mentioning that just because only three QI attributes of vehicles are selected for the modeling discussions, it does not mean that the proposed model is only applicable to data with three QI attributes. The more vehicle QI attributes are considered, the greater the tracking probability will be. Fortunately, the model that is presented in this paper applies to any number of QI attributes. Three QI attributes for vehicles are selected just to verify the validity of the \((m, n)\)-bucket model. Which column is the SA is not the focus of this paper, and this study will be one of the next research directions. Although the method presented in this paper can usually achieve good results, it still has some limitations. We cannot deal with vehicles that are moving too fast or too slow. Because the estimated travel time is estimated using the log-normal distribution function, if the vehicle moves too fast or too slow, it is hard to estimate.

The model can be used to evaluate other scenarios for protecting confidential information, such as personalized privacy protection and continuous data publication. It is recommended that these issues be studied according to the proposed model.
Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant no. 2018YFB1601000) and the Fundamental Research Funds for the Central Universities under Contract no. 19lgpy290.

References

[1] T. Litman, Distance-Based Vehicle Insurance Feasibility, Costs and Benefits, Victoria Transport Policy Institute, Victoria, Canada, 2011.

[2] T. Litman, Distance-Based Vehicle Insurance Feasibility, Costs and Benefits, Victoria Transport Policy Institute, Victoria, Canada, 2018.

[3] S. Gambis, M. Killijian, and M. Del Prado Cortez, “Show me how you move and I will tell you who you are,” Transactions on Data Privacy, vol. 4, pp. 34–41, 2011.

[4] Y. Matsuo, N. Okazaki, K. Izumi et al., “Inferring long-term user properties based on users’ location history,” in Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 2159–2165, Hyderabad, India, January 2007.

[5] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, “Limits of predictability in human mobility,” Science, vol. 327, no. 5968, pp. 1018–1021, 2010.

[6] Y. d. Montjoye, C. A. Hidalgo, M. Verleysen et al., “Unique in the Crowd: the privacy bounds of human mobility,” Scientific Reports, vol. 3, no. 1, pp. 1–5, 2013.

[7] M. Olurin, C. Adams, and L. Logippo, “Platform for privacy preferences (P3P): current status and future directions,” in Proceedings of the 2012 Tenth Annual International Conference on Privacy, Security and Trust, pp. 217–220, Paris, France, July 2012.

[8] R. Shokri, G. Theodorakopoulos, P. Papadimitratos et al., “Hiding in the mobile crowd: location privacy through collaboration,” IEEE Transactions on Dependable and Secure Computing, vol. 11, no. 3, pp. 266–279, 2014.

[9] A. Khoshgozaran, H. Shirani-Mehr, and C. Shahabi, “Blind evaluation of location based queries using space transformation to preserve location privacy,” GeoInformatica, vol. 17, no. 4, pp. 599–634, 2013.

[10] M. Terrovitis and N. Mamoulis, “Privacy preservation in the publication of trajectories,” in Proceedings of the Ninth International Conference on Mobile Data Management, pp. 65–72, Beijing, China, April 2008.

[11] M. Nergiz, M. Atzori, and Y. Saygin, “Towards trajectory anonymization,” in Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS-SPRINGL 08, pp. 52–61, Irvine, CA, USA, November 2008.

[12] B. Zan, P. Hao, M. Gruteser et al., “VTI zone-aware path cloaking algorithm,” in Proceedings of the 2011 14th International IEEE Conference on Intelligent Transportation Systems, pp. 1525–1530, Washington, DC, USA, October 2011.

[13] Z. Sun, B. Zan, X. Ban, and M. Gruteser, “Privacy protection method for fine-grained urban traffic modeling using mobile sensors,” Transportation Research Part B: Methodological, vol. 56, pp. 50–69, 2013.

[14] O. Abul, F. Bonchi, and M. Nanni, “Never walk alone: uncertainty for anonymity in moving objects databases,” in Proceedings of the 24th IEEE International Conference on Data Engineering, pp. 376–385, Cancun, Mexico, April 2008.

[15] J. Gao, L. Sun, and M. Cai, “Quantifying privacy vulnerability of individual mobility traces: a case study of license plate recognition data,” Transportation Research Part C: Emerging Technologies, vol. 104, pp. 78–94, 2019.

[16] H. Li, Z. Ding, Q. Shi et al., “Study on global guidance strategy for urban traffic flow based on average speed information,” Journal of Hefei University of Technology, vol. 41, no. 5, pp. 595–600, 2018.

[17] Z. Sun, B. Zan, J. Ban et al., “Evaluation of privacy preserving algorithms using traffic knowledge based adversary models,” in Proceedings of the 2011 14th International IEEE Conference on Intelligent Transportation Systems, pp. 1075–1082, Washington, DC, USA, October 2011.

[18] F. Castellares, L. C. Montenegro, and G. M. Cordeiro, “The beta log-normal distribution,” Journal of Statistical Computation and Simulation, vol. 83, no. 2, pp. 203–228, 2013.

[19] X. Li, P. Sui, Y. Bai, and L.-E. Wang, “M-generalization for multipurpose transactional data publication,” Frontiers of Computer Science, vol. 12, no. 6, pp. 1241–1254, 2018.

[20] S. Yaseen, S. M. A. Abbas, A. Anjum et al., “Improved generalization for secure data publishing,” IEEE Access, vol. 6, pp. 27156–27165, 2018.

[21] L. Oneto, S. Ridella, and D. Anguita, “Differential privacy and generalization: sharper bounds with applications,” Pattern Recognition Letters, vol. 89, pp. 31–38, 2017.

[22] S. Zhang, X. Li, Z. Tan, T. Peng, and G. Wang, “A caching and spatialK-anonymity driven privacy enhancement scheme in continuous location-based services,” Future Generation Computer Systems, vol. 94, pp. 40–50, 2019.

[23] L. Sweeney, “k-anonymity: a model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[24] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in Proceedings of the 22nd International Conference on Data Engineering, Atlanta, GA, USA, April 2006.

[25] K. Mancuhan and C. Clifton, “Support vector classification with ε-diversity,” Computers & Security, vol. 77, pp. 653–665, 2018.

[26] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “L-diversity: privacy beyond k-anonymity,” in Proceedings of the 22nd International Conference on Data Engineering, Atlanta, GA, USA, April 2006.

[27] X. Xiao and Y. Tao, “Anatomy: simple and effective privacy preservation,” in Proceedings of the 32nd International Conference on Very Large Data Base Endowment, Seoul, Korea, September 2006.

[28] T. Li, N. Li, J. Zhang, and I. Molloy, “Slicing: a new approach for privacy preserving data publishing,” IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 561–574, 2012.
[29] A. S. M. T. Hasan, Q. Jiang, J. Luo et al., "An effective value swapping method for privacy preserving data publishing," Security and Communication Networks, vol. 9, no. 16, pp. 3219–3228, 2016.

[30] B. Li, Y. Liu, X. Han, and J. Zhang, "Cross-bucket generalization for information and privacy preservation," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 3, pp. 449–459, 2018.

[31] H. Zakerzadeh, C. C. Aggarwal, and K. Barker, "Managing dimensionality in data privacy anonymization," Knowledge and Information Systems, vol. 49, no. 1, pp. 341–373, 2016.

[32] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness," IEEE Transactions on Dependable and Secure Computing, vol. 16, no. 4, pp. 580–593, 2019.