International Journal of Geo-Information

MDPI

Conference Report

An Effective Privacy Architecture to Preserve User Trajectories in Reward-Based LBS Applications †

A S M Touhidul Hasan [1,2], Qiang Qu [1,*], Chengming Li [1], Lifei Chen [3] and Qingshan Jiang [1,*]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; touhidul.hasan@siat.ac.cn (A.S.M.T.H.); cm.li@siat.ac.cn (C.L.)
[2] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[3] College of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350117, China; clfei@fjnu.edu.cn
* Correspondence: qiang.qu@siat.ac.cn (Q.Q.); qs.jiang@siat.ac.cn (Q.J.); Tel.: +86-755-8639-2299 (Q.Q.); +86-755-8639-2340 (Q.J.)
† Part of this manuscript was presented and published at the 6th International Conference on Communication and Network Security, Singapore, 26–29 November 2016.

Received: 11 December 2017; Accepted: 5 February 2018; Published: 7 February 2018

Abstract: How can training performance data (e.g., running or walking routes) be collected, measured, and published in a mobile program while preserving user privacy? This question is becoming important in the context of the growing use of reward-based location-based service (LBS) applications, which aim to promote employee training activities and to share such data with insurance companies in order to reduce the healthcare insurance costs of an organization. One of the main concerns of such applications is the privacy of user trajectories, because the applications normally collect user locations over time with identities. The leak of the identified trajectories often results in personal privacy breaches. For instance, a trajectory would expose user interest in places and behaviors in time by inference and linking attacks. This information can be used for spam advertisements or individual-based assaults. To the best of our knowledge, no existing studies can be directly applied to solve the problem while keeping data utility. In this paper, we identify the personal privacy problem in a reward-based LBS application and propose privacy architecture with a bounded perturbation technique to protect user's trajectory from the privacy breaches. Bounded perturbation uses global location set (GLS) to anonymize the trajectory data. In addition, the bounded perturbation will not generate any visiting points that are not possible to visit in real time. The experimental results on real-world datasets demonstrate that the proposed bounded perturbation can effectively anonymize location information while preserving data utility compared to the existing methods.

Keywords: privacy architecture; identified trajectory; anonymization; data utility; location-based service

1. Introduction

We are witnessing a proliferation of geo-positioning capabilities. Smartphones, navigation devices, some tablets, and other mobile devices are equipped with Global Positioning System (GPS) receivers. Other available positioning technologies exploit the communication infrastructures used by mobile devices, such as Wi-Fi, 3G, and 2G. As a result of this development, user locations are used in a wide range of location-based service (LBS) applications such as health and wellness programs where walking or running trajectories are measured for training purposes. Recently, an increasing number of companies are rewarding employees based on their training data. Such LBS applications are called reward-based LBS applications. Notably, research shows that a healthy and fit employee is more

productive and generates lower expenditure for healthcare costs [1]. Moreover, a company may save $300,000 on healthcare costs by providing data to the insurance company (www.businessinsider.com/company-saved-money-with-fitbits-2014-7). Therefore, reward-based LBS applications are of importance and becoming popular in real world situations due to the "healthy" functionalities they offer, including rewarding and data-sharing. Users often use smart devices together with fitness LBS applications. These smart devices continuously collect a variety of measurements concerning physiological features such as acceleration, respiration, and electrocardiography (ECG) records [2,3], including smartphones and wearable devices [4]. Moreover, they often have modules of GPS, Wi-Fi, accelerometer, and different functional sensors that can keep track of user activities in real-time. With this data, user behaviors including smoking, personal stress, and moving patterns can be learned and reported by the applications [5,6]. By accumulating data from different sources and applying sophisticated machine learning algorithms, LBS healthcare systems have been attracting attention [3,7]. Reward-based LBS applications constitute such systems for business use.

The data collected from reward-based LBS applications are of interest to both employers and insurance companies with strong economic motivations. Employers store and analyse the data to provide rewards and motivate employees for proper training; insurance companies purchase the data from employers for product design. However, the data may pose a serious issue of personal privacy breach because identified trajectories and other confidential information are easy to restore [8,9]. The USA Federal Trade Commission has shown that aspects of the data privacy of popular fitness apps are easily breached including user names, emails, search histories, dietary habits, and activity routes (i.e., trajectories) (www.smh.com.au/digital-life/digital-life-news/data-collection-wearable-fitness-device-information-tracking-your-life-20150416-1mmzbq.html).

The data logs generated by a reward-based LBS application have trajectories with exact locations and the corresponding user identities, which can be used for spam advertisements, individual-based assaults, and linking attacks [8,10,11]. The breached user trajectories could be used to identify a user and her points of interest (PoIs) in the other publicly available datasets. For instance, bike sharing datasets are published publicly by removing user identity [12]. Reward-based LBS application data might be used to identify the particular user by linking attacks in the bike sharing datasets. The breach of identified trajectories thus brings serious security and privacy issues for implementing reward-based LBS applications. However, traditional data privacy mechanisms [13–15] cannot be used for this study, as they focus on preventing de-anonymization of identity data [16,17]. The privacy of the reward-based LBS applications not only indicates user identity, but also the corresponding trajectories. Therefore, it requires a client-server privacy setting to anonymize the user trajectory. Studies show that the leak of trajectories may reveal the corresponding user identities [18]. This study proposes a privacy setting that considers a user's privacy requirements, the adversary's background knowledge, and the anonymized data utility. Data utility [8,19] refers to how effectively the anonymized dataset acts compared to the original trajectory data. For instance, it can be used for billboard advertisements or traffic analysis. The privacy architecture thus outputs privacy-preserved data without changing the length and duration of trajectories for the corresponding user, which can be utilized to obtain the points of interest and crowded spots for a particular area. However, physical activity logs with trajectories pose challenges for the data anonymizer to protect personal privacy [8].

In this study, a client-server privacy architecture is introduced that protects identified trajectories with higher data utility. The privacy architecture follows privacy design principles [20] to anonymize trajectories at the user end with a fixed global location. We propose a bounded perturbation method for anonymizing identified trajectories. Note that perturbation methods modify spatial coordinates by adding random noises [15,21]. However, these methods may have problems in preserving data utility, as the performance routes would be significantly changed. The proposed method can protect the data privacy while keeping the data utility by global location set. The generated trajectory is able to avoid user privacy breach and achieve approximately equal performance in terms of the original dataset.

The remainder of this paper is structured as follows. The studies on the privacy and anonymization of spatial data are reviewed in Section 2. In Section 3, we present the details of the proposed privacy architecture. The experimental results are discussed in Section 4, and the paper is concluded in Section 5.

2. Background and Related Work

Reward-based LBS applications leveraging identified spatial data are becoming popular. Numerous privacy breaches have been proposed to illustrate how to breach the user privacy to obtain trajectory data [18,22–24]. In this section, we briefly review the inference and linking attacks, anonymization techniques, adversary, and background knowledge of trajectory data analysis.

2.1. Inference and Linking Attacks

An inference attack takes the input of LBS application data along with external information to gain underlying knowledge about users that breaches their individual privacy. This knowledge could be workplaces, home addresses, social networks, places of interest, mobility patterns, physical conditions, and political views. For instance, recent studies [18,25,26] show that the movement patterns can be predicted from visited places. In the study by Song et al. [26], the authors analysed 50,000 users' anonymized trajectories obtained from a mobile phone company, and the results show that 93% of the data could be used for mobility prediction.

Based on movement patterns, there are some mobility models such as semantic trajectories that can be used to link places with semantic information [18]. A user's mobility behavior can be easily obtained from the user's frequently visited routes [27]. An adversary might use the semantic information and frequent routes to derive a clear understanding of user mobility behaviors. For example, a user leaves home (PoI_1) bringing their child to school (PoI_2) before going to work (PoI_3) on weekdays, which is more in-depth knowledge than simply knowing the movement pattern $PoI_1 => PoI_2 => PoI_3$. A social relation shows that two users are in contact with a non-negligible amount of time and they share some social links, which may lead to the inference attack on the trajectory dataset.

In reward-based LBS applications, we know that the location data has been submitted to the organization with the user identity. From the inference attack, an adversary could gain knowledge about mobility frequency for a user's particular route. Therefore, inference attacks help an adversary to link the user with the publicly available data sources (i.e., bike sharing transaction records [12]) to identify the user's corresponding sensitive information, which is called a linking attack [11].

The related work shows that trajectories are of importance to user privacy. In this paper, we are focusing on hiding user visiting trajectories which are generated by reward-based LBS applications in order to avoid inference and linking attacks.

2.2. Anonymization Techniques

In relational databases, although identifiers are removed, a set of quasi-identifiers can re-identify a person in a published table [28]. To protect the re-identification of identity, *k*-anonymity [16] and *l*-diversity [17], among other methods, have been proposed to publish relational data with privacy preservation. Both *k*-anonymity and *l*-diversity use perturbation, suppression, and generalization to anonymize data. In these methods, the authors focus on de-anonymization attack protection. The privacy metrics and methods work well with relational data. However, the identified trajectories pose challenges to the methods, especially considering the preservation of data utility [2]. Similarly, in the studies of References [13–15,19,29–31], data privacy is ensured while user trajectories are published, so they cannot be used in a straightforward way in this work.

Spatial and temporal cloaking [30] is an extension of the *k*-anonymity for spatio-temporal data. The primary idea is to guarantee that at each time-stamp a user is located in a location that is shared by at least *k-1* other users. A conceivable approach to achieve the property of spatial and temporal clocking is to split the space into areas of different sizes until further splitting is necessary in order

to violate k-anonymity. Swapping [31] refers to exchanging the trajectories of a period between users. For example, by swapping trajectories of Alice and Bob for today, their behaviors become unrecognizable and predictable. Both methods are of a relational nature, and for the studies, the data privacy is ensured while user trajectories are published. Thus, they cannot be used to solve this problem in a straightforward way.

An anonymization algorithm A gets an individual LBS application daily trajectory L as input, introduces ambiguity by adding or removing some information from L, and generates the output L', an anonymized version of the original dataset L. The LBS application data for the proposed problem does not follow the relational nature. Therefore, the trajectory data needs to be anonymized with anonymization techniques such as perturbation, generalization, and suppression, and it does not satisfy any privacy model (namely k-anonymity). After anonymization, the trajectory data can ensure that the contributed user in the LBS application dataset will be indistinguishable. Considering individual trajectory data, we discuss pseudonymization, generalization, suppression, and perturbation anonymization techniques.

Pseudonymization [29] substitutes identifiers of mobility traces by creating an arbitrary pseudonym or combining unknown values. Pseudonymization is often insufficient for privacy protection because the trajectory data can identify the person. To limit the disclosure of a user trajectory, we can apply generalization [32], suppression [16], and perturbation [15,21,33] techniques.

Generalization is a process that modifies a value to a more generalized one [32]. If the value is numeric, this value may be changed to a range of values. For example, value 52 can be replaced by range 51–55. If the value is a categorical value, it may be transferred to another categorical value denoting a widespread concept of the original categorical value. For example, the country Japan can be changed to region Asia and the country Canada can be changed to North America. We can incorporate the generalization idea to anonymize the location data, but it will reduce the data utility [11]. Suppression is a process that changes a particular value in an attribute to a suppressed value, denoted by *, and it also reduces the data utility.

Conversely, perturbation modifies the location coordinates by adding some random noise. For example, random noise can be generated by a Gaussian or uniform distribution. In perturbation, if the surrounding area is not taken into consideration, the perturbed coordinates might have no physical sense (e.g., in a middle of a lake or on a cliff), and this reduces the data utility. In the following Table 1, we summarize the anonymization techniques which can be used to anonymize personal trajectory data.

Table 1. Summary of the anonymization techniques.

Anonymization	Methodology	Privacy Breach	Data Utility
Pseudonymization [29]	Substitutes the identity of the individual with arbitrary values.	High	High
Generalization [32]	Generalizes the trajectory data.	Medium	Low
Suppression [16]	Suppresses the trajectory data by a suppressed value, namely (*).	Low	Low
Perturbation [15,21,33]	Appends random noise to the trajectory data, and does not consider the surroundings.	Low	Medium

* denotes the suppressed value.

In this paper, we study the protection of identified trajectories from reward-based LBS applications. As this requires the preservation of data utility, generalization and suppression are not applicable. However, other methods based on pseudonymization have a higher privacy breach, and perturbation does not provide higher data utility. We thus propose a bounded perturbation for this work, which is a utility-preserving privacy technique for identified trajectories of reward-based LBS applications.

2.3. Adversary and Background Knowledge

The characterization of an adversary is essentially done by specifying the actions he intends to perform, the goal of his attack, and the way he can interact with the system. The adversary is an attacker who aims to retrieve identified trajectories to discover user points of interest and frequent routes which might breach the privacy of a user. In reward-based LBS applications, the data analyzer might have the potential to become an adversary, and would use his data access authorization to breach user points of interest and frequent routes.

Background knowledge can be described as an adversary's experience which has been obtained from life experience or discovered formally from prior rules coordinates in the data analysis. However, for LBS applications, the user's trajectory is be stored in the organization. We thus assume that the adversary's background knowledge is the collected trajectory data generated by the LBS users. This background knowledge helps the adversary to determine an individual's points of interest and frequent routes, which may lead to inference and linking attacks.

3. A Client-Server-Based Privacy Methodology

In this section, we introduce the client-server privacy architecture, followed by the proposed bounded perturbation technique to effectively anonymize identified trajectories with utility preservation.

3.1. Client-Server Privacy Architecture

The fundamental functionality of the proposed privacy architecture lies in anonymizing the user trajectory in a client-server privacy setting to ensure that the data contributors are safe in a dataset. Figure 1 illustrates the overview of the proposed client-server privacy architecture. In the privacy architecture, it has the following components: end-user (or employee), intermediary device (client), central server, business organization (or employer), and data processing organization (or insurance company).

Figure 1. Client-server privacy architecture.

An employee generates LBS application-specific data by health fitness devices that link with the intermediary device (e.g., computer, cell-phone) for the trajectory anonymization. An intermediary device could be a cell phone, a computer, or any suitable device that executes the procedure to complete the anonymization processes. It anonymizes the trajectory at the user end by applying the bounded perturbation technique. For anonymization, the intermediary device requests that the global location is set to the central server and anonymizes the visiting locations of the user.

The central server supports the anonymization process, which generates the global location set and keeps the records from users. An employer is the facilitator to introduce a reward-based LBS application. Moreover, an employer uses the application-specific data to give rewards to end-users for being active in daily life and to negotiate with insurance companies to reduce insurance cost.

An insurance company is a third-party business organization that would use end-user data for analysis.

3.2. Anonymization

Identified trajectories would breach user privacy, and we must anonymize such trajectories before sharing them with the central server. To anonymize the trajectories, we introduce a global location set to perform bounded perturbation.

Let u denote a user in the reward-based LBS system. A movement of user u updates a tuple $< id, (x, y, t) >$, where id represents the identity of the user u, and the tuple describes that the user u visited point (x, y) at time t. Here, x is the longitude, y is the latitude, and t is the detailed recording of time. The user movement history can be used to draw the movement patterns of the user, and it is defined as follows.

Definition 1 (Identified trajectory)**.** *An identified trajectory is a sequence of successive PoIs visited by an identified user along time t, represented as*

$$L = \{id, (x_1, y_1, t_1), (x_2, y_2, t_2), \ldots, (x_n, y_n, t_n)\},$$

where id is the identity of the user u, (x_i, y_i) *represents the longitude and latitude of* PoI_i*, and* t_i *is the corresponding time.*

The strategy of reward-based LBS applications is to preserve the trajectory data L along with other application-specific features. In the preserved trajectory L, any independent part of a user u with a start and end point is called a route. In an urban area, a route is called a frequent route if a route appears a number of times in the trajectory dataset. Because of the frequent nature, the route is considered as an unsafe route for the user u [27]. In addition, a unique route might identify a user and threaten his privacy.

For the reward-based LBS application, we consider that the trajectory data L might have frequent and unique user routes. Therefore, identified trajectories would breach user privacy, and we must anonymize user daily trajectories before sharing them with the business organization. To anonymize the trajectories, a global location set is introduced to perform the anonymization operation.

Definition 2 (Global location set (GLS))**.** *A GLS consists of all the points of interest (PoIs) in a region. Every PoI has a location and a description for its semantic meaning, such as walkway, highway, residential house, lake, mountain, and landmark. A GLS is represented as*

$$(loc_1, location_1), (loc_2, location_2) ... (loc_i, location_j),$$

where $loc_i = (latitude, longitude)$ *is a pair of coordinates of PoI i, and* $location_j$ *represents its semantic description.*

GLS is generated by using OpenStreetMap API [34]. Figure 2 is an example of a global location set, where nodes represent the *PoIs*. Generally, *PoIs* are connected with each other by the road networks. Therefore, for the analysis, we consider that the *PoIs* are connected with each other by a distance d.

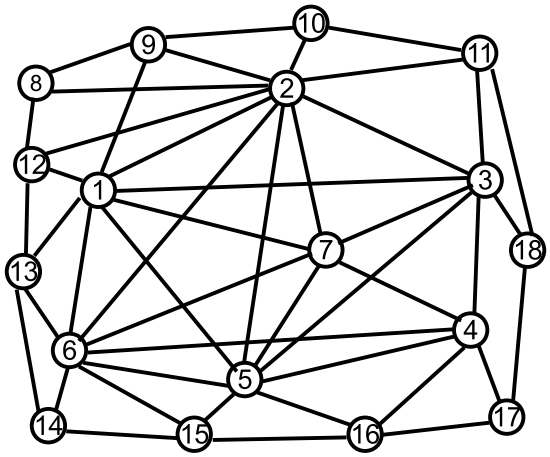

Figure 2. Illustration of global location set.

The user's visited trajectory is L, and we know that L would breach user privacy by revealing the *PoIs* and frequent routes. For trajectory anonymization, we may append the random noise $r = (rd, ra)$ with the visited *PoIs*, which will result in new $PoIs'$ that are impossible to visit in reality (e.g., in the middle of a lake or on a mountain cliff). Therefore, we propose *bounded perturbation* techniques with the global location set GLS to anonymize the location trajectory L.

Definition 3 (Bounded perturbation)**.** *Unlike the perturbation methods in References [15,21], bounded perturbation includes random noise $r = (rd, ra)$ with the actual data points, but it takes the global location set GLS into consideration to guarantee its availability. Therefore, the generated data points would not include invalid places. This reduces the utility of the generated trajectories.*

For instance, the new *PoIs* are generated by adding some random noise r; i.e., $(x'_i = x_i + r_x, y'_i = y_i + r_y)$, where r_x is the random noise for x_n and r_y is the random noise for y_n. To calculate random noise for longitude and latitude points, we use the Earth's radius; we assume the Earth's radius is R, random distance is rd in $[l \leq rd \leq k]$, where l is the minimum and k is the maximum distance in meters which depends on the area of the city, and random uniform direction ra over $[0, 2\pi]$, respectively. Then, the random noise r_x, r_y is

$$r_x = \frac{rd}{R(\cos(\pi \times \frac{y_n}{ra}))} \times \frac{ra}{\pi}, \tag{1}$$

$$r_y = \frac{rd}{R} \times \frac{ra}{\pi}. \tag{2}$$

We compare the generated point (x'_i, y'_i) with the global location set GLS by the map matching techniques [35]. If it is in the GLS, we keep it as a newly generated anonymized point. Otherwise, we recompute the point with new random noise until it is in the GLS.

Given an identified trajectory L and applying bounded perturbation technique to generate the anonymized trajectory L', represented by

$$L' = \{id, (x'_1, y'_1, t_1), (x'_2, y'_2, t_2), \ldots, (x'_n, y'_n, t_n)\},$$

where id represents the identity of the user u, (x'_i, y'_i) represents the generated longitude and latitude PoI_i, and t represents the corresponding time.

An anonymized trajectory L' is the newly generated trajectory bounded by GLS with the visited timing information, and it excludes original visited places from the identified trajectories, so that the anonymized trajectory L' will protect the user's *PoIs*, unique routes, and frequent routes.

The primary goal for preserving user privacy is anonymizing the trajectory, and we must also consider the data utility. Therefore, in the anonymized dataset, we measure the data utility by the relative distortion.

Definition 4 (Relative distortion). *The relative distortion measures the quality of the anonymized dataset compared with the original dataset. Let the original count be O and the anonymized count be A, then the relative distortion is*

$$Relative\ Distortion = \frac{|O - A|}{O}. \tag{3}$$

To calculate relative distortion, we consider the positive distortion (i.e., calculate the absolute difference of original and anonymization count).

For instance, if a *PoI* is visited 10 times in the original dataset and 6 times in the anonymized dataset then the relative distortion of the *PoI* is $|10 - 6|/10$, which is 0.4.

3.3. Anonymization Algorithms

In this section, we present two algorithms to perform anonymization of identified trajectories. We introduce the global location set to do the anonymization process effectively, and Algorithm 1 produces a global location set. Taking the daily identified trajectory L and the global location set as arguments, the Algorithm 2 outputs the anonymized trajectory L' for the corresponding L.

3.3.1. Generation of Global Location Set

The algorithm is introduced to complete the perturbation process for user-visited locations over a particular time period. As a perturbation process may generate invalid *PoIs*, we introduce a global location set GLS, and the generation of *PoIs* will be within the set in order to avoid problems. We thus call the perturbation process a *bounded perturbation*. Algorithm 1 produces the GLS at the server side, and an intermediary device requests it in order to complete the anonymization process. In addition, Algorithm 1 helps to generate those locations that are valid. An employer specifies the possible areas to analyze the employees' data, and then the algorithm determines the bounding box of the global location set. In line 1, we initialize a bounding box B and a global location set GLS. A global location set GLS is generated in lines 1 to 4, and finally returns the GLS.

Algorithm 1 Generation of a global location set.

Input: A bounding box B in the OpenStreetMap [34] with four parameters (x_1, y_1, x_2, y_2) where x_1 and x_2 are longitude coordinates and y_1 and y_2 are latitude coordinates;
Output: Generate the possible visited locations GLS ={all possible visited locations} in the bounding box B;

1: Initialize $B = \{(x_1, y_1), (x_2, y_2)\}, GLS = \{\}$, VL = {highway, landmark, etc.};
2: **for** Each point $\{loc_i, location_j\}$ in the bounding box B **do**
3: **if** $(B(loc_i, location_j) == VL)$ **then**
4: $GLS = \{GLS\} \cup \{loc_i, location_j\}$;
return GLS

3.3.2. Anonymized Location Trajectory

The *anonymized location trajectory* algorithm is executed at the user end in a suitable device called the intermediary device (e.g., a computer or smart-phone) to anonymize the user's visited locations. Please note that we assume that wearable devices only collect physiological features with locations

and do not conduct any further processing, as the devices often have limited memory and processing power to execute the anonymization algorithm.

The algorithm takes the global location set GLS and user-visited locations L as input, and it produces the anonymized locations L'. We initialize the temporary global location set GLS_T, anonymized location trajectory L', random distance rd in $l \leq rd \leq k$, and random direction ra over $[0, 2\pi]$ to calculate random noise r. To generate the anonymized trajectory L', the original visited locations in L are deducted from the global location set GLS to make sure the generated locations do not have any original visited locations. In line 3, the algorithm generates the new anonymized trajectory L' by adding random noise r. In line 4, the timing information is added with the anonymized trajectory, and the algorithm returns the anonymized trajectory L' in line 5.

Algorithm 2 Anonymized location trajectory.

Input: Identified trajectory L, global location set GLS.
Output: Anonymized location trajectory L' in the global location set GLS.

1: Initialize $GLS_T = \{\}, L' = \{\}$, random noise r;
2: Deduct the real visited locations from the global location set $GLS_T = GLS - L$;
3: Generate new locations L' in GLS_T for each visited point L with random noise r_x and r_y from Equations (1) and (2).
4: Adding the timing information for each point in L' as in the L;
return L'

3.4. Discussion on the Anonymization Technique

In the client-server privacy architecture, two algorithms are introduced to complete the anonymization process. Figure 2 is an example of a *global location set*, and we extracted it from OpenStreetMap data by using Algorithm 1. The visited path with timing information is called the identified trajectory and routes, and from the problem definition, it is known that an identified trajectory, unique, and frequent routes might breach the user's privacy. Therefore, it is necessary to anonymize identified trajectories to protect the user's *PoIs*, unique routes, and frequent routes before submitting them for further processing.

Suppose a user trajectory is $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$, and Algorithm 2 is used to anonymize the visited locations. For simplicity, we consider that the distance d is uniform between points of interest. In anonymization, to add noise to the trajectory, distance rd and direction ra are selected randomly to generate the new anonymized visited locations. In Algorithm 2 (line 2), the real visited path L is excluded from the global location set GLS, and line 4 generates the new anonymized visited locations in L' for the user. The algorithm could generate $6 \rightarrow 7 \rightarrow 5 \rightarrow 6$ as the anonymized trajectory, which could be submitted for publishing. Supposing another trajectory $5 \rightarrow 7 \rightarrow 3 \rightarrow 5$, the generated anonymized visited trajectory could be $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$. Both visited $1, 2, 3, 5, 7$ in reality and anonymized points of interest are $1, 2, 3, 5, 6, 7$.

By the existing studies, we know that perturbation can be used to solve the user privacy issues, but perturbation may generate some places that are practically impossible to visit, which reduces the data utility. We thus proposed a bounded perturbation with the help of global location set GLS, and this does not generate any locations that are not practically possible to visit. Therefore, our proposed bounded perturbation method has more data utility than the classical perturbation method.

4. Experimental Evaluation

In this section, we demonstrate experiments on real-world datasets. The experiments are divided into two parts: the first part was designed to present the personal privacy breach and test the effectiveness of the proposed bounded perturbation algorithm for trajectory anonymization in comparison with the perturbation methods. Our experimental results show that the bounded

perturbation method can successfully anonymize the trajectory points at the intermediary device. The results of this experiment are presented in Sections 4.2 and 4.3.

In the second part, we measure the effectiveness of bounded perturbation and perturbation techniques in preserving data utility as compared with the original trajectory data. The experimental results demonstrate that the bounded perturbation preserves more data utility than the perturbation method. The results of this experiment are presented in Section 4.4.

4.1. Data Set

In the experiment, we use the Geolife [36,37] project dataset to simulate the reward-based LBS application. This dataset was collected in the Geolife project from 182 users over three years, and it contains 17,621 identified trajectories.

To assemble the experimental environment, we used OpenStreetMap API [34], R and several R packages [38–40], QGIS [41], and Google Fusion Tables [42] for the analysis of trajectory data.

4.2. Privacy Breach

In this section, we present the significance of protecting the user privacy of identified trajectories in a real-life setting. Figure 3 demonstrates user trajectories over a period of time, and it gives the confidence to the adversary to learn about the particular user and can initiate an inference attack or conduct a linking attack to an available dataset. From the trajectories for a period, the adversary can find the user's frequent routes and the points of interest. The moving path has the user identity, timing, and latitude-longitude values. Therefore, the identified trajectories from a reward-based LBS application might breach user privacy.

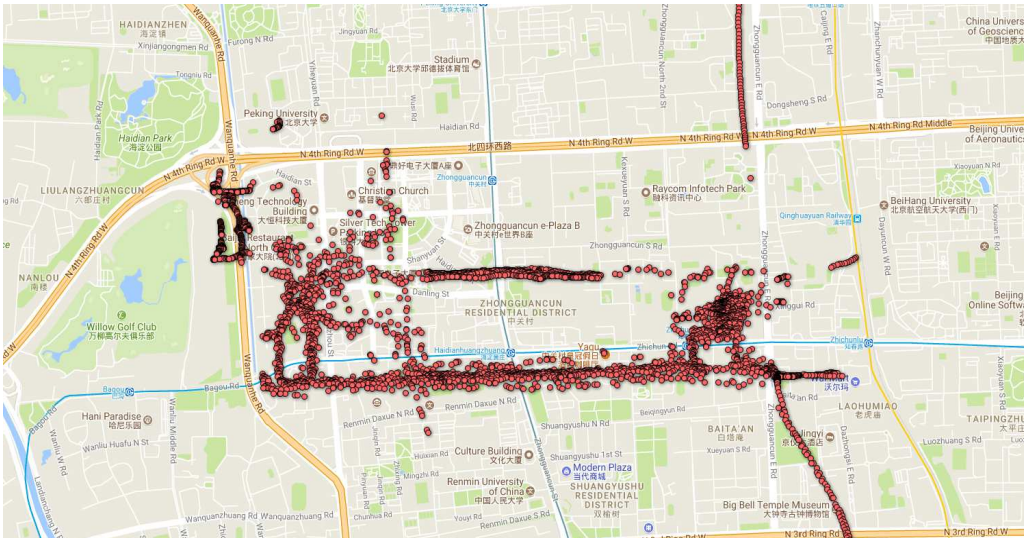

Figure 3. Identified user trajectories.

By analyzing particular user data from the dataset, it is possible to find a user's points of interest and frequent routes to exhibit the privacy breach. Figure 3 exposes the particular user's visited trajectory for a period of time. An adversary might use this particular user dataset to find the user's points of interest and frequent routes and apply this background knowledge to breach the user's privacy.

In terms of the Geolife [36,37] project dataset description, the location data were collected at five-second intervals; i.e., the location's latitudes and longitudes would be collected if the user's speed was at least 3.0 miles per hour. This means that if the user's speed was below 3.0 miles per

hour, the data collection was stopped (e.g., taking a tour in the garden, watching a movie in a theater, exercising in the gym, or visiting a friend's home). The recorded last location point signifies the user's point of interest. Suppose a user was riding a bike to the gym and parked her bike at the parking place and the last location was recorded. Then, the adversary might conclude that the gym is one of the user's points of interest (considering that she is not running).

As a consequence, we can use the time interval properties of two consecutive location points to determine a user's visited points of interest. For instance, the time between two consecutive location records may determine if the location is a given user *PoI*. By analyzing the user dataset, we divide the time interval (in sec) into: $time\ interval \geq 300$, $time\ interval \geq 600$, $300 \leq time\ interval \leq 1800$, and $600 \leq time\ interval \leq 3600$. By the intervals and location information, we can easily find the points of interest of a particular user. In the experiment, we observe that a particular person visited 42 places in $600 \leq time\ interval \leq 3600$. Therefore, we can conclude that the user was more interested in those places.

4.3. Location Trajectory Anonymization

From the privacy breach section, it is observed that the user's privacy was breached by revealing points of interest and frequent routes, which may lead to an inference attack and a linking attack. In this case, it is necessary to anonymize the identified trajectory before presenting it to the organization's central server or a third-party service provider. In the privacy setting, we anonymized the user's daily trajectory and submitted it to the central server. Figure 4 shows the anonymized locations of a user for a period of time, and the figure demonstrates that it has no frequent routes which the adversary may use to breach user privacy (i.e., inference and linking attacks).

In a reward-based LBS application, for instance, a health/wellness program's primary objective is to increase the productivity of the employees by keeping them healthy. In addition, the organization may utilize the physiological features to reduce the increasing medical insurance cost. The numerous demands for health/wellness programs and the data they produce might breach user privacy. Therefore, we proposed bounded perturbation techniques to generate logical locations to ensure users' privacy.

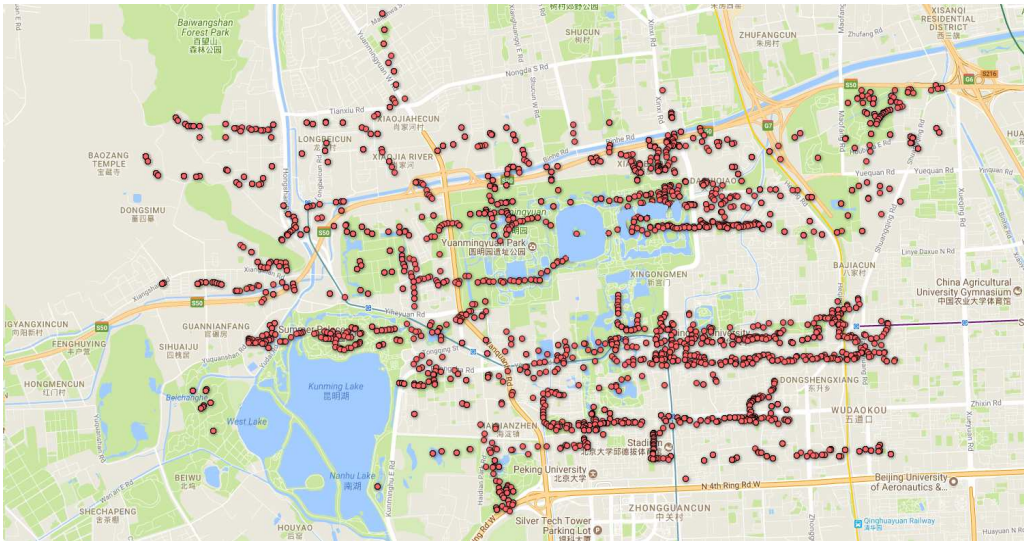

Figure 4. Anonymized locations of a user.

4.4. Data Utility

Data utility is of great importance to reward-based LBS applications. In this section, we evaluate the data utility of the anonymized trajectories of the proposed bounded perturbation method. Anonymization at the user end might reduce the data utility regarding traffic or other data mining analysis because generated locations would be different locations. This can be true if the dataset contains only one user, but reward-based LBS applications are often supposed to have more than one user. By analyzing trajectory data, one can identify the points of interest and visiting users, which may be used by the advertising company to place a billboard advertisement. To determine mass population movements and points of interest, we need to consider the number of visitors, not the individuals who visited.

To see the effectiveness of the generated trajectory, we conducted two data utility measurement experiments on the anonymized and original datasets. In the first experiment, we identified the point of interest, and for each point of interest, we calculated the relative distortion. In the second experiment, the visiting users were counted for a particular area, and relative distortion was calculated. These experimental results may determine the quality of the anonymized dataset, and could be used to find the crowded spot in an area to place a billboard advertisement.

In the visited locations, not all points are considered as *PoIs*. To find the time-specific visited *PoIs*, we grouped the places into time intervals between two consecutive visited points, and we have divided the time intervals (in sec) into: *time interval* $\geq$ 300, *time interval* $\geq$ 600, 300 $\leq$ *time interval* $\leq$ 1800, and 600 $\leq$ *time interval* $\leq$ 3600.

In Figure 5, on the Y-axis we plotted the *PoIs* in every 10,000 data points for the original data, bounded perturbation-generated data, and perturbation-generated data. It shows that bounded perturbation method had more *PoIs* than the perturbation methods [15] for the particular area.

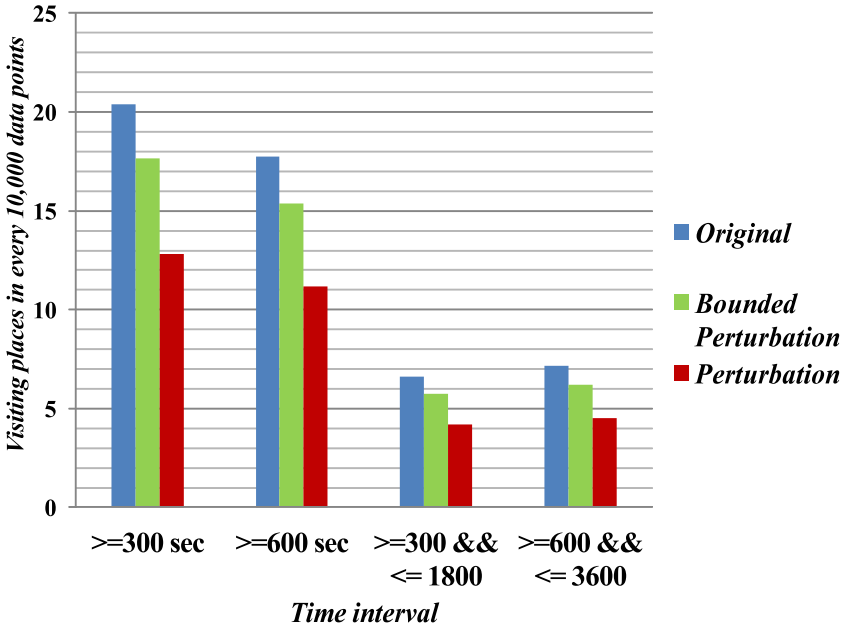

Figure 5. Visited points of interest (*PoIs*) in different time intervals.

The use case of Definition 4 is as follows. The DBSCAN [43] algorithm was applied to the original dataset to obtain the most visited *PoIs* in a region and rank them based on the number of times they were visited [44–46]. For each *PoI* from the original dataset, the count was measured in the anonymized datasets. After getting the anonymized visited count for individual *PoIs*, the relative distortion was

calculated by Equation (3). Furthermore, we calculated all individual *PoIs*' relative distortion and the average distortion. Let the relative distortion of a *PoI* be D; then, the average distortion of all *PoIs* can be determined by

$$Average\ Relative\ Distortion = \frac{1}{n} \times \sum_{i=1}^{n} D, \tag{4}$$

where n is the number of *PoIs*.

The relative distortion on individual *PoIs* describes the quality of the anonymized dataset in comparison with the original dataset. Table 2 demonstrates the average relative distortion for all *PoIs*. Experimental results demonstrate that the bounded perturbation has lower relative distortion than the perturbation method.

Table 2. Average relative distortion for *PoIs*.

Anonymization Technique	Average Relative Distortion
Bounded Perturbation	0.206357345
Perturbation	0.48617868

We conducted the visiting users' experiment to find the most crowded spot in an area. To determine which visiting users were in a selected area, we used the following query based on the criteria in the study [47]:

$$SELECT\ COUNT(*)\ FROM\ Dataset\ WHERE\ Area \in PoI\ AND\ Date\&Time \in Session, \tag{5}$$

where *Dataset* is the original or the anonymized dataset. *Area* is based on individual *PoIs*, which was obtained by the DBSCAN algorithm. From the individual *PoIs*, we obtained the longitude and latitude values and appended a 100-m radius to select the particular area. For the *Data&Time*, 30 individual days were selected with morning (07:00–10:00) and evening (17:00–20:00) sessions.

In the experiment, the individual query was made by using Equation (5), and a search was conducted on the original and anonymized datasets. Suppose that we want to count the number of users in Area 1, Day 1, and the Morning session. Then, the query returns the count for original and anonymized datasets. From the results of the original and anonymized datasets for Area 1, Day 1, and the Morning session, the relative distortion of the visiting users for the particular session was calculated by Equation (3). In the experiment, we selected 30 individual days, morning and evening sessions, and all possible areas to calculate the distortion. After that, the average distortion was calculated by Equation (4). Table 3 presents the average relative distortion in the anonymized dataset. Experimental results demonstrate that the bounded perturbation achieves better data quality.

Table 3. Average relative distortion.

Anonymization Technique	Average Relative Distortion in the Morning	Average Relative Distortion in the Evening
Bounded Perturbation	0.240973419	0.280139043
Perturbation	0.561570944	0.52706014

All of the users of the reward-based LBS applications followed the same anonymization method and submitted their data to the central server. In the bounded perturbation process, all the users from the same employer used the same global location set *GLS* to finish the anonymization process. Therefore, the anonymized locations that were generated by the proposed bounded perturbation model had more data utility than the perturbation method.

From the experimental results, an advertising company would know the most crowded spot to post a billboard advertisement. Thus, the bounded perturbation-generated trajectory data could be used by the advertising company in the same way as the original trajectory data. In summary, it is

possible to say that the generated trajectory data by the bounded perturbation has more data utility compared with the perturbation method.

5. Conclusions and Future Work

In this study, we showed that location data might breach a user's points of interest and frequent routes, which would lead to inference and linking attacks. This research demonstrated the significance of the anonymization of identified trajectories. In this paper, the proposed client-server privacy architecture was able to preserve user privacy while keeping the data utility of the identified trajectories. The global location set-based bounded perturbation techniques could anonymize the identified trajectory to protect the user's points of interest and frequent routes. Therefore, the anonymized trajectory is defended from inference and linking attacks. Experimental findings showed that the proposed privacy architecture was effective in terms of privacy concerns and data utility compared to the conventional perturbation methods.

In the future, we aim to improve the bounded perturbation method by completing the anonymization process locally. Moreover, the collected data from the reward-based LBS applications may have other properties in addition to identified trajectories that may have issues of privacy. We would thus extend the current work by supporting various data attributes in the setting.

Acknowledgments: This research work was supported by Shenzhen Technology Development Grant No. CXZZ20150813155917544, Shenzhen Fundamental Research Foundation Grant No. JCYJ20150630114942277, Guangdong Province Research Grant No. 2015A030310364; and sponsored by the CAS-TWAS President's Fellowship for International Ph.D. students.

Author Contributions: A S M Touhidul Hasan, Chenming Li conceived and designed the experiments and performed the experiments; A S M Touhidul Hasan and Qingshan Jiang analyzed the data and contributed analysis tools; A S M Touhidul Hasan, Qiang Qu and Lifei Chen wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mattke, S.; Liu, H.; Caloyeras, J.P.; Huang, C.Y.; Van Busum, K.R.; Khodyakov, D.; Shier, V. *Workplace Wellness Programs Study*; Rand Corporation: Santa Monica, CA, USA, 2013.
2. Raij, A.; Ghosh, A.; Kumar, S.; Srivastava, M. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; pp. 11–20.
3. Choi, H.; Chakraborty, S.; Charbiwala, Z.M.; Srivastava, M.B. Sensorsafe: a framework for privacy-preserving management of personal sensory information. In *Secure Data Management*; Springer: Berlin, Germany, 2011; pp. 85–100.
4. Issa, H.; Shafaee, A.; Agne, S.; Baumann, S.; Dengel, A. User-sentiment based evaluation for market fitness trackers-evaluation of fitbit one, jawbone up and nike+ fuelband based on amazon.com customer reviews. In Proceedings of the 1st International Conference on Information and Communication Technologies for Ageing Well and e-Health, ICT4AgeingWell 2015, Lisbon, Portugal, 20–22 May 2015; SCITEPRESS: Setúbal, Portugal, 2015; pp. 171–179.
5. Plarre, K.; Raij, A.; Hossain, S.M.; Ali, A.A.; Nakajima, M.; al'Absi, E.; Ertin, T.; Kamarck, T.; Kumar, S.; Scott, M.; et al. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In Proceedings of the 2011 10th International Conference on Information Processing in Sensor Networks (IPSN), Chicago, IL, USA, 12–14 April 2011; pp. 97–108.
6. Reddy, S.; Burke, J.; Estrin, D.; Hansen, M.; Srivastava, M. Determining transportation mode on mobile phones. In Proceedings of the 12th IEEE International Symposium on Wearable Computers, ISWC 2008, Pittsburgh, PA, USA, 28 September–1 October 2008; pp. 25–28.
7. Kotz, D.; Avancha, S.; Baxi, A. A privacy framework for mobile health and home-care systems. In Proceedings of the First ACM Workshop on Security and Privacy in Medical and Home-Care Systems, Chicago, IL, USA, 13 November 2009; pp. 1–12.

8. Krumm, J. Inference attacks on location tracks. In *Pervasive Computing*; Springer: Berlin, Germany, 2007; pp. 127–143.
9. Aïvodji, U.M.; Gambs, S.; Huguet, M.-J.; Killijian, M.-O. Meeting points in ridesharing: A privacy-preserving approach. *Transp. Res. Part C Emerg. Technol.* **2016**, *72*, 239–253.
10. Hasan, A.S.M.T.; Jiang, Q.; Li, C.; Chen, L. An effective model for anonymizing personal location trajectory. In Proceedings of the 6th International Conference on Communication and Network Security, Singapore, 26–29 Novemerbr 2016; pp. 35–39.
11. Hasan, A.S.M.T.; Jiang, Q.; Li, C. An effective grouping method for privacy-preserving bike sharing data publishing. *Future Internet* **2017**, *9*, 65, doi:10.3390/fi9040065.
12. Citi Bike Daily Ridership and Membership Data. Available online: https://www.citibikenyc.com/system-data (accessed on 3 April 2017).
13. Fan, L.; Xiong, L.; Sunderam, V. Fast: Differentially private real-time aggregate monitor with filtering and adaptive sampling. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 22–27 June 2013; pp. 1065–1068.
14. Cao, Y.; Yoshikawa, M. Differentially private real-time data release over infinite trajectory streams. In Proceedings of the 2015 16th IEEE International Conference on Mobile Data Management (MDM), Pittsburgh, PA, USA, 15–18 June 2015; Volume 2, pp. 68–73.
15. Armstrong, M.P.; Rushton, G.; Zimmerman, D.L. Geographically masking health data to preserve confidentiality. *Stat. Med.* **1999**, *18*, 497–525.
16. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzz. Knowl. Based Syst.* **2002**, *10*, 557–570.
17. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3.
18. Gambs, S.; Killijian, M.-O.; del Prado Cortez, M.N. Show me how you move and i will tell you who you are. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, San Jose, CA, USA, 2 November 2010; pp. 34–41.
19. Zhang, Z.; Sun, Y.; Xie, X.; Pan, H. An efficient method on trajectory privacy preservation. In *Big Data Computing and Communications*; Springer: Berlin, Germany, 2015; pp. 231–240.
20. Langheinrich, M. Privacy by design principles of privacy aware ubiquitous Systems. In *International Conference on Ubiquitous Computing*; Springer: Berlin, Germany, 2001; pp. 273–291.
21. Kwan, M.-P.; Casas, I.; Schmitz, B. Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartogr. Int. J. Geogr. Inform. Geovisual.* **2004**, *39*, 15–28.
22. Hansell, S. Aol removes search data on vast group of web users. *N. Y. Times* **2006**, *8*, C4.
23. Narayanan, A.; Shmatikov, V. De-anonymizing social networks. In Proceedings of the 2009 30th IEEE Symposium on Security and Privacy, Oakland, CA, USA, 17–20 May 2009; pp. 173–187.
24. Krumm, J. A survey of computational location privacy. *Pers. Ubiquitous Comput.* **2009**, *13*, 391–399,
25. Gonzalez, M.C.; Hidalgo, C.A.; Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782.
26. Song, C.; Qu, Z.; Blumm, N.; Barabási, A.-L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021.
27. Gkoulalas-Divanis, A.; Verykios, V. A free terrain model for trajectory k–Anonymity. In *Database and Expert Systems Applications*; Springer: Berlin, Germany, 2008; pp. 49–56.
28. Hasan, A.S.M.T.; Jiang, Q.; Luo, J.; Li, C.; Chen, L. An effective value swapping method for privacy preserving data publishing. *Secur. Commun. Netw.* **2016**, *9*, 3219–3228.
29. Pfitzmann, A.; Hansen, M. Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management-a Consolidated Proposal for Terminology; version v0.25, December 2005, Citeseer. Available online: https://www.freehaven.net/anonbib/cache/terminology.pdf (accessed on 6 April 2017).
30. Gruteser, M.; Grunwald, D. Anonymous usage of location-based services through spatial and temporal cloaking. In Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, San Francisco, CA, USA, 5–8 May 2003; pp. 31–42.
31. Domingo-Ferrer, J.; Sramka, M.; Trujillo-Rasúa, R. Privacy-preserving publication of trajectories using microaggregation. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, San Jose, CA, USA, 2 November 2010; pp. 26–33.

32. Samarati, P.; Sweeney, L. Generalizing data to provide anonymity when disclosing information. In Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, Seattle, DC, USA, 1–3 June 1998; Volume 98, p. 188.
33. Liu, S.; Qu, Q.; Chen, L.; Ni, L.M. SMC: A practical schema for privacy-preserved data sharing over distributed data streams. *IEEE Trans. Big Data* **2015**, *1*, 68–81.
34. OpenStreetMap Contributors. Available online: https://www.openstreetmap.org (accessed on 3 March 2017).
35. Greenfeld, J.S. Matching gps observations to locations on a digital map. In Proceedings of the Transportation Research Board 81st Annual Meeting, Washington, DC, USA, 13–17 January 2002.
36. Zheng, Y.; Zhang, L.; Xie, X.; Ma, W.-Y. Mining interesting locations and travel sequences from gps trajectories. In Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, 20–24 May 2009; pp. 791–800.
37. Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; Ma, W.-Y. Understanding mobility based on gps data. In Proceedings of the 10th International Conference On Ubiquitous Computing, Seoul, Korea, 21–24 September 2008; pp. 312–321.
38. R Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
39. Eugster, M.J.A.; Schlesinger, T. osmar: OpenStreetMap and R Journal, 2010. Accepted for Publication on 14 August 2012. Available online: http://osmar.r-forge.r-project.org/RJpreprint.pdf (accessed on 4 January 2017).
40. Wickham, H.; Francois, R. dplyr: A Grammar of Data Manipulation, 2015, r Package Version 0.4.3. Available online: https://CRAN.R-project.org/package=dplyr (accessed on 4 January 2017).
41. QGIS Development Team, QGIS Geographic Information System, Open Source Geospatial Foundation, 2009. Available online: http://qgis.osgeo.org (accessed on 3 March 2017).
42. Gonzalez, H.; Halevy, A.Y.; Jensen, C.S.; Langen, A.; Madhavan, J.; Shapley, R.; Shen, W.; Goldberg-Kidon, J. Google fusion tables: web-centered data management and collaboration. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, Indianapolis, IN, USA, 6–10 June 2010; pp. 1061–1066.
43. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 2–4 August 1996; Volume 96, pp. 226–231.
44. Peixoto, D.A.; Xie, L. Mining Trajectory Data. 2013. Available online: https://www.researchgate.net/profile/Douglas_Peixoto/publication/275381558_Mining_Trajectory_Data/links/553b4e320cf245bdd76468c5.pdf (accessed on 28 December 2017).
45. Zheng, Y. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 29.
46. Li, Q.; Zheng, Y.; Xie, X.; Chen, Y.; Liu, W.; Ma, W.-Y. Mining user similarity based on location history. In Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine, CA, USA, 5–7 November 2008; p. 34.
47. Zhang, Q.; Koudas, N.; Srivastava, D.; Yu, T. Aggregate query answering on anonymized tables. In Proceedings of the IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007; pp. 116–125.