future internet

MDPI

Article

An Effective Grouping Method for Privacy-Preserving Bike Sharing Data Publishing

A S M Touhidul Hasan [1,2], Qingshan Jiang [1,*] and Chengming Li [1]

1 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; touhidul.hasan@siat.ac.cn (T.H.); cm.li@siat.ac.cn (C.L.)
2 Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: qs.jiang@siat.ac.cn; Tel.: +86-0755-8639-2340

Received: 20 September 2017; Accepted: 16 October 2017; Published: 18 October 2017

Abstract: Bike sharing programs are eco-friendly transportation systems that are widespread in smart city environments. In this paper, we study the problem of privacy-preserving bike sharing microdata publishing. Bike sharing systems collect visiting information along with user identity and make it public by removing the user identity. Even after excluding user identification, the published bike sharing dataset will not be protected against privacy disclosure risks. An adversary may arrange published datasets based on bike's visiting information to breach a user's privacy. In this paper, we propose a grouping based anonymization method to protect published bike sharing dataset from linking attacks. The proposed Grouping method ensures that the published bike sharing microdata will be protected from disclosure risks. Experimental results show that our approach can protect user privacy in the released datasets from disclosure risks and can keep more data utility compared with existing methods.

Keywords: bike sharing; identity disclosure; linking attacks; data publishing; privacy preservation

1. Introduction

Modern and urban lifestyles have directed physical activity out of everyday life, and this has resulted in the rise of warnings about people's health caused by inactive lifestyles [1]. An increment of physical activity, such as walking and cycling, is considered a preventative health measures. More importantly, walking and cycling represent one of the most active forms of daily life activities [2]. Urban computing implies an emerging interdisciplinary field that helps to integrate and analyze complex data produced by a variety of sources in urban environments. In cities, bike sharing systems establish the transportation systems that are becoming more and more popular [3,4]. Analysis of bike sharing data can significantly help to plan and improve policies of cities, and advancing the lives of their citizens [5]. Many bike sharing companies have started to publish their user data to the public without taking any precaution that published datasets might breach user's privacy. In addition, publishing of bike sharing data leads to disclosing the individual's privacy with their in-depth behavior and movement pattern, which is unacceptable.

The pervasive adoption of GPS-enabled smartphones and various location-based devices in combination with social media have led to an explosion of spatiotemporal datasets, i.e., mobile health-care records, and bike sharing transaction records [6]. Cities all over the world are collecting these data and making them open to the public [7]. For example, the U.S. government's open data site [8] implements executive orders on making government data available, and it also affirms that numerous countries, cities, and counties have launched open data sites to make data available. Collecting and analyzing of pervasive computing data lead to numerous challenges including

communication security, information security, and privacy violation, especially in health-care and personal data management [9,10]. To implement the pervasive based system, the service provider should introduce a robust framework and reliable encrypted communication system for protecting user privacy and security [11,12].

The data publisher has the microdata table in the form of *T*(*Identifier, Quasi-identifier, Sensitive attributes*) [13]. In the microdata Table *T*, some attributes comprise the *Identifier* that can identify a person, such as the social security number (SSN) or name. Some attributes determine *Quasi-identifier* (*QI*) such as Birth Year, Gender, Start Station and End Station, and when these *Quasi-identifier* attributes aggregate together it can identify a person. Some attributes express *Sensitive attributes*, which are unknown to the adversary and known as sensitive such as Start and End Station with Start and End Time. For data publishing, the identifier has been removed from the microdata table.

In the last decade, numerous incidents of data privacy breaches occurred because of personal data sharing resulting in financial and reputation disasters for organizations [14,15]. Table 1 exhibits an example of bike sharing transaction microdata, which remains open to the public [16]. User *ID* and *name* are never released during bike sharing data publishing. A user's sensitive information might still be leaked due to the presence of location and timing information [17]. Therefore, the requirement of anonymization implies breaking the relations among attribute values, so that a person will be indistinguishable in the released bike sharing transaction microdata table. Table 2 illustrates the *k-anonymized* [13] version of bike sharing transaction microdata Table 1.

Table 1. Bike-sharing transaction table.

Birth Year	Gender	Start Station	End Station	Start Time	End Time
1970	M	E 47 St & Park Ave	W 49 St & 8 Ave	4/1/2016 6:53	4/1/2016 6:59
1970	M	W 49 St & 8 Ave	E 47 St & Park Ave	4/1/2016 16:12	4/1/2016 16:19
1986	F	E 17 St & Broadway	E 27 St & 1 Ave	4/5/2016 13:48	4/5/2016 13:56
1986	F	E 27 St & 1 Ave	E 17 St & Broadway	4/6/2016 0:16	4/6/2016 0:25
1970	M	E 47 St & Park Ave	W 49 St & 8 Ave	4/7/2016 18:49	4/7/2016 18:55
1988	M	Broadway & E 22 St	E 20 St & 2 Ave	4/7/2016 6:06	4/7/2016 6:11
1970	M	W 49 St & 8 Ave	E 47 St & Park Ave	4/7/2016 17:56	4/7/2016 18:02
1981	M	8 Ave & W 52 St	W 63 St & Broadway	4/7/2016 6:47	4/7/2016 6:52
1986	F	E 27 St & 1 Ave	E 17 St & Broadway	4/20/2016 0:06	4/20/2016 0:15
1988	M	E 20 St & 2 Ave	Broadway & E 22 St	4/20/2016 8:41	4/20/2016 8:45
1983	M	Carlton Ave & Flushing Ave	Front St & Gold St	4/20/2016 11:16	4/20/2016 11:20
1983	F	W 21 St & 6 Ave	E 20 St & 2 Ave	4/21/2016 7:53	4/21/2016 8:00

In this paper, we present a new approach called Grouping which preserves user privacy against a linking attack [13,18]. In the anonymization, the Grouping method groups the dataset both vertically and horizontally. In the vertical grouping, location information and timing information have been grouped into the separate columns. In the horizontal grouping, the user travel records have been grouped in the buckets based on the trip duration. In a bucket, the attribute values randomly permute to break the correlation in between different columns. Hence, each *QI* value will be linked with *l* distinct sensitive values and will reduce the confidence that the adversary will have when breaching personal privacy [19,20]. Our proposed approach only generalizes the gender attribute column to ensure the user privacy. Therefore, it has less information loss and provides better data utility of the published datasets. A series of experiments on real-world dataset were conducted to support the effectiveness of the Grouping method.

The remainder of this paper is structured as follows. Section 2 presents background and related works. Section 3 describes the details of the Grouping method. Section 4 discusses the experimental analysis. Section 5 gives the conclusion.

Table 2. Anonymization table of bike sharing transaction table (*k-anonymity*).

Birth Year	Gender	Start Station	End Station	Start Time	End Time
(1970–1988)	Person	[E 47 St & Park Ave, Broadway & E 22 St, W 49 St & 8 Ave]	[W 49 St & 8 Ave, E 20 St & 2 Ave, E 47 St & Park Ave]	4/1/2016 6:53	4/1/2016 6:59
(1970–1988)	Person	[E 47 St & Park Ave, Broadway & E 22 St, W 49 St & 8 Ave]	[W 49 St & 8 Ave, E 20 St & 2 Ave, E 47 St & Park Ave]	4/7/2016 6:49	4/7/2016 6:55
(1970–1988)	Person	[E 47 St & Park Ave, Broadway & E 22 St, W 49 St & 8 Ave]	[W 49 St & 8 Ave, E 20 St & 2 Ave, E 47 St & Park Ave]	4/7/2016 18:06	4/7/2016 18:11
(1970–1988)	Person	[E 47 St & Park Ave, Broadway & E 22 St, W 49 St & 8 Ave]	[W 49 St & 8 Ave, E 20 St & 2 Ave, E 47 St & Park Ave]	4/7/2016 17:56	4/7/2016 18:02
(1970–1986)	Person	[E 20 St & 2 Ave, Carlton Ave & Flushing Ave, W 49 St & 8 Ave, E 17 St & Broadway]	[Broadway & E 22 St, Front St & Gold St, E 47 St & Park Ave, E 27 St & 1 Ave]	4/20/2016 8:41	4/20/2016 8:45
(1970–1986)	Person	[E 20 St & 2 Ave, Carlton Ave & Flushing Ave, W 49 St & 8 Ave, E 17 St & Broadway]	[Broadway & E 22 St, Front St & Gold St, E 47 St & Park Ave, E 27 St & 1 Ave]	4/20/2016 11:16	4/20/2016 11:20
(1970–1986)	Person	[E 20 St & 2 Ave, Carlton Ave & Flushing Ave, W 49 St & 8 Ave, E 17 St & Broadway]	[Broadway & E 22 St, Front St & Gold St, E 47 St & Park Ave, E 27 St & 1 Ave]	4/1/2016 16:12	4/1/2016 16:19
(1970–1986)	Person	[E 20 St & 2 Ave, Carlton Ave & Flushing Ave, W 49 St & 8 Ave, E 17 St & Broadway]	[Broadway & E 22 St, Front St & Gold St, E 47 St & Park Ave, E 27 St & 1 Ave]	4/5/2016 13:48	4/5/2016 13:56
(1970–1986)	Person	[E 27 St & 1 Ave, 8 Ave & W 52 St, W 21 St & 6 Ave]	[E 17 St & Broadway, W 63 St & Broadway, E 20 St & 2 Ave]	4/6/2016 0:16	4/6/2016 0:25
(1970–1986)	Person	[E 27 St & 1 Ave,8 Ave & W 52 St, W 21 St & 6 Ave]	[E 17 St & Broadway,W 63 St & Broadway, E 20 St & 2 Ave]	4/7/2016 6:47	4/7/2016 6:52
(1970–1986)	Person	[E 27 St & 1 Ave, 8 Ave & W 52 St, W 21 St & 6 Ave]	[E 17 St & Broadway, W 63 St & Broadway, E 20 St & 2 Ave]	4/20/2016 0:06	4/20/2016 0:15
(1970–1986)	Person	[E 27 St & 1 Ave,8 Ave & W 52 St, W 21 St & 6 Ave]	[E 17 St & Broadway, W 63 St & Broadway, E 20 St & 2 Ave]	4/21/2016 7:53	4/21/2016 8:00

2. Background and Related Works

In this section, we describe the privacy-preserving context, privacy-preserving data publishing methodology, and existing anonymization techniques. In addition, we discuss background knowledge with the privacy threats in bike sharing data publishing.

2.1. Privacy-Preserving Context

For preserving user privacy, we have to define a meaningful privacy context for the privacy-preserving data publishing. To determine specific privacy context, recently published research [21–24] identified the necessary privacy terms for the cyberspace, and these are the sender, recipient, attacker, identifiability, anonymity, pseudonymity, unlinkability, undetectability, unobservability, identity confidentiality and identity management. Pfitzmann and Hansen [22–24] describe a privacy setting that defines the relationship between essential privacy terms.

In the privacy context, a sender sends his data to a recipient where an attacker cannot gain any information about that data. This privacy setting could be followed in the privacy-preserving data publishing circumstance. In the privacy-preserving data publishing setting, a data publisher releases the data to the public, and it is open to everyone. An attacker also receives that published data, and he might use some background knowledge to identify a person by linking with some publicly available data sources [13]. Hence, the demand for anonymity is necessarily present in the privacy-preserving data publishing context [21]. Anonymity is the anonymous properties of a dataset in which an attacker cannot identify the record owner within a set of other records, which is called the anonymity set [24]. By applying some anonymizations operation on the published dataset, we can create the anonymity set which will protect the dataset from creating such link to identify a person. Consequently, the anonymous dataset will be protected from linking attacks and it will ensure the identity confidentiality in the published dataset.

2.2. Privacy-Preserving Data Publishing

To publish a dataset, there are trusted and untrusted model of data publishing [25]. In the untrusted publishing model, a data publisher might attempt to identify a user record and corresponding sensitive information from the dataset. Several anonymous communications and cryptographic solutions were proposed for collecting user data anonymously [26]. In the trusted privacy model, a data publisher remains trustworthy, and the record owners are reliable to provide their personal information for further processing. For example, a patient is ready to give her medical records to a hospital to obtain the needed medical service. Privacy-preserving data publishing is a trusted model of data publishing.

The data publishing states that an organization is the data owner and the public is the data miner who wants to do significant research on the published dataset. An organization wants to publish its own microdata table T to the public. Microdata table T could be released directly to the public if it contains no sensitive information. Usually, a microdata table T contains sensitive information and the data owner cannot give T to the public in the raw format. When a dataset has been published to the other parties for data mining, privacy-preserving techniques are mandatory to reduce the possibilities of identifying the sensitive information about a person [27]. In privacy-preserving data publishing, one conjecture is that the data receiver could be an adversary. For example, a data mining research center is a responsible entity, but every staff in that organization will not be accountable as well. This hypothesis makes the privacy-preserving data publishing problems and solutions to be very distinct from the encryption and cryptographic methods, in which only authorized and reliable receivers are allowed for the private key toward accessing the cipher text [27,28]. For privacy-preserving data publishing, the published microdata tables stand open to everyone. A significant challenge in privacy-preserving data publishing is to protect the privacy of a user without

disclosing their sensitive information. In addition, we have to ensure the data utility with the data privacy, namely the published dataset can be used for data mining and knowledge discovery.

To preserve sensitive information, an anonymization techniques need to be applied to a published microdata table. The anonymization approach tries to protect the identity and the sensitive information of a user, assuming that sensitive data must be preserved for data analysis. At the time of data publishing, unique identifiers of a user must be removed from the datasets. Even after all unique identifiers being removed, Sweeney [13] showed a real-life privacy threat to a former governor of the state of Massachusetts. In Sweeney's example, a user's name in a public voter list was linked to his record in a released medical dataset through the combination of zip code, date of birth, and sex. Each of these attributes does not uniquely identify a record owner. However, the aggregation of these attributes, which is *QI*, usually find a unique or a small number of record holders. According to Sweeney [13], 87% of the residents of the USA could be uniquely identified by using *QI* attributes.

In the above example, the user is identified by linking his *QI* attributes. To perform such linking attacks [13,18], the adversary needs two pieces of prior knowledge: (1) published dataset of a record holder and (2) the *QI* of the user. To limit linking attacks on the published dataset, we can provide an anonymization version of microdata table T^*(*Quasi-identifier, Sensitive attributes*) by applying anonymization operation such as randomization [19], generalization [29] and perturbation [30]. In the modified microdata table T^*, *QI* is an anonymous version of the original *QI* of the primary dataset. Anonymization processes preserve some particular information so that numerous records become indistinguishable with respect to *QI* values. If a person remains linked to a record through *QI*, and the same person is connected to all other records that have the same *QI* values, then the association will be ambiguous. The anonymization process produces an anonymous version of T such that it satisfies a given privacy model like *k-anonymity* [13] or *l-diversity* [31], and preserves as much data utility as possible.

2.3. Anonymization Techniques

For data anonymization, there are several privacy models, such as the partition-based model [13,31], the randomization-based model [19], and the differential privacy based model [32]. Among them, partition-based and randomization-based techniques are popular for privacy-preserving data publishing. In addition, recently differential privacy has received significant consideration for privacy-preserving data publishing. In the partitioning and randomization methods, the data values of *QI* (e.g., birth year, gender, start and end location) are generalized to construct a *QI* group. Therefore, an individual cannot be identified with their sensitive values in the group. Conversely, differential privacy answers the statistical queries based on the user request.

In partition and randomization based anonymization techniques, there are some popular anonymization methods which have been proposed for privacy preserving in one-time data publishing. Among them, *k-anonymity* [13] and *l-diversity* [31] methods are more popular. *k-anonymity* [13] is the first proposed privacy model for data publishing. It requires that all records in a published dataset cannot be distinguished from at least *k-1* other records. *k-anonymity* does not consider sensitive attributes so that attackers may learn the relationship between sensitive data and individuals through a background knowledge attack. Background knowledge attack was proposed in [31]. Background knowledge attack means that an adversary could use background knowledge to discover sensitive information, such as general understanding or professional knowledge about the published dataset.

To address the drawbacks of the *k*-anonymity methodology, the *l*-diversity privacy model was proposed [31]. A *QI* group is *l*-diverse if the probability that any tuple in this group remains associated with a sensitive value is at most $1/l$. Both *k*-anonymity and *l*-diversity use the generalization technique to anonymize the microdata. For the generalization, the data table loses an enormous amount of information, particularly for higher dimensional data [19]. Generalization breaks the correlation between attributes and it assumes that any possible combinations of attribute values are equally

possible, i.e., Person (Figure 1) expresses the generalized value for the attribute values Female and Male.

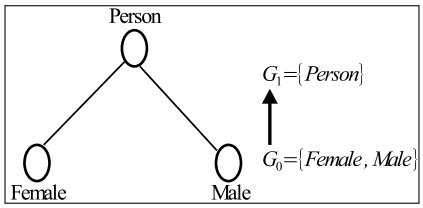

Figure 1. A taxonomy for the Gender attribute.

Generalization implies an anonymization technique to coarsen the values by mapping a value to an interval or a particular concept to a more general one [29]. It has been widely used for data anonymization. An advantage of generalization holds its faithfulness that a generalized value remains coarse but semantically consistent to its original value. For example, when age 24 is generalized to (20–30), we know that the age assumes a value in between 20 to 30 and it cannot assume a value of 32. Due to the faithfulness, when an adversary sees a set of records containing the values of a user ID, the adversary knows that the user's record remains in the set. Such information facilitates the disclosure risks. Therefore, generalization alone cannot prevent a dataset from a privacy attack [14].

The bike sharing companies published their datasets in every month, and each released dataset may contain multiple riding information of a single user. We consider that a registered user frequently rides a bike and, for every ride, the dataset will have her riding information. The bike sharing dataset contains the start and end locations of a riding path with the start and end time. A user future movement patterns can be predicted from the past locations [33]. In addition, some research works show that our actions are easily predictable by nature [34,35]. When a dataset contains user information more than one time then, by arranging the *QI* values, an adversary may know the user's identity with the visiting locations, which may lead to a privacy breach [33].

Privacy protection for a single dataset has been extensively studied where we have considered the information of a user remains only one time. When the information of an individual remains multiple times in multiple datasets or even on the same dataset, an adversary may reveal the privacy of the individual [36–39]. Recently, published *hybrid* [14], and *sequential* [38] methods have been proposed for privacy-preserving sequential data publishing. These methods used generalization [29] and perturbation [30] to anonymize the *QI* values and Sensitive values. Therefore, it reduces the published data utility. In this paper, we propose a Grouping anonymization method based on partition and randomization approach to anonymize the published bike sharing transaction microdata table.

In differential privacy, ϵ-*differential* [32] provides a strong privacy guarantee for statistical query answering. A survey on differential privacy can be found in [40]. Most of the differential privacy methods support interactive settings to satisfy the ϵ-*differential* privacy requirements. Mohammed [41] proposed the first non-interactive setting based algorithm for differentially private data release that protects information for classification analysis. In differential privacy, datasets play a vital role to check the effectiveness of the anonymization techniques. Li [42] proposed DPSynthesizer, an open-source toolkit for differentially private data synthesis. DPSynthesizer performs a set of state-of-the-art techniques for building differentially private histograms from which synthetic data can be created, and it is eligible for low-dimensional data.

Recently, the cloud platform has become preferred for data management. Cloud computing facilitates end-users to outsource their dataset to a third-party service provider for data management. In the cloud platform, the security and privacy become a major concern for outsourcing data. One of the significant security interests of the outsourcing paradigm is how to protect sensitive information in the outsourced dataset. Dong [43–45] proposed data-cleaning-as-a-service (DCaS) paradigm focusing on functional dependency (FD) constraints against data security attack by encrypting a small amount

of non-sensitive data. In addition, they designed an FD preserving encryption algorithm that can provide a provable security guarantee against the frequency analysis attack.

2.4. Background Knowledge and Privacy Threats

The adversary's background knowledge is described as the experience that he already learned and discovered formally from the prior rules of published datasets, or informally from the life experiences. For example, some sensitive attribute values such as ovarian cancer and breast cancer are associated with females only. The adversary's background knowledge assists with learning relevant sensitive information and finding sensitive records to breach individual privacy in published datasets.

An adversary may know an individual who rides a public bicycle to a bus stop from his home. Consequently, the adversary knows the person's start and end station, approximates start and end time, and the gender value. This information might be used as the *QI* values to search in the published bike sharing dataset to find all of the user's probable visiting places that would breach the user's privacy.

For bike sharing data publishing, published datasets work as a background knowledge because of its identity nature. An adversary may arrange released bike sharing datasets based on start and end station and could find a person who frequently rides a bike, and the adversary is not sure about the actual identity of the person. Therefore, an adversary might use video surveillance systems [46–49] to know the particular bike user.

2.5. Problems in Bike Sharing Data Publishing

Bike sharing data publishing presents a new challenge for privacy and utility for the published microdata table. Bike sharing data publishing is distinct from the traditional multiple time data publications, such as multiple view data publication [50,51] and series data publication [38,39,52], since, in a bike sharing dataset, a single user's records exist multiple times in the dataset. An adversary may use only a single release to conduct the privacy attacks and carry on with every other version of the microdata table. In addition, traditional multiple time data publishing [38,39,52] uses Generalization [29] to anonymize the microdata table that decreases the data utility.

A bike sharing microdata table consists of user's riding transaction records that are visited locations with timing information, and these are called points of interests (*POI*). A *POI* can be any place for a person such as home, workplace, sports center or political party's office. In an attack against user privacy, the attacker applies user-specific travel information to breach the user's privacy [53]. The objective of this attack is to identify a user's house, workplace, and behavior. Analysis of bike sharing application data could cause a serious privacy breach of any user that might reconstruct her social networks, knowledge of her favorite visited places, her political and religious views.

In the bike sharing published microdata table, an adversary may reveal a user identity and sensitive information by arranging user's records. For example, a public bike sharing user rides a bicycle to reach a bus stop to go to her workplace and come back home on weekdays. Therefore, the bike sharing transaction database will have the user's records for every time she rides a bicycle. An adversary may know her bus stop and arrange her records based on the bus stop and can know her Birth Year, Gender, Start Station and Start Time, which may lead to identifying the Stop Station and Stop Time. By using Birth Year, Gender, Start Station and Stop Station, the adversary may further search in the published microdata table and can arrange all of the user's available records for that particular route. The adversary could use these details of visiting information to initiate a physical or financial harassment to the user [54].

We explain an example of how a bike sharing transaction dataset may breach a user's privacy. Table 1 presents the data segment from a bike sharing company [16]. By arranging the Birth Year, Gender, Start Station and End Station, we may find the person who frequently uses the bike sharing service. It is observed that a bike sharing transaction microdata table has multiple records of a single user. From Table 1, it can be recognized that a male user born in 1970 often rides a bike from E 47 St and Park Ave to W 49 St and 8 Ave during the morning session and W 49 St and 8 Ave to E 47 St and

Park Ave during the afternoon session. By arranging a particular user's bike riding information, an adversary might visit those places in person to infer the actual user identity. At present, the adversary has the identity of the person with his *POI*. The adversary might be a thief, and from the summary of the above information, the adversary could steal from his house. Therefore, we can conclude that, for bike sharing datasets, an adversary could breach the user privacy based on the published bike sharing transaction microdata table.

3. Methodology

In this section, we present the bike sharing data publishing methodology and anonymization algorithm, which will ensure user's privacy in the published dataset.

3.1. Preliminaries and Problem Definitions

Typically, a bike sharing data publisher has a table T of records with $d+1$ attributes, $A = \{(A_1, ..., A_d), S\}$ and the attribute domains are $\{ID, D[A_1], ..., D[A_d], D[S]\}$, where ID is the unique identity of a user. ID is removed before every release of a dataset to the public. A tuple $t \in T$ can be expressed as $t = (t[A_1], t[A_2], ..., t[A_d], t[S])$.

A set of attributes $t[A_i](1 \leq i \leq d)$ of a bike sharing transaction table T is called a QI if these attributes together can uniquely identify at least one user. Let Table T be the bike sharing transaction Table as shown in Table 1. A QI of T is $\{Birth\ Year, Gender, Start\ Station, End\ Station\}$ might uniquely identify a user.

Some attributes in the bike sharing dataset are regarded as *Sensitive attributes* because they reveal the *POI* of a user. For example, Start and End Station with Start and End Time together can identify a person's visiting places and which we call a *POI* [53,55]. In bike sharing dataset, $t[S]$ defines as *sensitive attributes*. Let Table T be the bike sharing transaction Table as shown in Table 1. A *sensitive attributes* of T is $\{Start\ Station, End\ Station, Start\ Time, End\ Time\}$ might breach the user privacy by revealing *POI*. The bike sharing dataset QI and *sensitive attributes* share some common attributes with each other, i.e., $\{Start\ Station, End\ Station\}$.

In order to publish bike sharing datasets, we have to satisfy a privacy requirement. A user U rides any bike at different times t_i; $i \geq 1$ of a day D. Therefore, the table T has the multiple entries of a user U and publishing of the table T may violate the privacy requirements. We assume that, in a bike sharing transaction table where a user U exists more than once and as a subscriber of the bike sharing system, her records would appear in every release. Published bike sharing datasets and QI values should satisfy the privacy requirements, and, in this research, we adopt the *l-diversity* privacy requirements [31]. A QI group is said to be *l-diversity* [31], if the probability that any tuple in this group is associated with a sensitive value is at most $1/l$. For the bike sharing data publishing, we know that QI values and *sensitive values* are overlapped with each other. Hence, we define the *l-diversity* as in Slicing [19], where the $\{Birth\ Year, Gender\ Start\ Station, End\ Station\}$ will be associated with more $\{Start\ Time, End\ Time\}$.

3.2. Grouping Methods

In this section, we develop solutions from the problem definitions and design anonymization algorithms. The anonymization algorithm consists of the following steps: attribute grouping, tuple grouping with random permutation, and column generalization.

In attributes grouping, the related attributes are grouped in a fashion such that each attribute belongs to one subset, i.e., start location and end location form a subset. Each grouped subset forms a column. Particularly in a bike sharing table T, there will be c columns $C_1, C_2, ..., C_c$ satisfying $\bigcup_{i=1} C_i = A$ and for any $1 \leq i_1 \neq i_2 \leq c$, $C_{i_1} \cap C_{i_2} = \emptyset$.

We group the related attributes by measuring the correlation between them. For the bike sharing data publishing, we consider most of the attributes to be categorical. The correlation between attributes are calculated by the mean square contingency coefficient [19,56]. Given two attributes A_1 and A_2

with value domains $\{v_{11}, v_{12}, ..., v_{1d_1}\}$ and $\{v_{21}, v_{22}, ..., v_{2d_2}\}$, respectively, and domain sizes d_1 and d_2, the mean square contingency coefficient between attributes A_1 and A_2 is calculated as follows:

$$\phi^2(A_1, A_2) = \frac{1}{min\{d_1, d_2\} - 1} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{(n_{ij} - n_{i.}n_{.j})^2}{n_{i.}n_{.j}}$$

where $n_{i.}$ and $n_{.j}$ are the fractions of occurrence of v_{1i} and v_{2j} in the data, respectively. n_{ij} is the fraction of cooccurrence of v_{1i} and v_{2j} in the data. Therefore, $n_{i.}$ and $n_{.j}$ are the marginal totals of $n_{ij} : n_{i.} = \sum_{j=1}^{d_2} n_{ij}$ and $n_{.j} = \sum_{i=1}^{d_1} n_{ij}$. It can be shown that $0 \leq \phi^2(A_1, A_2) \leq 1$.

After estimating the correlation between attributes, we use the widely-known *k-medoid* clustering algorithm PAM (Partition Around Medoids) [57] to partition attributes into columns. In the algorithm, each attribute constitutes a point in the cluster space. The difference between two attributes is determined as $d(A_1, A_2) = 1 - \phi^2(A_1, A_2)$, which lies between 0 and 1. In the space, two correlated attributes will have a smaller dissimilarity between the corresponding data points.

We group attributes so that highly correlated attributes are placed in the same column. This approach is useful for privacy and data utility [19]. Regarding data utility, grouping highly correlated attributes protects the relationships among those attributes. In terms of data privacy, associating correlated attributes reduces the identification risks. Because relationships between uncorrelated attribute values are less common and hence more identifiable, it is desirable to break the associations between uncorrelated attributes to preserve privacy.

The tuple grouping consists of different subsets of Table T, such that each tuple relates to a specific subset based on their trip duration. Every subset of tuples denotes a bucket B [19]. Let T be a bike sharing table that contains b buckets $B_i (1 \leq i \leq b)$. Every bucket B contains n tuples $t_j (1 \leq j \leq n)$. The Mondrian [58] algorithm is used, and it follows a top–down approach without generalization for separating tuples to create a bucket. For each bucket, the values in every Column are randomly permuted to break the cross-column correlations. Hence, it will break the linkage between the *QI* values and sensitive values and reduce the adversary's confidence in linking with the *POI*.

In the bike sharing datasets, the Gender column holds an essential factor to breach personal privacy. Therefore, we use generalization methodology [29] to generalize the Gender column to preserve user privacy in the published bike sharing dataset.

3.3. Algorithms

In this section, we present our algorithm from protecting published datasets from disclosure risks. Two algorithms are introduced to perform the anonymization process. Algorithms 1 and 2 can successfully finish the anonymization process for protecting user privacy in published datasets. The primary objective is to ensure the user privacy by satisfying the *l-diversity* [19] privacy requirements, and increases the data utility as well.

Algorithm 1 Anonymization

Input: Bike Sharing Data set T
Output: Anonymized Data set T*

1: For a given dataset T generates an anonymized table T*, which will satisfy the privacy concept with the privacy requirement R of *l*-diversity;
2: $B = \emptyset$;
3: **for** each tuples in T **do**
4: group tuples into i buckets $\{B_1, ..., B_i\}$ as in Mondrian [58].
5: permute attribute values in each bucket B_i
6: Privacy Check (B_i, R)
7: $T^* = T^* \cup B_i$;
return T*.

Algorithm 2 Privacy Check(B, *R*)

Input: Bucket *B*
Output: TRUE, if the bucket satisfies privacy requirement *R*

1: **for** each tuple in bucket *B* **do**
2: Check the *l-diversity* of every tuple to satisfy privacy requirement *R* as in [19];
return TRUE.

3.3.1. Anonymization Algorithm

Our anonymization algorithm performs the anonymization process in the following ways. Initially, *B* contains no bucket. In each iteration (lines 3 to 7), the anonymization algorithm groups the table into buckets according to the Mondrian [58] criteria. In line 5, we permute the attribute values in each column to break the correlation between cross columns. In line 6, privacy is checked by the Privacy Check algorithm and appends the bucket with *T**. Finally, the anonymized table *T** is published.

3.3.2. Privacy Check Algorithm

Our privacy check algorithm checks the privacy requirements *R* in each bucket. In line 2, we have checked the *l-diversity* privacy requirement as in slicing [19].

3.3.3. Time Complexity

The time complexity of anonymization algorithm depends on Mondrian [58], and, for *l-diversity*, privacy checking depends on Slicing [19]. The Mondrian algorithm requires $O(nlogn)$ because at each level the whole dataset needs to scan $O(n)$ times and the Mondrian algorithm requires n heights of the tree is $O(logn)$. To check *l-diversity* requirements, it takes $O(n)$. Therefore, the total time complexity is $O(nlogn)$.

3.4. Discussion on the Anonymization Techniques

In this subsection, we illustrate how the proposed Grouping method can protect the user *POI* by satisfying the *l*-diversity [19] privacy requirements. From the bike sharing transaction table (Table 1), we see that a user with the Birth Year of 1970 has four records in total. Therefore, this user is selected as a representative to find his *POI*, and we would like to verify whether our proposed method can protect his privacy or not.

We consider that a bike sharing user with the Birth Year 1970 and an adversary wish to infer the user's locations and timing information (i.e., *POI*) from the published anonymized Table 3. In order to determine the user's *POI*, the adversary has to examine its matching bucket. By checking the Birth Year column, it is found that the user exists in both buckets. From the nature of a bike sharing transaction table, we know that a person's record will appear multiple times in the published dataset. From the published microdata table, the adversary finds that, in the first bucket, there are two records for the Birth Year 1970. The adversary is not sure that the user is a male or a female because the Gender column carries generalized values.

By examining the third column (Start station, End Station) first bucket, the adversary finds that there are six records and among them two records (record number two and five) are similar. The adversary might assume that these records represent the locations where the particular user visited. The adversary needs to link the locations with the timing information to breach the user's privacy.

Table 3. Anonymization table of bike sharing transaction table (Grouping).

Birth Year	Gender	(Start Station, End Station)	(Start Time, End Time)
1983	Person	(E 20 St & 2 Ave, Broadway & E 22 St)	(4/7/2016 6:49, 4/7/2016 6:55)
1988	Person	(E 47 St & Park Ave, W 49 St & 8 Ave)	(4/20/2016 8:41, 4/20/2016 8:45)
1970	Person	(8 Ave & W 52 St, W 63 St & Broadway)	(4/20/2016 11:16, 4/20/2016 11:20)
1988	Person	(Carlton Ave & Flushing Ave, Front St & Gold St)	(4/7/2016 18:06, 4/7/2016 18:11)
1981	Person	(E 47 St & Park Ave, W 49 St & 8 Ave)	(4/7/2016 6:47, 4/7/2016 6:52)
1970	Person	(Broadway & E 22 St, E 20 St & 2 Ave)	(4/1/2016 6:53, 4/1/2016 6:59)
1986	Person	(E 17 St & Broadway, E 27 St & 1 Ave)	(4/5/2016 13:48, 4/5/2016 13:56)
1970	Person	(W 49 St & 8 Ave, E 47 St & Park Ave)	(4/7/2016 17:56, 4/7/2016 18:02)
1983	Person	(W 49 St & 8 Ave, E 47 St & Park Ave)	(4/1/2016 16:12, 4/1/2016 16:19)
1986	Person	(W 21 St & 6 Ave, E 20 St & 2 Ave)	(4/21/2016 7:53, 4/21/2016 8:00)
1970	Person	(E 27 St & 1 Ave, E 17 St & Broadway)	(4/6/2016 0:16, 4/6/2016 0:25)
1986	Person	(E 27 St & 1 Ave, E 17 St & Broadway)	(4/20/2016 0:06, 4/20/2016 0:15)

In the first bucket of the fourth column (Start Time, End Time), the adversary finds that there are a total of six different time stamps. In order to determine the *POI*, the adversary needs to associate the location information with the timing information. However, the location information associated with six different time stamps make the adversary uncertain about the user's actual visiting time in that place. Therefore, we can say that the proposed Grouping method might preserve user privacy by breaking the linking amongst the columns.

By carefully observing the first bucket, we discover that the Birth Year 1970 is linked with two persons (the Gender column is generalized, and, in generalization, we have to consider all possible values), one visiting place, and six different time stamps. It makes a total of 12 possible combinations or records, which satisfy the *l-diversity* privacy requirements as in Slicing [19].

4. Experimental Analysis

In this section, we conduct the experiments on the bike sharing dataset received from Citi Bike of New York City (NY, USA) [16]. Citi Bike publishes their dataset on a monthly basis. For the analysis, we use the published dataset from January 2016 to May 2016. It has a total of 4,215,675 records with fifteen attribute values. We consider six essential attributes in the dataset to conduct the anonymization simulation, and datasets are described in the Table 4.

Table 4. Data set description of bike sharing transaction table.

	Attribute	Type	Number of Unique Values
1	Birth Year	Continuous	87
2	Gender	Categorical	2
3	Start Station	Categorical	475
4	End Station	Categorical	483
5	Start Time	Date	2,331,112
6	End Time	Date	2,330,582

The experiments are divided into two parts: the first part is designed to see the effectiveness of the proposed Grouping method against the disclosure risks, and, in the second part, data utility is measured on the published anonymized dataset. In the experimental analysis, we have fixed 10-*anonymity* for the Generalization method [13] and 10-*diversity* for our proposed Grouping method.

4.1. Disclosure

At the time of bike sharing transaction table publishing, the data publisher releases the dataset in a raw format by removing identifiers. To calculate the disclosure risks in the bike sharing dataset, we count all the records as the *QI* values. We group the vulnerable *QI* values based on their birth year,

gender, start locations and end locations. Figure 2 shows the representation of datasets with unique *QI* values and probable disclosure risks from January 2016 to May 2016. The primary objective of the anonymization technique continues to break the correlation between these *QI* values to ensure personal privacy.

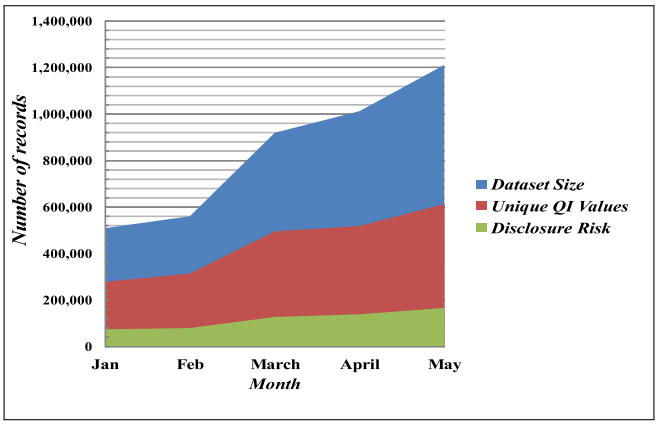

Figure 2. Bike sharing dataset properties.

We further calculate the disclosure risks in the published anonymized dataset. To calculate the disclosure risks, we compare the original dataset *QI* values with the anonymized dataset *QI* values. If a record in the anonymized dataset matches the original dataset *QI* values, then we count it as disclosed *QI* values. Disclosure risks are calculated by the following formula [59]:

$$Disclosure = \frac{Matched\ records}{Total\ records} \times 100\%.$$

Figure 3 shows disclosure risks for the *k-anonymity* [13] and the Grouping method. On the *x*-axis, it shows datasets from January to May and on the *y*-axis represents the corresponding disclosure risks. The experimental results show that the Grouping method has less disclosure risks than the *k-anonymity* method:

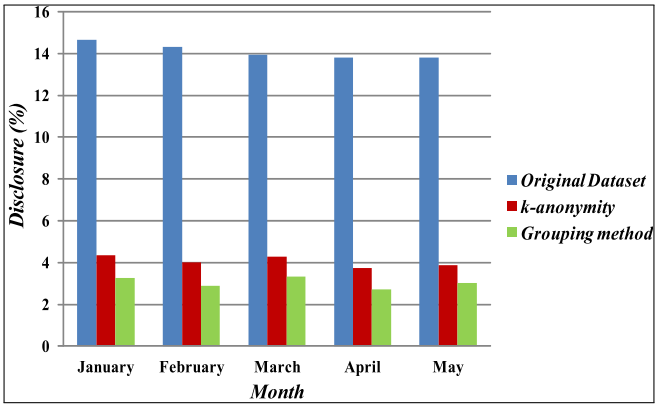

Figure 3. Disclosure risks.

4.2. Data Utility Comparison

The data utility experiment extends into two parts. In the first part, the data utility is measured based on the distortion ratio during the anonymization process for *k-anonymity* [13] and the proposed Grouping method. In the second part, the data utility is computed using the relative error in the aggregate query.

4.2.1. Data Utility

For publishing the microdata table, privacy preservation is an important issue, and we have to consider the data utility as well. Data loss metric shows how much data utility remains in the published bike sharing dataset. There are many methods to calculate the data loss in the published dataset. The distortion ratio describes a process to calculate the data loss in the published bike sharing transaction table [27].

To calculate the distortion ratio, we have to consider that every attribute of the bike sharing dataset associated with a generalized taxonomy [27]. Figures 1 and 4 present the taxonomy for attributes Gender and Birth Year, respectively. If the value of a tuple does not generalize, the distortion of that value is 0, and if it is generalized, the distortion is defined by the height of the taxonomy tree. For instance, the Gender Male/Female is not generalized, and it resides in the leaf node, so the height indicates 0, and thus the distortion equals to 0. If it is generalized one level up in the taxonomy tree, the distortion equals $1/H$. Here, H indicates the height of the taxonomy tree. Let $d_{j,k}$ be the distortion of the value of attribute A_j of tuple t_k. The distortion of the whole published table corresponds to the sum of the distortions of all values in the generalized dataset. Thus, the distortion is denoted as:

$$\sum_{j=1,k=1}^{n,m} d_{j,k}.$$

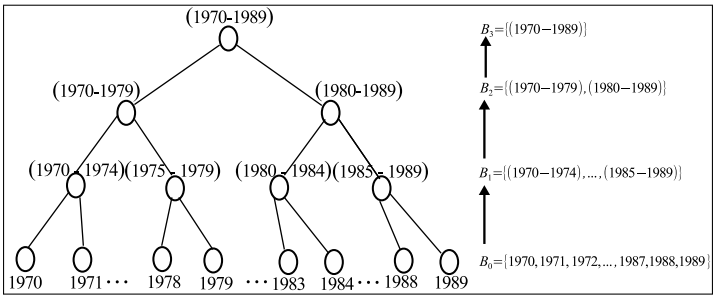

Figure 4. A taxonomy for attribute Birth Year.

The distortion ratio is $D_R = \frac{D_P}{D_G}$, where D_R denotes the distortion ratio, D_P means the distortion of the datasets in the published table, and D_G signifies the distortion of the fully generalized (i.e., all values generalized to the root of the taxonomy trees) table. Data utility is calculated by subtracting distortion ratio D_R from 100% [27]:

$$Data\ utility = (100 - D_R)\%.$$

Figure 5 shows the experimental results of the data utility based on information loss from the published datasets. In the original dataset, the data utility is 100% because it has 0 distortions. By contrast, our proposed method and *k-anonymity* [13] have 25% and 60% distortion, respectively. Therefore, we can say that the Grouping method has more data utility than the generalization approach for the *k-anonymity* method.

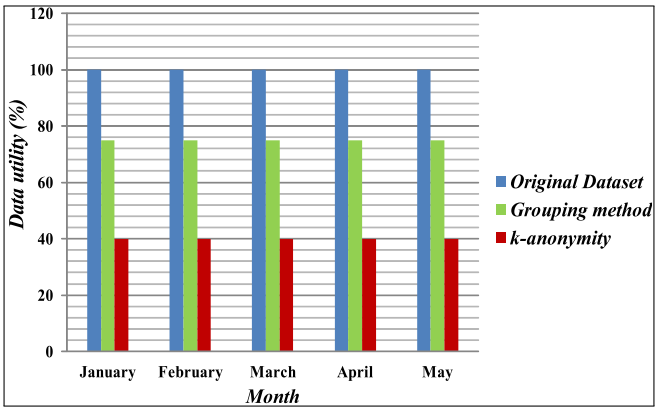

Figure 5. Data utility on published table.

4.2.2. Aggregate Query Error

The effectiveness of aggregate query [60] is evaluated by estimating the data utility in the published datasets. In the experiment, "COUNT" operator is computed where the query predicate comprises the start and end locations with time in the following form:

$SELECT\ COUNT(*) FROM\ Table$

$WHERE\ v_{i_1} \in V_{i_1} \text{AND} \ldots v_{i_{dim}} \in V_{i_{dim}} \text{AND}\ s \in V_s$

where $v_{i_j} (1 \leq j \leq dim)$ indicates the QI value for attribute A_{i_j}, $V_{i_j} \subseteq D_{i_j}$ and D_{i_j} comprises the domain for attribute A_{i_j}, s represents the sensitive attribute value and $V_s \subseteq D_s$ and D_s implies the domain for the sensitive attribute S. A query predicate can be characterized by predicate dimension dim and query selectivity sel. Dimension dim indicating the number of QIs exist in the predicate and selectivity sel indicating number of values in each $V_{i_j} (1 \leq j \leq dim)$. Specifically, the size of $V_{i_j} (1 \leq j \leq dim)$ is randomly chosen from $0, 1, \ldots, sel * |D_{i_j}|$. Each query executes on the original table and all anonymized tables. Original count is given as org_{count} and anonymized count is given as anz_{count}, where anz_{count} is considered for all anonymization methods, respectively. The average relative error is computed over all queries as [60]:

$$Relative\ error = \frac{|anz_{count} - org_{count}|}{org_{count}} \times 100\%.$$

In Figure 6, relative query error is plotted on the y-axis based on the QI selection. In the experiment, we have selected birth year, gender, birth year and gender, birth year, gender, start and end location, respectively, as QI attributes and calculated the relative query error on the anonymized dataset by *k-anonymity* [13], and the proposed Grouping method. For the experiment, we have selected the January 2016 dataset, and all possible combinations of the query were made and executed through the anonymization table while we calculated the average relative query error. The relative query error is calculated and presented in Figure 6, where the value on the y-axis denotes relative percentage error and those on the x-axis stands for different QI selection. The experimental result shows that proposed anonymization has small relative errors compared with the generalization approach for the *k-anonymity* method.

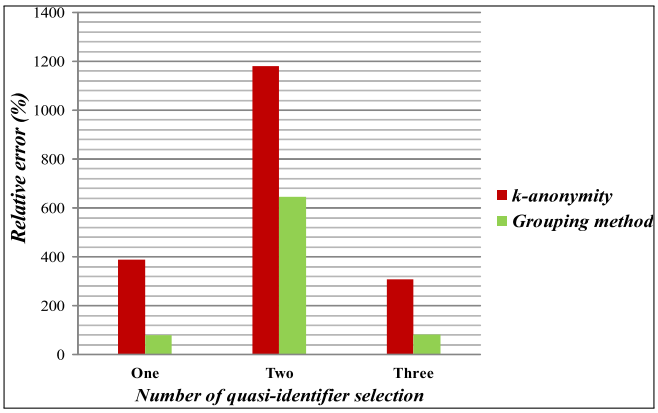

Figure 6. Aggregate query answering error.

5. Conclusions

This paper proposed an anonymization technique for bike sharing datasets to minimize the probability of a successful linking attack. Our proposed Grouping anonymization method can successfully break the correlation between the *QI* values with the sensitive values in order to protect users from disclosure risks. We experimentally showed that the Grouping method has smaller privacy disclosure risks while it is published to the public. The Grouping method only generalizes the gender attribute. Therefore, it provides better data utility for the published datasets. The experimental results demonstrate that the Grouping method has higher data utility and smaller relative query error as compared to the other methods.

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. 2015A080804019; and sponsored by the CAS-TWAS President's Fellowship for International Ph.D. students.

Author Contributions: A S M Touhidul Hasan 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, Qingshan Jiang and Chenming Li wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Fishman, E.; Washington, S.; Haworth, N. Bikeshare's impact on active travel: Evidence from the united states, great britain, and australia. *J. Transp. Health* **2015**, *2*, 135–142.
2. Scheepers, E.; Wendel-Vos, W.; van Kempen, E.; Panis, L.I.; Maas, J.; Stipdonk, H.; Moerman, M.; den Hertog, F.; Staatsen, B.; van Wesemael, P.; et al. Personal and environmental characteristics associated with choice of active transport modes versus car use for different trip purposes of trips up to 7.5 kilometers in the netherlands. *PLoS ONE* **2013**, *8*, e73105, doi:10.1371/journal.pone.0073105.
3. Chourabi, H.; Nam, T.; Walker, S.; Gil-Garcia, J.R.; Mellouli, S.; Nahon, K.; Pardo, T.A.; Scholl, H.J. Understanding smart cities: An integrative framework. In Proceedings of the 2012 45th Hawaii International Conference on System Science (HICSS), Maui, HI, USA, 4–7 January 2012; pp. 2289–2297.
4. Meijer,A.; Bolívar, M.P.R. Governing the smart city: A review of the literature on smart urban governance. *Int. Rev. Adm. Sci.* **2016**, *82*, 392–408.
5. Dyson, L.; *Beyond Transparency: Open Data and the Future of Civic Innovation*; Code for America Press: San Francisco, CA, USA, 2013.
6. Dobre, C.; Xhafa, F. *Pervasive Computing: Next Generation Platforms for Intelligent Data Collection*; Morgan Kaufmann: Burlington, MA, USA, 2016.

7. Douriez, M.; Doraiswamy, H.; Freire, J.; Silva, C.T. Anonymizing nyc taxi data: Does it matter? In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016; pp. 140–148.
8. The Home of the U.S. Government's Open Data. 2017. Available online: https://www.data.gov (accessed on 14 October 2017).
9. Lu, R.; Lin, X.; Shen, X. Spoc: A secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency. *IEEE Trans. Parallel Distrib. Syst.* **2013**, *24*, 614–624.
10. Castiglione, A.; DAmbrosio, C.; De Santis, A.; Castiglione, A.; Palmieri, F. On secure data management in health-care environment. In Proceedings of the 2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), Taichung, Taiwan, 3–5 July 2013; pp. 666–671.
11. Gligoric, N.; Dimcic, T.; Drajic, D.; Krco, S.; Chu, N. Application-layer security mechanism for m2m communication over sms. In Proceedings of the 2012 20th Telecommunications Forum (TELFOR), Belgrade, Serbia, 20–22 November 2012; pp. 5–8.
12. Pizzolante, R.; Carpentieri, B.; Castiglione, A.; Castiglione, A.; Palmieri, F. Text compression and encryption through smart devices for mobile communication. In Proceedings of the 2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), Taichung, Taiwan, 3–5 July 2013; pp. 672–677.
13. Sweeney, L. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570.
14. Li, J.; Baig, M.M.; Sattar, A.S.; Ding, X.; Liu, J.; Vincent, M. A hybrid approach to prevent composition attacks for independent data releases. *Inf. Sci.* **2016**, *367–368*, 324–336.
15. Narayanan,A.; Shmatikov, V. Shmatikov how to break anonymity of the netflix prize dataset. *arXiv* **2006**, arXiv:cs/0610105.
16. Citi Bike Daily Ridership and Membership Data. Available online: https://www.citibikenyc.com/system-data (accessed on 3 April 2017).
17. 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, .
18. Bayardo, R.J.; Agrawal, R. Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering, 2005, ICDE 2005, Tokoyo, Japan, 5–8 April 2005; pp. 217–228.
19. Li, T.; Li, N.; Zhang, J.; Molloy, I. Slicing: A new approach for privacy preserving data publishing. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 561–574, .
20. Hasan, 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.
21. Kambourakis, G. Anonymity and closely related terms in the cyberspace: An analysis by example. *J. Inf. Secur. Appl.* **2014**, *19*, 2–17.
22. Pfitzmann, A.; Köhntopp, M. Anonymity, unobservability, and pseudonymity—A proposal for terminology. In *Designing Privacy Enhancing Technologies*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 1–9.
23. Pfitzmann, A.; Hansen, M. A Terminology for Talking About Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. 2010. Available online: http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf (accessed on 14 October 2017).
24. Hansen, M.; Smith, R.; Tschofenig, H. CA Privacy terminology and concepts. In *Internet Draft, March 2012*; Technical Report; Network Working Group, IETF: Fremont, CA, USA, 2011.
25. Gehrke, J. Models and methods for privacypreserving data publishing and analysis. In Proceedings of the 22nd International Conference on Data Engineering (ICDE), Atlanta, GA, USA, 3–7 April 2006; Volume 105.
26. Yang, Z.; Zhong, S.; Wright, R.N. Anonymity-preserving data collection. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA, 21–24 August 2005; pp. 334–343.
27. Wong, R.C.-W.; Fu, A.W.-C. Privacy-preserving data publishing: An overview. *Synth. Lect. Data Manag.* **2010**, 2, 1–138.
28. Taric, G.J.; Poovammal, E. A survey on privacy preserving data mining techniques. *Indian J. Sci. Technol.* **2017**, *8*, doi:10.17485/ijst/2017/v10i5/111138.
29. Samarati, P.; Sweeney, L. Generalizing data to provide anonymity when disclosing information. *PODS* **1998**, *98*, 188, doi:10.1145/275487.275508.

30. Kwan, M.-P.; Casas, I.; Schmitz, B. Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Int. J. Geogr. Inf. Sci.* **2004**, *39*, 15–28, .
31. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3, doi:10.1145/1217299.1217302.
32. Dwork, C. Differential privacy. In *IN ICALP*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.
33. 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.
34. Gonzalez, M.C.; Hidalgo, C.A.; Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782,
35. Song, C.; Qu, Z.; Blumm, N.; Barabási, A.-L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021,
36. Hasan, T.; Jiang, Q. A general framework for privacy preserving sequential data publishing. In Proceedings of the 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA), Taipei, Taiwan, 27–29 March 2017; pp. 519–524.
37. Ganta, S.R.; Kasiviswanathan, S.P.; Smith, A. Composition attacks and auxiliary information in data privacy. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 265–273.
38. Wang, K.; Fung, B. Anonymizing sequential releases. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 414–423.
39. Xiao, X.; Tao, Y. M-invariance: Towards privacy preserving re-publication of dynamic datasets. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 11–14 June 2007; pp. 689–700.
40. Dwork, C. Differential privacy: A survey of results. In *5th International Conference on Theory and Applications of Models of Computation*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 1–19.
41. Mohammed, N.; Chen, R.; Fung, B.; Yu, P.S. Differentially private data release for data mining. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 493–501.
42. Li, H.; Xiong, L.; Zhang, L.; Jiang, X. Dpsynthesizer: Differentially private data synthesizer for privacy preserving data sharing. *Proc. VLDB Endow.* **2014**, *7*, 1677–1680.
43. Dong, B.; Wang, W. Frequency-hiding dependency-preserving encryption for outsourced databases. In Proceedings of the IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA, 19–22 April 2017; pp. 721–732.
44. Dong, B.; Wang, W.; Yang, J. Secure data outsourcing with adversarial data dependency constraints. In Proceedings of the IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), New York, NY, USA, 9–10 April 2016; pp. 73–78.
45. Dong, B.; Liu, R.; Wang, W.H. Prada: Privacy-preserving data-deduplication-as-a-service. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai, China, 3–7 November 2014; pp. 1559–1568.
46. Wang, X. Intelligent multi-camera video surveillance: A review. *Pattern Recogn. Lett.* **2013**, *34*, 3–19,
47. Albano, P.; Bruno, A.; Carpentieri, B.; Castiglione, A.; Castiglione, A.; Palmieri, F.; Pizzolante, R.; Yim, K.; You, I. Secure and distributed video surveillance via portable devices. *J. Ambient Intell. Hum. Comput.* **2014**, *5*, 205–213.
48. Yadav, D.K.; Singh, K.; Kumari, S. Challenging issues of video surveillance system using internet of things in cloud environment. In *International Conference on Advances in Computing and Data Sciences*; Springer: Singapore, 2016; pp. 471–481.
49. Albano, P.; Bruno, A.; Carpentieri, B.; Castiglione, A.; Castiglione, A.; Palmieri, F.; Pizzolante, R.; You, I. A secure distributed video surveillance system based on portable devices. In *CD-ARES*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 403–415.

50. Yao, C.; Wang, X.S.; Jajodia, S. Checking for k-anonymity violation by views. In Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, 30 August–2 September 2005; pp. 910–921.
51. Yang, B.; Nakagawa, H.; Sato, I.; Sakuma, J. Collusion-resistant privacy-preserving data mining. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 483–492.
52. Wong, R.C.-W.; Fu, A.W.-C.; Liu, J.; Wang, K.; Xu, Y. Global privacy guarantee in serial data publishing. In Proceedings of the IEEE 26th International Conference on Data Engineering (ICDE), Long Beach, CA, USA, 1–6 March 2010; pp. 956–959.
53. Zhang, S.; Freundschuh, S.M.; Lenzer, K.; Zandbergen, P.A. The location swapping method for geomasking. *Cartogr. Geogr. Inf. Sci.* **2017**, *44*, 22–34,
54. Li, H.; Zhu, H.; Du, S.; Liang, X.; Shen, X. Privacy leakage of location sharing in mobile social networks: Attacks and defense. *IEEE Trans. Dependable Secur. Comput.* **2016**, doi:10.1109/TDSC.2016.2604383.
55. Hasan, 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 November 2016; ACM: New York, NY, USA; pp. 35–39.
56. Cramér, H. *Mathematical Methods of Statistics (PMS-9)*; Princeton University Press: Princeton, NJ, USA, 2016; Volume 9.
57. Kaufman, L.; Rousseeuw, P.J. Finding Groups in Data: An Introduction to Cluster Analysis; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 344.
58. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering, Atlanta, GA, USA, 3–7 April 2006; p. 25.
59. Geeen, K.; Tashman, L. Percentage error: What denominator? *Foresight Int. J. Appl. Forecast.* **2009**, *12*, 36–40.
60. 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.