Article

Data Re-Identification—A Case of Retrieving Masked Data from Electronic Toll Collection

Hsieh-Hong Huang ^{1,*}, Jian-Wei Lin ² and Chia-Hsuan Lin ¹¹ Department of Information Science and Management Systems, National Taitung University, Taitung 95092, Taiwan; u10212149@ms102.nttu.edu.tw² Department of International Business, Chien Hsin University of Science and Technology, Taoyuan 32097, Taiwan; jwlin@uch.edu.tw

* Correspondence: kory@nttu.edu.tw; Tel.: +886-8951-7685

Received: 30 March 2019; Accepted: 14 April 2019; Published: 16 April 2019

Abstract: With the growth of big data and open data in recent years, the importance of data anonymization is increasing. Original data need to be anonymized to prevent personal identification from being revealed before being released to the public. There is a growing variety of de-identification methods which have been proposed to reduce the privacy issues, however, there is still much to be improved. The purpose of this study is to demonstrate the possibilities of re-identification from masked data, and to compare the pros and cons of different de-identification methods. A set of electronic toll collection data from Taiwan was used and we successfully re-identified vehicles with specific patterns. Four de-identification methods were performed and finally we compared the strengths and weaknesses of these methods and evaluated their appropriateness.

Keywords: data anonymity; open data; de-identification; re-identification; electronic toll collection

1. Introduction

Cyber-physical road systems have made considerable progress through the development of technologies such as vehicle communications, internet of vehicles (IoVs), and smart transportation. For example, the IntelliDrive system in United States allows cars on the road to communicate over short distances in real-time through roadside facilities. Vehicle identification codes are also obtained when vehicles pass the system's sensors; this data is then used to measure traffic flows [1]. Although the installation of GPS in specific vehicles (e.g., buses for public transport) allows for the collection of information related to their driving dynamics, these systems have been installed by transportation enterprises and the data is only used internally. This prevents the development of new usages for this data or complementary linkages between different enterprises to supplement data-related deficiencies, thus limiting the usefulness of the data.

Although the original purpose of the electronic toll collection (ETC) system was to replace manual toll collection, this system has provided the opportunity to use data mining techniques to extract real-time information on public transportation. The analysis of variables such as time, place, vehicles, and trips for the construction of models based on this data can be used to improve traffic management and assist travelers in understanding traffic data more accurately [2]. For example, the ETC system in Taiwan records 3.2 million trips and 20 million instances of toll data every day, on average. Immense benefits could be obtained if the information hidden within this data could be used properly. However, in the acquisition of toll collection data using the current ETC system, there are privacy concerns regarding this data as it includes road user identities. Therefore, this data must be de-identified before it is released to the public.

In the case of Taiwan's ETC system, the raw data acquired by system is only being used internally. The vehicle identification code field is deleted from the ETC data before it is openly released, which prevents the identification of vehicular information. However, whereas excessive de-identification measures can protect the privacy of road users, this also harms the usability of the data and makes it difficult to derive useful information from the data. In this study, we used the ETC data which have been de-identified provided by the freeway bureau of Taiwan to demonstrate that the re-identification of specific vehicles, i.e., the Freeway Schedule Bus Service (FSBS), is possible. In addition, we used different de-identification methods on the re-identified information of FSBS buses, to investigate the efficacy of each de-identification method and their effects on data granularity. The strengths, weaknesses, and scenarios for each method are investigated by validating the accuracy of the re-identified data after de-identification.

2. Background

2.1. Data Release and Open Data

Information systems are being used widely in numerous fields including business, the retail industry, banks, the telecommunications industry, and information industries. An immense quantity of data is generated in each of these fields; for example, the retail industry accumulates significant amounts of trade data and the telecommunications industry accumulates tremendous quantities of user-related data. Using telecommunications enterprises as an example, the information analysis platforms developed for this industry include functions such as network quality management, customer analysis, and analyses on business loss prevention. The use of data mining techniques and statistical analysis on this data allows different customer groups to be distinguished from the data [3]. Using the aforementioned analyses on each customer group, it is possible to produce marketing strategies tailored to each group to retain customers and improve service quality, hence improving the value of the data [4,5]. Big data is highly developable and generates tremendous value for enterprises as the usage of this data assists managers in making better decisions, resulting in increased benefits for enterprises.

Open data is a relatively recent trend in the information age. The release of open and transparent governmental data on network platforms for free public usage allows the public to understand the workings of the government. Further, this also allows the public to use their creativity to create multiple usages for these data that improve their value. However, the open release of data also presents certain challenges, particularly on the requirement for restrictions on open data. To respect privacy laws, data that can be traced to individuals must not be released openly. However, if the released information is severely degraded and the data does not possess an appropriate level of information quality, controversies, misunderstandings, and opacity can arise. The release of degraded information can reduce, in fact, the credibility of the information and that of the provider of this information [6].

Open data can be used in different fashions that add value to the data. Moreover, owing to the importance of governmental open data, big data trends in data usage and the ever-increasing body of data being accumulated by different industries, the collection, analysis, and usage of data are now issues of extreme importance. In particular, big data refers to enormous quantities of complex and unprocessed information, which is impossible to manage and process by manual means; even the storage and processing of this data is difficult to perform with current software [7].

2.2. Data De-Identification and Privacy Issues

The publishing of open data or analyzed data on the internet for public browsing, downloading, and usage can involve personal privacy issues. This data must therefore be processed using specific techniques to prevent the leakage of individual data and to prevent infringements of individual law. De-identification is the process for protecting personal information; this refers to the use of randomization, masking, anonymity, and concealment to prevent personal information leakage, by making it impossible for others to identify the personal data of specific persons [8]. Personal data that

has been de-identified to render impossible the identification of individuals from this data by direct or indirect means are no longer recognized as personal data in the eyes of the law; data can only be released openly for public usage after this has been achieved.

As de-identified data is no longer treated as personal data, it should be impossible to acquire the identities of specific persons from this data. That is, de-identification is only considered successful if it is impossible to identify individuals even with tracking or comparative techniques [9]. In view of data privacy issues, Xu et al. (2014) highlighted privacy problems that are possible at each level (data providers, data collectors, data miners, and decision makers) and provided guidance for the circumvention of these issues [10]. For example, data providers can provide information to data collectors to obtain certain benefits; the solution is to remind these providers that the release of their data online could compromise their privacy and reveal their tracks. Privacy problems and leakages of the providers' personal data can also occur if data collectors are not careful in their processing of private data during data collection, or if the information they release is collected and analyzed by their competitors in attempts to restore the original information. The release of any data to the public should be screened by testing mechanisms to determine if privacy leakages could result from these releases. Decision makers can contract with data miners to ensure that the mined data cannot be leaked to competitors or lead to personal data leakages, to protect the company's business [11].

2.3. De-Identification Methods and Standards

There are a number of de-identification techniques existed for processing personal information to de-identified, and k-anonymity is the most famous one [12]. Sweeney (2002) proposed a model to protect privacy when disclosing information, termed k-anonymity. K-anonymity is a method for data protection privacy, which lets all the data have at least k-1 item which is similar to the original data [13]. By replacing attribute values with a generalized version of them (generalization) or removing sensitive information (suppression), k-anonymity can prevent the data from being identified and reduce the ability to aggregate, analyze, and draw sensitive inferences from individuals' data [14]. In this study, some of the de-identification methods we used for evaluation, e.g., averaging privacy field data, are variants of k-anonymity. Through anonymity, people are able to protect personal information and to manage the boundaries between their private and public spheres in advance [15].

To aim at the requirements of anonymity and help organizations implementing data de-identification processes for privacy enhancing purposes, the International Organization for Standardization (ISO) has proposed a series of methods for data de-identification, such as the ISO 20889 and ISO 29100 series [16]. The ISO 29100 and ISO 29191 standards provide an additional level of protection for big data linkages and open data [17] and also alleviate the concerns of the public and scientific researchers on privacy invasions or inadvertent illegal invasions of personal data [18]. ISO 29192-1 to ISO 29192-5 are technical standards for the security of small quantities of information and include mechanisms such as block cipher, stream cipher, and asymmetric encryption [19].

3. Materials and Methods

The primary focus of this work is to re-identify the patterns of the FSBS buses in the ETC data. The knowledge discovery in the database procedure proposed by Fayyad et al. [11] is used to analyze and process the ETC data. This includes procedures such as the selection of data, preprocessing, data conversion, data mining, and explanation and evaluation.

3.1. Research Materials: Etc Data

When Taiwan constructed the ETC system, support for the collection of traffic data was considered, which resulted in the ETC system being used to collect traffic data. This data is openly published on governmental open data platforms to promote its usage and to enhance its value and quality [20]. Owing to privacy considerations, only highly de-identified versions of raw data on paths taken during each trip (M06A) is released by the governmental open data platforms. The crucial Unicode field has

been deleted, thus removing direct and indirect associations between the data; the granularity of the data consequently is large because of the impossibility of direct or indirect discrimination of the data. The ETC data used in this study is moderately de-identified versions of the M06A, which retains the Unicode field, however, with its identification numbers converted into unique 32-bit values using an unpublished method. The M06A dataset used in this work records the route information of each vehicle from its point of entry up to its exit from an interchange. The content of this data is included in the following Table 1.

Table 1. Electronic toll collection (ETC) data schema.

Field Name	Description
UniCode:	Vehicle identification code. eTagID or license plate number of the vehicle was converted into a unique code using an unpublished encryption method.
VehicleType:	Vehicle type code, which has values such as “31” (small passenger vehicle), “32” (small truck), “41” (bus), “42” (truck), or “5” (trailer).
DetectionTime_O:	Time when the vehicle passes its first detection station during this trip.
GantryID_O:	Code number of the first detection station passed by the vehicle during this trip.
DetectionTime_D:	Time when the vehicle passes its final detection station during this trip.
GantryID_D:	Code number of the last detection station passed by the vehicle during this trip.
TripLength:	Total travelled distance during this trip.
TripEnd:	Trip notes, “Y” denotes a trip with a normal ending, “N” denotes a trip with an abnormal ending.
TripInformation:	Code numbers of the detection stations passed by the vehicle during this trip and the times corresponding to each of these passes.

The UniCode field was originally the vehicle license plate or eTagID, which is unique information that can be traced to the vehicle owner who registered the license plate or ID; the granularity of this information is therefore extremely small. To protect this personal data, the eTagIDs or vehicle license plates were encrypted into 32-bit vehicle identification codes. The information in the first field was hidden in the M06A dataset; this is a de-identification measure that is also intended to remove direct and indirect associations between the data.

3.2. Research Design and Procedure

This study was performed over two stages. During the high-frequency stage, it was proven that the conversions used for the de-identification of vehicle identification codes in the ETC data did not preclude the risk of specific subjects being identified via comparative methods. In the low-frequency stage, four different methods of de-identification were compared. In the low-frequency stage, the previously identified ETC data of specific vehicles was used as test data. The test data was then de-identified using four different methods and attempts were then made to re-identify the data. The de-identification efficacies, data granularity, and strengths and weaknesses of these methods were then compared, and the appropriate usage scenarios for each method were determined through observations and comparisons. We then discussed how de-identification measures for data privacy protection could be balanced with the retention of data granularity. The findings of this study can therefore act as a reference for subsequent technical developments in data de-identification.

3.3. Model Explanation and Evaluation

In this study, after a model for re-recognizing the FSBS buses was developed, a confusion matrix was used to evaluate the accuracy of the re-identified data and cross validation was used to evaluate the strengths and weaknesses of the re-identification model’s algorithms. In cross validation, the acquired datasets were divided into sub-datasets of appropriate size. One of the sub-datasets was

used as testing data for validating the effectiveness of the model and the remaining sub-datasets were used as training data for model construction. In this manner, each sub-dataset was validated using test data and the reliability of the model was guaranteed through repeated validations.

3.4. Background Analysis of the Dataset

We obtained the following information by analyzing the raw data by vehicle type: in the data interval from 16/10/2015 to 23/10/2015, 25,561,407 trips were made over eight days, where 17,372,861 of these trips were made by small passenger cars. This accounts for 67.79% of the total number of trips, making small passenger cars the most common type of vehicle traveling on the national highways. This was followed by small trucks, with 5,581,615 records (21.84%), making them the second largest group of vehicles on national highways. Small trucks and small passenger cars therefore accounted for approximately 90% of the trips made on national highways. Buses made 386,604 trips, or 1.51% of the total number of trips, making them the least numerous of the vehicles traveling on the national highways. Trucks made 1,230,397 trips, or 4.81% of the total number of trips. Trailers made 989,929 trips, or 3.87% of the total number of trips. In Table 2, this data is organized to display the number of trips by each type of vehicle on each day as a percentage of the total number of trips made on that particular day. The data was compartmentalized into the first four days and last four days, which were used as training and testing/validation datasets, respectively.

Table 2. Trip data analysis (by vehicle type and day).

	Day	31 Cars	32 Small Trucks	41 Buses	42 Trucks	5 Trailers	Total
Training set	Day 1	2,322,209 (66.16%)	794,931 (22.65%)	53,517 (1.52%)	187,276 (5.34%)	152,250 (4.34%)	3,510,183
	Day 2	2,412,406 (70.95%)	708,370 (20.83%)	52,673 (1.55%)	123,066 (3.62%)	103,598 (3.05%)	3,400,113
	Day 3	2,385,249 (76.03%)	595,603 (18.99%)	51,576 (1.64%)	54,770 (1.75%)	49,965 (1.59%)	3,137,163
	Day 4	2,181,427 (66.29%)	741,507 (22.53%)	49,563 (1.51%)	175,877 (5.34%)	142,274 (4.32%)	3,290,648
Testing set	Day 5	2,097,515 (65.02%)	736,207 (22.82%)	48,963 (1.52%)	190,739 (5.91%)	152,290 (4.72%)	3,225,714
	Day 6	2,072,792 (66.75%)	711,646 (22.57%)	46,599 (1.48%)	176,445 (5.60%)	145,156 (4.60%)	3,152,638
	Day 7	2,116,849 (65.57%)	734,594 (22.75%)	47,797 (1.48%)	184,132 (5.70%)	144,969 (4.49%)	3,228,341
	Day 8	1,784,414 (68.20%)	558,757 (21.35%)	35,916 (1.37%)	138,092 (5.28%)	99,427 (3.80%)	2,616,606
Summary		17,372,861 (67.97%)	5,581,615 (21.84%)	386,604 (1.51%)	1,230,397 (4.81%)	989,929 (3.87%)	25,561,407

The data analysis and data mining performed in this work is focused mainly on bus data. Compared to other types of vehicles, buses can be divided into company buses, FSBS buses, and tour buses, many of which tend to have regular routes or time patterns. FSBS buses in particular must apply for right-of-way, and only drive along approved routes, which makes their patterns even more distinct. As mentioned above, buses account for 1.51% of the total number of trips of all vehicles. By dividing the data into hourly units, and comparing the distribution of buses with the overall traffic flow at each time point, it can be observed that there are distinct peak and off-peak periods in the data. The comparison between the distributions of buses and overall traffic flow is illustrated in Figure 1.

Figure 1. Comparison between distribution of buses and overall traffic flow.

4. Results and Discussion

4.1. Identification of the Regularly Driven Buses

The data for the first four days was used as the training dataset for the data mining performed to re-identify the FSBS buses. The last four days was used as the validation dataset, which was used to test if the re-identified FSBS buses retained the same patterns.

First, the interchange routes traveled by every “41” bus on each day were counted, with each vehicle being distinguished using their fixed and unique vehicle identification codes. If a bus traveled through the same starting interchanges more than four times, it was selected as a high frequency FSBS bus. After the data was compartmentalized into four-day blocks, all of the data that did not belong to a previously identified FSBS vehicle was aggregated and analyzed. Any bus that had driven through the same starting interchanges more than four times within four days was then selected as a low frequency FSBS bus. The low frequency and high frequency vehicle identification codes (buses) were compared to the data corresponding to the last four days to observe if the aforementioned codes reappeared in the validation dataset. After this comparison, the buses that reappeared on each of the last four days were then counted to calculate the reappearance rate of the FSBS buses and to validate the accuracy of the determinations.

We performed re-identification using a two-staged approach. 1067 and 1011 vehicles were selected as FSBS buses during the low and high frequency stages, respectively, to obtain 2078 FSBS buses. In the comparison with the validation data (the last four days), the reappearance rate of vehicles was highest on the Day 5, with 970 and 934 vehicles reappearing in the low and high frequency stages, respectively, for a total of 1904 vehicles. The probabilities on the Day 6, Day 7, and Day 8 were also 88% or greater. 1480 vehicles reappeared on all four days, giving a probability of 71.22%. It is thus demonstrated that the FSBS buses were re-identified with a high level of accuracy, the re-identification and validation results are described in detail in Table 3.

Table 3. Results of re-identified FSBS buses and validation.

	Period	Re-Identified		Validation	
		Low Frequency	High Frequency	Total	Hitting Rate
Training set	Days 1–4	1067	1011	2078	
Testing set	Day 5	970	934	1904	91.63%
	Day 6	917	925	1842	88.64%
	Day 7	926	918	1844	88.74%
	Day 8	926	932	1858	89.41%
	Continuous hit in all 4 days	702	778	1480	71.22%

4.2. De-Identification Methods

4.2.1. Deletion of Privacy Fields

The deletion of privacy fields, as the name suggests, simply refers to the deletion of privacy data fields that must be encrypted. We have chosen to delete the UniCode field, which contains fixed and unique vehicle identification codes for each vehicle. This leaves six fields: DetectionTime_O, GantryID_O, initial interchange, DetectionTime_D, GantryID_D, and the final interchange. Note that the VehicleType, TripLength, TripEnd, and TripInformation fields, which are not necessary for this study, were deleted in advance, during the preparation of the research data. The “initial interchange” and “final interchange” fields were added to the data and the original interchange codes were converted into Mandarin to facilitate observation and comparison. As the results of deleting the UniCode field, it is impossible to distinguish any single vehicle and only general information remained for traffic analysis.

4.2.2. Cryptographic Salting

Salting refers to the use of an additional input to a one-way function to hash data. The code derived from a hash function with salt would be considered an identifying element because the resulting value would be susceptible to compromise by the recipient of such data [21]. In this study, the UniCode field was encoded using the MD5 algorithm to obtain a fixed-length 128-bit value, making it difficult to recover the original value.

The results of this experiment are as follows: 1011 and 1067 vehicles were selected as FSBS buses during the high and low frequency stages, respectively, to obtain 2078 vehicles. From the analysis of this data (Table 4), it can be observed that these results are identical to those obtained from the original data (Table 3). The original UniCode field contained fixed and unique values; although the MD5 conversion increased the difficulty of identifying vehicle identification codes and made it almost impossible to recover the original codes, the uniqueness and constancy of the original data was not altered by this procedure. Therefore, the previous computational re-identifications operations were able to discover the tracks of the FSBS buses.

Table 4. Results of re-identification via cryptographic salting.

	Period	Re-Identified		Validation	
		Low Frequency	High Frequency	Total	Hitting Rate
Training set	Days 1–4	1011	1067	2078	
Testing set	Day 5	934	970	1904	91.63%
	Day 6	925	917	1842	88.64%
	Day 7	918	926	1844	88.74%
	Day 8	932	926	1858	89.41%
	Continuous hit in all 4 days	778	702	1480	71.22%

4.2.3. Modifying Privacy Field Data

The modification of privacy field data entails the removal of the last three codes from the information being encrypted, or the replacement of these codes with an arbitrary symbol. We replaced the last three codes of the UniCode field with “000” (for example: “12345” became “12”, and any value less than 1000 became zero). Consequently, the range (i.e., number of vehicles represented by each value) of every value expands and the uniqueness of the vehicle identification code is lost. The codes themselves became considerably simplified; for example, the highest valued vehicle identification code, which was “1690126” became “1690”. The data blocks with the same identification code increased significantly and the results of the statistical analysis were significantly altered.

As indicated in Table 5, 560 and 384 vehicles were identified in the high and low frequency stages, respectively, for a total of 944 vehicles. The number of vehicles determined to travel on all four days also exhibited an extremely high rate of reappearance, with all vehicles having reappearance rates

of 95% and greater. However, the vehicle identification codes no longer represented unique vehicles. The ambiguity in data boundaries and difficulties in discrimination caused by excessively wide data ranges severely distorted this statistical analysis. Therefore, these results were essentially meaningless.

Table 5. Results of re-identification via privacy field.

	Period	Re-Identified		Validation	
		Low Frequency	High Frequency	Total	Hitting Rate
Training set	Days 1–4	560	384	944	
Testing set	Day 5	555	380	935	99.05%
	Day 6	556	382	938	99.36%
	Day 7	556	379	935	99.05%
	Day 8	555	376	931	98.62%
	Continuous hit in all 4 days	549	369	918	97.25%

4.2.4. Averaging Privacy Field Data

To average the privacy field data, the data to be encrypted is separated and compartmentalized, and the average value of each compartment is used to represent all of the privacy field data within the compartment. The GantryID_O and GantryID_D fields were selected for this method because averaging the UniCode field would expand the interval of vehicle identification codes and yield the same results as the previous de-identification method. It can be observed from comparisons that the averaging of interchange IDs in a route will expand the area of vehicle routes and make it difficult to determine vehicle positions, which subsequently renders the acquisition of vehicle routes via comparisons impossible. First, the values of the GantryID_O field were divided into three groups; every 15 to 20 codes were then grouped into sets according to the order of the codes. All of the codes in these groups were uniformly represented by the codes of the first interchange in each group. The same process was then performed on the GantryID_D field.

As indicated in Table 6, 1648 and 1054 vehicles were identified as FSBS buses in the high and low frequency stages, respectively, for a total of 2702 vehicles, which is 624 more than the original number of identified vehicles. This indicates that the interchange area was expanded, such that the probability of the same car driving through the same initial interchanges repeatedly also increased, thus increasing the number of selected vehicles. It can be observed that the identification of FSBS buses was not sufficiently accurate; although the first set of reappearance data appeared to be highly accurate, the final number of vehicles that reappeared on all four days was only 73.25% of the total.

Table 6. Results of re-identification via privacy field averaging.

	Period	Re-Identified		Validation	
		Low Frequency	High Frequency	Total	Hitting Rate
Training set	Days 1–4	1648	1054	2702	
Testing set	Day 5	1460	1241	2701	99.96%
	Day 6	1426	1163	2589	95.82%
	Day 7	1419	1151	2570	95.11%
	Day 8	1422	1154	2576	95.34%
	Continuous hit in all 4 days	1130	757	1887	73.25%

In the first iteration of this de-identification method, 15 to 20 codes were assigned to each division based on an ascending order of codes. However, errors could have occurred in the sequence of codes assigned to each region due to variations in the speed of construction of each division. Therefore, we chose to adopt a different method, where the GantryID_O fields were divided according to the location of each interchange code into 15 major regions. The same was then performed for GantryID_D. We investigated the results obtained when an improved definition of the de-identification procedure

was provided. The results are shown in Table 7. It is demonstrated that division of the interchange codes into 15 regions expanded the intervals for the selection of vehicles traveling repeatedly (four times or more) on the same interchanges, and increased the number of data instances within each interval. However, when comparisons were made with the data on Day 7 and Day 8, it was revealed that the accuracy the data was relatively low, as the percentage of vehicles reappearing on all four days was only 56.01%. This result demonstrates that the re-identification of FSBS buses was not accurate in this case.

Table 7. Results of re-identification via privacy field data division and averaging.

	Period	Re-Identified		Validation	
		Low Frequency	High Frequency	Total	Hitting Rate
Training set	Days 1–4	1642	1768	3410	
Testing set	Day 5	1395	1417	2812	82.46%
	Day 6	1384	1329	2713	79.56%
	Day 7	1338	1304	2642	77.48%
	Day 8	1360	1295	2655	77.86%
	Continuous hit in all 4 days	1064	846	1910	56.01%

4.3. Discussions

The deletion of privacy fields method is the most secure of the four methods discussed in this paper. The data that remains after the deletion of this field cannot be used to uniquely represent a vehicle or to observe a unique vehicle. All that can be accomplished with the remaining data are quantity-based analyses as the data cannot be used to describe the routes travelled by a vehicle, which makes the granularity of the data large. Consequently, the usability of this data becomes limited. Although this method ensures that the private parts of the data cannot be deduced through comparisons, it also makes the release of this data meaningless because further work on this data cannot improve its usability.

The cryptographic salting method, where spurious information is added to the privacy field, followed by encryption using other algorithms to make it more difficult to recover the original data. This improves the security of the privacy field and obfuscates the contents of the original data. However, this does not alter the one-to-one correspondence of the original de-identification method. Consequently, the results obtained with this method are the same as those of the original data. All that is changed is the time required for the comparisons due to the increased complexity of the privacy field. Because the uniqueness of the privacy field is unaltered, the FSBS buses and other related information can be identified with ease from the de-identified data. Hence, this method does not improve on the shortcomings of the currently available ETC data as it has not altered the granularity of the data. The only change is that the privacy field's contents are more complex and more difficult to observe.

The modifying privacy field data method changes the original one-to-one de-identification to a many-to-one privacy field conversion. The de-identified vehicle identification codes no longer represent unique vehicles, as they now represent a range of codes, thus expanding the range of each code value. This increase in range seemingly causes the probability of repetitions to increase, and superficially suggests that further research can be performed using this data. This data can only be used as a reference however, because vehicles cannot be identified with any accuracy. The extreme lack of accuracy in the determinations makes it impossible to validate any of the conclusions drawn from this data. Although this method of de-identification improves on the privacy-related weaknesses of the original method and makes it impossible for FSBS buses to be identified using association rules, the granularity of the resulting data is overly coarse. The resulting distortions to the data and the ambiguity in the definition of the data intervals make it exceedingly challenging to draw useful conclusions from this data. Consequently, in-depth studies and applications based on this data are also extremely difficult to implement.

The averaging of privacy field data method is another way that improves on one-to-one de-identification measures. In this study, the interchange codes were first sorted according to their numerical values and compartmentalized for averaging; it was hoped that the codes would be ordered from north to south. However, the results exhibited excessively high reappearance rates in the first four days, which were not consistent with the actual number of vehicles that reappeared continuously over the last four days. It was later determined that this could have been caused by differences in the newly constructed order of interchanges. Therefore, we divided the interchanges into divisions based on their geographical locations and averaged their values using 15 regions. The results indicate that this method increased data granularity; however, it did not lead to ambiguity-related distortions and inaccuracies. The uniqueness of the vehicle identification codes, which is an important indicator in this study, was retained in this method. Although the usability of this data for detailed studies was been improved over the original data, the results we obtained demonstrate a level of credibility, as the FSBS buses were ultimately identifiable using comparative methods. Further, in this method, the low accuracy of the determinations is sufficient for preventing privacy leakages from the open data being investigated.

Comparisons were then made between the degree of re-identification, reappearance rates, and validated accuracies obtained from the data produced by each de-identification method. We summarize the empirical results of this study in Table 8. Also, we examined and compared each de-identification method to observe its strengths and weaknesses. The results of this comparison are summarized in Table 9.

Table 8. Results of re-identification and validation using different methods.

		Re-Identification Using Training Set		Validation Using Testing Set	
		Number of Re-Identified Vehicles	Re-Identified Ratio	Hitting Rate	Validated Accuracy
De-identification methods	Original data	2078	Medium	Low	High
	Deletion of privacy fields	0	None	High	High
	Cryptographic salting	2078	Medium	Low	High
	Modifying privacy fields	944	Low	High	Low
	Averaging privacy fields	2702	Medium	Medium	Medium
	Division-averaging privacy fields	3410	High	Medium	High

Table 9. Comparison of benefits of each de-identification method.

De-Identification Methods	Processing Time	Code Length	Coding Difficulty	Data Usability	Privacy Security	Data Credibility
Original data	-	-	-	High	Low	High
Deletion of privacy fields	Short	Short	Easy	Low	High	High
Cryptographic salting	Long	Short	Easy	High	Low	High
Modifying privacy fields	Short	Short	Easy	Low	High	Low
Averaging privacy fields	Medium	Medium	Medium	Medium	Medium	Medium
Division-averaging privacy fields	Medium	Medium	Hard	Medium	Medium	High

Processing time: time required for each de-identification method to de-identify the original data. Code length: length of the codes written for the data de-identification programs. Coding difficulty: difficulty writing the code for the data de-identification programs. Data usability: usability of the data for in-depth usage. Privacy security: likelihood of personal data leakages caused by comparative methods. Data credibility: value of the conclusions drawn from in-depth study of the data.

5. Conclusions

In this study, we successfully re-identified FSBS buses from the open data of the ETC system. Four different methods of de-identification were examined via re-identification tests, where the accuracies of the data re-identifications were validated using test data. In the case of ETC data, we can observe that vulnerabilities may remain in these de-identification methods. Large quantities of data can be compared to acquire private data, which poses an issue for information security.

In this study, we used different de-identification methods to investigate how improvements could be made in this aspect. Each type of de-identification method produced different effects on data granularity and indirectly affected the extensibility of subsequent studies using the de-identified data. Each de-identification method also had a different effect on privacy security issues. Through the investigations performed in this study, we determined that the ideal usage environment for the original data is to entrust it to trustworthy experts and scholars, as the details of the original data can provide the most useful information in big data mining; however, there are sources of concern for information security. It was found that the salting method for field conversion and obfuscation yielded the same results as the original de-identification method. De-identification via the removal of privacy field data leaves a limited amount of usable information, which can only be used for rough statistical analyses. This method is highly secure as it largely precludes the possibility of privacy leakages; however, the extensibility of the resulting data is extremely limited. Data that has been de-identified through the modification of privacy fields can be provided to students for educational purposes, or released to the public to help them understand the data collected by the ETC system. However, in-depth studies using this data are not recommended owing to the ambiguities of the data, which can lead to erroneous results. The averaging of privacy field data converts detailed parts of the information into simplified classifications; however, these classifications continue to be affected by certain defects. By improving the definition of the divisions, we ultimately determined that privacy field division and averaging is the optimal method of de-identification; this does not result in excessively coarse levels of data granularity and is sufficiently detailed for use by the public or scholars in other fields. Although FSBS buses can be re-identified from the de-identified data, the low level of accuracy of these identifications cannot lead to a significant leakage of private data.

Author Contributions: Conceptualization, H.-H. H.; Data curation, C.-H. L.; Project administration, H.-H. H.; Validation, J.-W. L.; Writing – original draft, C.-H. L.; Writing – review & editing, H.-H. H.

Funding: This study was partially supported by Ministry of Science and Technology of Taiwan, grant number MOST 105-2815-C-143-009-H.

Acknowledgments: Data for this study came from the Freeway Bureau, Ministry of Transportation and Communications, Taiwan.

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

References

1. Zhou, Y.; Mo, Z.; Xiao, Q.; Chen, S.; Yin, Y. Privacy-Preserving Transportation Traffic Measurement in Intelligent Cyber-physical Road Systems. *IEEE Trans. Veh. Technol.* **2016**, *65*, 3749–3759. [CrossRef]
2. Weng, J.; Yuan, R.; Wang, R.; Wang, C. Freeway Travel Speed Calculation Model Based on ETC Transaction Data. *Comput. Intell. Neurosci.* **2014**, *2014*, 48. [CrossRef] [PubMed]
3. Hand, D.J.; Mannila, H.; Smyth, P. *Principles of Data Mining*; MIT Press: Cambridge, MA, USA, 2001.
4. Tan, P.-N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Pearson Education India: Chennai, India, 2005.
5. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2011.
6. Janssen, M.; Charalabidis, Y.; Zuiderwijk, A. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Inf. Syst. Manag.* **2012**, *29*, 258–268. [CrossRef]
7. Snijders, C.; Matzat, U.; Reips, U.-D. “Big Data”: Big Gaps of Knowledge in the Field of Internet Science. *Int. J. Int. Sci.* **2012**, *7*, 1–5.
8. Van Devender, M.S.; Glisson, W.B.; Benton, R.; Grispos, G. Understanding De-identification of Healthcare Big Data. *Proc. Twenty-Third Am. Conf. Inf. Syst.* 2017. Available online: <https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1457&context=amcis2017> (accessed on 16 April 2019).
9. Bettini, C.; Riboni, D. Privacy Protection in Pervasive Systems: State of the Art and Technical Challenges. *Pervasive Mob. Comput.* **2015**, *17*, 159–174. [CrossRef]
10. Xu, L.; Jiang, C.; Wang, J.; Yuan, J.; Ren, Y. Information Security in Big Data: Privacy and Data Mining. *IEEE Access* **2014**, *2*, 1149–1176.

11. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **1996**, *17*, 37–54.
12. Ito, K.; Kogure, J.; Shimoyama, T.; Tsuda, H. De-identification and Encryption Technologies to Protect Personal Information. *Fujitsu Sci. Tech. J.* **2016**, *52*, 28–36.
13. Sweeney, L. k-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2002**, *10*, 557–570. [[CrossRef](#)]
14. Babu, K.S.; Jena, S.K. Balancing between Utility and Privacy for k-Anonymity. *Commun. Comput. Inf. Sci.* **2011**, *191*, 1–8. [[CrossRef](#)]
15. Acquisti, A.; Brandimarte, L.; Loewenstein, G. Privacy and Human Behavior in the Age of Information. *Science* **2015**, *30*, 509–514. [[CrossRef](#)]
16. Politou, E.; Michota, A.; Alepis, E.; Pocs, M.; Patsakis, C. Backups and the Right to be Forgotten in the GDPR: An Uneasy Relationship. *Comput. Law Secur. Rev.* **2018**, *34*, 1247–1257. [[CrossRef](#)]
17. Fal', O.M. Standardization in Personal Data Protection. *Cybern. Syst. Anal.* **2014**, *50*, 324–326. [[CrossRef](#)]
18. Yu, S. Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. *IEEE Access* **2016**, *4*, 2751–2763. [[CrossRef](#)]
19. Mitchell, C.J. Challenges in Standardising Cryptography. *Int. J. Inf. Secur. Sci.* **2016**, *5*, 29–38.
20. Fan, S.-K.S.; Su, C.-J.; Nien, H.-T.; Tsai, P.-F.; Cheng, C.-Y. Using Machine Learning and Big Data Approaches to Predict Travel Time Based on Historical and Real-Time Data from Taiwan Electronic Toll Collection. *Soft Comput.* **2018**, *22*, 5707–5718. [[CrossRef](#)]
21. U.S. Department of Health and Human Services. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act. (HIPAA) Privacy Rule*; U.S. Department of Health and Human Services: Washington, DC, USA, 2012.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).