Risk of Re-identification from Payment Card Histories in Multiple Domains

Satoshi Ito, Reo Harada, Hiroaki Kikuchi
Graduate School of Advanced Mathematical Sciences, Meiji University
4-21-1 Nakano, Nakano Ku, Tokyo, 164-8525 Japan
cs172032@meiji.ac.jp, cs172050@meiji.ac.jp, kikn@meiji.ac.jp

***Abstract*—Anonymization is the process of modifying a data set to prevent the identification of individual people from the data. However, most studies consider only the anonymization of data from a single domain. No study has been made on the risk of re-identification from combined data sets involving more than one domain. This paper proposes an evaluation of the risk of re-identification from payment card histories in multiple domains. First, we model the correlation between two histories from different usage domains in terms of information entropy and use mutual information to quantify the risk of identification from the data. Second, we describe an experiment to evaluate the risk in payment card data. The results validated the proposed method for real payment card data from 31 subjects. Metrics for the privacy and utility of 47 anonymized data items were evaluated. Overall, we found that there was a correlation between the histories of transportation and item purchases stored in the payment card data and established that most (44 of 47) of the anonymized data enabled correct identification with more than 45% accuracy for any privacy metric. This indicates that the risk of re-identification from payment card data is very high.**

***Index Terms*—re-identification, payment card, anonymization, risk evaluation**

I. Introduction

Companies should evaluate the re-identification risks when anonymizing personal-identification data included in payment transactions before employing big data extensively in their business. For example, it is known that 87% of population of America are identified from combination of some information (zip code, sex, and birthday) [1]. Evaluation of anonymized data can be made from both privacy and utility perspectives. Anonymization is a process of modifying data to prevent individual people from being identified via information in the original data. Many evaluation indexes and anonymization methods for purchase-oriented data have been proposed [9], [8], [11], [10].

However, most studies consider the anonymization of data from only a *single domain*. The re-identification risk from data histories has been studied only in terms of transportation data alone [11], [10] or purchase data alone [12]. Not so many studies have been made yet on the risk of re-identification from *combined data sets* involving more than one domain. One of the reasons for this omission is that agreement has not been obtained for any personal data sets to be combined with other data. Moreover, public data sets related to the combined data are limited. In addition, it is not trivial to formalize the mathematical model for combined data resources with totally different features.

Therefore, it is necessary to consider any correlation between transportation and payment records to clarify the risk of re-identification. Such histories are unlikely to be completely independent, leading to some correlation. Because the best mathematical property for modeling the correlation of two data sources is not well known yet, we propose a new method to evaluate the risk of re-identification from *multiple domain* combined history of transportation and purchase records. Our proposed method allows us to quantify the risk of re-identification from the transportation history given purchase history and to take steps for calibrating the level of de-identification the combined data efficiently.

In this paper, we empirically study the payment card "Suica"[13], a major payment card service used in Japan, which records not only the history of railway stations visited but also the history of item purchases, deposit, bus charges, and other uses. We can observe both the history of stations visited and item purchases for the same person through this data. We model the correlation between histories involving two different domains using information entropy, thereby quantifying the risk of re-identification from the data via the mutual information about the person. We measured this risk for synthesized data and evaluated the correlation between transportation and purchase histories. It should be noted that we found such a correlation for actual payment card data, and the risk is increased greatly above the risk when only one history is used.

In addition, we conducted an experiment to evaluate the risk using actual payment card data and validated our proposed method for data from 31 subjects. Finally, after proposing some metrics for the privacy and utility of anonymization, we have evaluated the risk for these data.

Our main contributions are as follows.

- We have proposed new metrics based on entropy for evaluating the re-identification risk in anonymized data that comprises histories from multiple domains.
- We have conducted an experiment to evaluate the risk for real payment card data collected from 31 subjects.
- We have developed a platform to evaluate the privacy and utility of anonymized data and have reported on the results of an evaluation.

TABLE I: Example of data from payment card

Date	Detail	Fare(JPY)
Oct 30 2016	in : Ueno (JR-EAST) out : Tokyo (JR-EAST)	-194
Oct 30 2016	in : Tokyo (JR-EAST) out : Ueno (JR-EAST)	-194
Oct 8 2016	Deposit in ticket vending machine	2000
Oct 1 2016	Purchase in vending machine	-150

TABLE II: Details of data from payment card

	Class	Number	Attribute	Detail
Personal Data	user data M	n 31	user ID	2 digit number
			sex	M/F
			grade	1 digit number
			address	place
			range of season ticket 1	place
			range of season ticket 2	place
	history data T	ℓ 584	user ID	2 digit number
			date	yyyy/mm/dd
			times	value
			entraining point	name of station
			alighting point	name of station
			entraining route	name of route
			alighting route	name of route
			usage	category
			location of use	category
			fare	value

TABLE III: Example of user data M

User ID	Sex	Grade	Address	Range of season ticket 1	Range of season ticket 2
1	M	1	Chiba	NA	NA
2	F	3	Tokyo	Nakano	Shinzyuku

This paper is organized as follows. In Section 2, we describe the data collected from 31 subjects. In Section 3, we propose a new entropy-based method for the evaluation of the re-identification risk for data from multiple domains. In Section 4, we measure the risk via experiment. In Section 5, we describe related works. Section 6 concludes the paper.

II. Payment card data

For our research, we collected personal data and payment card histories from 31 subjects, with their agreement. Let M and T be tables of registered passengers and records of boarding, respectively. We used the Android application "IC-card reader by MoneyForward"[14] to collect this information. We obtained 19 records of use per user. Table I shows an example of the history that we retrieved from the application.

Table II shows a statistical summary of M and T. M has 31 records for six attributes, and T has 584 records for 10 attributes. Tables III and IV show examples of M and T, respectively. From the payment card, there are three main attributes (date, details of usage, and fare), which may involve some subordinate attributes. We therefore divided the details attribute into six sub-attributes; namely, entraining point, alighting point, entraining route, alighting route, usage, and location. Table IV shows the subdivided data from Table I. The usage attribute has five possible values; namely, traffic, purchases, deposit, bus charges, and other uses. The location attribute has eight possible values; namely, simple re-chargers, ticket vending machine, vending machine, fare adjustment machine, fare adjustment machine for riding past or connections, commodity sales terminal unit, on board terminal, or nothing.

We first created M by reading the subject directory. We then added to M the entraining and alighting stations specified in the season ticket because this information is not stored in the payment card.

III. Evaluation of the risk of re-identification

A. Measuring re-identification risk using entropy

We propose a method for evaluating the risk of re-identification from personal data by using entropy. Table V shows a fragment E_S of such data, which gives the number of times that three stations (s_1, s_2, and s_3) have been passed through by three users, (u_1, u_2, and u_3). For example, u_1 passed through station s_1 twice, and s_2 did so once. First, let $P(U = u_i)$ be the probability of occurrence of u_i in a history E_S. Let n be the number of users. The entropy of users $H(U)$ when the history of passing stations is not available is given by

$$H(U) = -\sum_{i=1}^{n} P(U = u_i) \log_2 P(U = u_i).$$

In this case, we have $H(U) = 1.47$ [bit/history] because E_S gives $P(U = u_1) = 3/19$, $P(U = u_2) = 8/19$ and $P(U = u_3) = 8/19$.

Second, let $P(S = s_i)$ be the probability of occurrence of s_i in E_S. Let m be the number of stations. We then have the conditional entropy of users, given history s_i, as $H(U|S = s_i) = -\sum_{j=1}^{n} P(U = u_j|S = s_i) \log_2 P(U = u_j|S = s_i)$.

The entropy of users, given the history of use of stations S, $H(U|S)$ is given by

$$H(U|S) = \sum_{i=1}^{m} P(S = s_i) H(U|S = s_i).$$

Here, this evaluates to

$$\begin{aligned} H(U|S) &= \sum_{i=1}^{3} P(S = s_i) H(U|S = s_i) \\ &= \frac{10}{19} 1.52 + \frac{5}{19} 0.72 \\ &= 0.99. \end{aligned}$$

Finally, we calculate the mutual information. The mutual information $I(U; S)$ is the expected value of the amount of information obtainable from one record in the history of passing through stations. $I(U; S)$ is given by

$$I(U; S) = H(U) - H(U|S)$$

Here, this evaluates to

$$I(U; S) = 1.47 - 0.99 = 0.48.$$

We can interpret the semantics of $H(U)$, $H(U|S)$, and $I(U; S)$ as follows. For the situation when the history of passing stations is completely unknown, we have $H(U) = 1.47$,

TABLE IV: Example of history data T

User ID	Date	Times	Ent. point	Ali. point	Ent. route	Ali. route	Usage	Location	Fare
1	Oct 30 2016	2	Ueno	Tokyo	JR-EAST	JR-EAST	traffic	NA	-194
1	Oct 30 2016	1	Tokyo	Ueno	JR-EAST	JR-EAST	traffic	NA	-194
1	Oct 8 2016	1	NA	NA	NA	NA	deposit	ticket vending machine	2000
1	Oct 1 2016	1	NA	NA	NA	NA	purchase	vending machine	-150

TABLE V: Totalization table for example E_S

User \Station	s_1	s_2	s_3	Sum	$P(U=u_i)$
u_1	2	1	0	3	3/19
u_2	4	0	4	8	8/19
u_3	4	4	0	8	8/19
$H(U\|S=s_i)$	1.52	0.72	0		
$P(S=s_i)$	10/19	5/19	4/19		

TABLE VI: Values of entropy of usage

	Station(S)	Purchase(B)	Deposit(C)	Station and Purchase(S,B)
$H(U)$	4.900	4.338	4.736	4.412
$H(U\|x)$	1.814	0.948	3.256	0.182
$I(U;x)$	3.085	3.389	1.479	4.230
$P(U\|x)$	0.284	0.518	0.105	0.881
n_x	31	25	29	31
m_x	138	58	17	8004

which is equivalent to the average probability of identifying each user being $1/2^{H(U)} = 0.36$. Given one record of history of passing stations, we have $H(U|S) = 0.99$ and the average probability of identifying each user as $1/2^{H(U|S)} = 0.5$. In this case, we gain 0.48 bit of information from one history of passing stations. Noting that

$$H(U) = 1.47 < 1.92 = 4I(U;S),$$

we obtain the average probability of identifying each user, given four records of histories of passing stations, as about 1.

B. Entropy of payment card data

Consider a history data set comprising 31 users U, 138 stations S, 58 fares to purchase B, and 17 deposits to make C. To simplify the computations, assume that the number of fares to purchase equals that for goods values.

Table VI shows the entropies classified by usage. x represents a specific usage. If x is "station", then $H(U) = 4.900$ and $I(U;S) = 3.085$, giving the average probability of identifying each user, given no history of use of the payment card, as $1/2^{H(U)} = 0.033$. One record of usage by the payment card conveys only one usage history. Therefore, given one history of passing stations, the average probability of identifying each user rises to $1/2^{H(U)-H(U|S)} = 0.284$. Let N_x be the number of users who use a payment card for usage x. For example, for the 31 users of a payment card, 31 users used the payment card for "traffic", and 25 users used it for "purchase". Let $P(U|x)$ be the average probability of identifying each user, given usage x.

TABLE VII: Example of fares to purchase for E_B

User \Fare	b_1	b_2	Sum	$P(U=u_i)$
u_1	2	0	2	2/7
u_2	1	3	4	4/7
u_3	0	1	1	1/7
$H(U\|B=b_i)$	0.92	0.31		
$P(B=b_i)$	3/7	4/7		

C. Correlation between usages

In this section, we analyze the relationships among usages to assess the re-identification risk with the history data of a payment card. Figure 1 shows a scatter plot between traffic frequency and amounts of deposit. Figure 2 shows a scatter plot between traffic fares and amounts of deposit. The correlation coefficients for amounts of deposit with frequency and traffic are 0.469 and 0.315, respectively, indicating that these two factors are slightly correlated. This means that the history data carry the risk that the traffic history could be predicted from the deposit history, and vice versa.

Table VII shows a fragment of E_B, giving totals for the purchases of three users. Users in E_B are the same users as in E_S. We obtain $H(U) = 0.98$, $H(U|B) = 0.57$, and $I(U;B) = 0.41$ for E_B, in a similar way to that described in Section 3.1.

Table VIII shows a fragment of $E_{S,B}$, giving totals for the combination of "traffic" and "purchases" of three users. In this case, we have

$$P(u_1|s_1,b_1) = \frac{P(u_1|s_1)P(u_1|b_1)}{\sum_{i=1}^{n} P(u_i|s_1)P(u_i|b_1)} = \frac{4}{4+4} = \frac{1}{2}$$

and $H(U) = 1.19$, $H(U|S,B) = 0.46$, $I(U;S,B) = 0.73$. Table IX shows the entropies of E_S, E_B, $E_{S,B}$. We obtain

$$I(U;S) + I(U;B) = 0.89 > 0.73 = I(U;S,B)$$

from Table IX. This shows that "traffic" is not independent of "purchases".

Table VI shows the entropies for the combination of "traffic" and "purchases". Comparing them with the values for $I(U;S)$ and $I(U;B)$, we have

$$I(U;S) + I(U;B) = 6.474 > 4.230 = I(U;S,B).$$

For example, $m = 8004$ is the number of combinations of "traffic" and "purchase".

IV. Evaluation

In this section, we consider the re-identification risk for data that contain a history of traffic and a history of purchases stored in payment cards. We propose some metrics to evaluate

TABLE VIII: Totalization table when obtaining history from E_S and E_B one at a time

	s_1, b_1	s_1, b_2	s_2, b_1	s_2, b_2	s_3, b_1	s_3, b_2	Sum	$P(U = u_i)$
u_1	4	0	2	0	0	0	6	6/46
u_2	4	12	0	0	4	12	32	32/46
u_3	0	4	0	4	0	0	8	8/46
$H(U\|S = s_i, B = b_j)$	1	0.81	0	0	0	0		
$P(S = s_i, B = b_j)$	8/46	16/46	2/48	4/46	4/46	12/46		

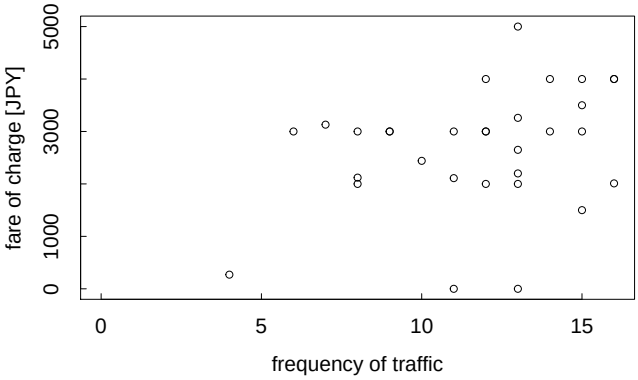

Fig. 1: Scatter diagram for times of traffic and amounts of deposit

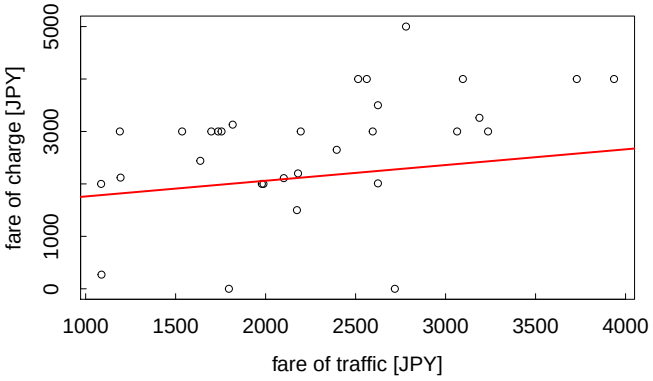

Fig. 2: Scatter diagram for traffic fares and amounts of deposit

the re-identification risk. A metrics of privacy is an index that identifies a person from anonymized data. The degree of privacy is evaluated by a re-identification ratio, which is defined as the fraction of correctly identified records out of the whole number of records in the original data. Utility metrics are indexes of differences between the values of attributes in the original data and the anonymized data. The index of utility is the mean absolute error between the original data and the anonymized data. We use *Python* and *R* to implement these indexes. We report on the results of our analysis of payment

TABLE IX: Values of E_S,E_B,$E_{S,B}$

$\backslash x$	s	b	s, b
$H(U)$	1.47	0.98	1.19
$H(U\|x)$	0.99	0.57	0.46
$I(U;x)$	0.48	0.41	0.73
$P(U\|x)$	0.50	0.67	0.73

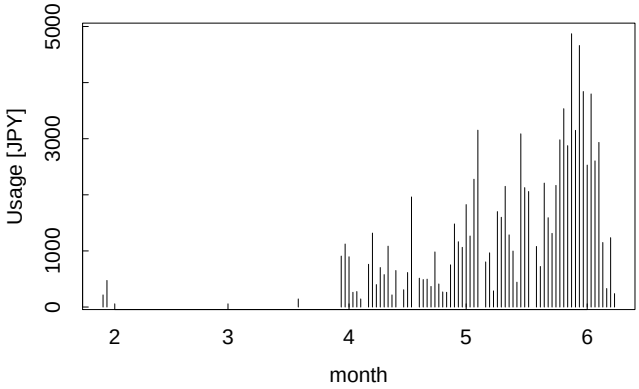

Fig. 3: Change of fare in days

card data in Section 4.1 and explain the proposed indexes of privacy and utility in Sections 4.2 and 4.3. Finally, we show the result of an experiment to evaluate the re-identification risk for payment card data in Section 4.4.

A. Analysis of payment card data

1) Statistics of payment card data: Figure 3 shows the monthly sum of fares used from April to June. The sum of fares in June is the largest because the storage size of the smart-card application that we used to get data from payment card is limited up to the latest 20 histories and too many records were collected in June when we got these data. Hence, this figure does not mean that subjects used payment card more often in June than any other months. Figure 4 shows the sum of fares for all students. The most frequently used subject spent 4,633 JPY (equivalent to 40 USD) in his latest 20 history and the least used subject spent 2,393 JPY. The average fare is 3133.9 JPY.

We speculated that the information from a payment card would contain not only the history of traffic but also the history of other usage. Table X shows the rate of usage in such a history. Figure 5 shows the rate of usage for each student. The history of traffic accounts for 62.3% of all histories, and we note that some students use a payment card to purchase items more often than for traffic use. The usage by students varies considerably.

Figure 6 shows the distribution of the Jaccard distance that represents the similarity among subjects in terms of passing stations. The mean Jaccard distance for the payment card data is 0.933. Therefore, the subjects using this payment card are totally different in term of passing stations.

2) Frequency of appearance of values: The top three usages for the payment card are traffic, purchases, and deposits,

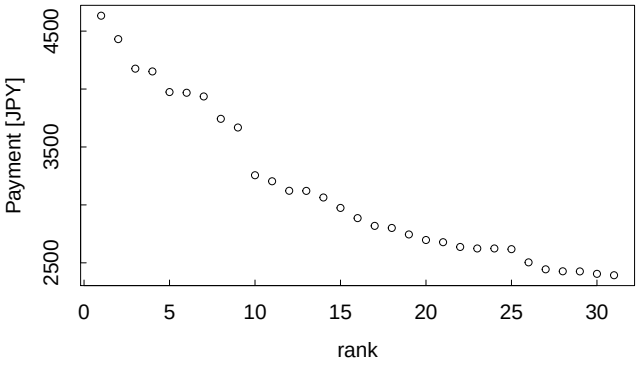

Fig. 4: Details of fare for all users

TABLE X: Breakdown of usage of payment card data

Usage	No. of records	Rate
Traffic	364	62.3%
Purchase	100	17.1%
Deposit	84	14.4%
Bus charge	2	0.3%
Others	34	5.8%
Sum	584	100.0%

accounting for about 94% of all histories. In this section, we investigate the frequency of these usages.

Figures 7, 8, and 9 show the frequency of appearance in the history for the top three usages, traffic, purchases, and deposits, respectively. Table XI shows the rates for rare values; i.e., those whose frequency of appearance is less than two for each usage. For example, the history data of traffic have 364 records, 138 stations, 727 histories, and 78 rare values. These rare values could be used for identification of an individual. Therefore, we need to deal with these values.

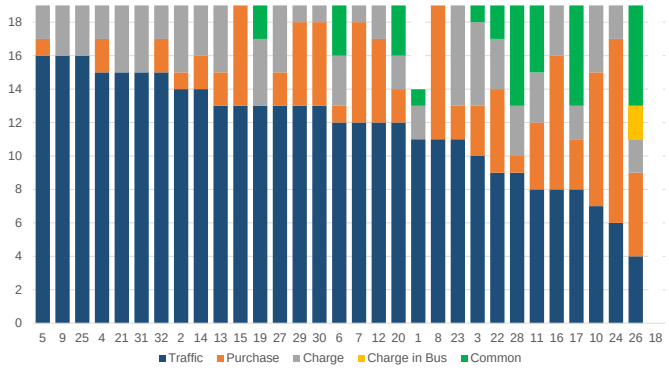

Fig. 5: Breakdown of usage of users

TABLE XI: Rate for rare data

Times/Usage	Traffic	Purchase	Deposit
1	4.8%	43.0%	11.9%
2	16.6%	55.0%	19.0%

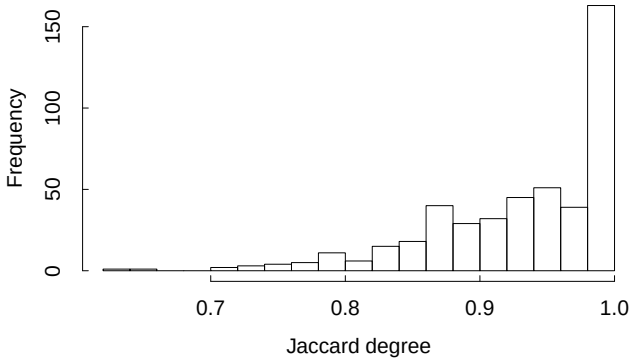

Fig. 6: Distribution of Jaccard distance between users for stations

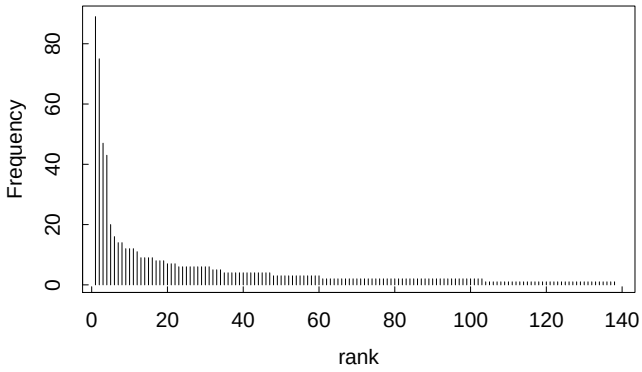

Fig. 7: Frequency of appearance of stations

B. *Anonymization methods*

We anonymized the payment card data with many anonymization methods. In this section, we introduce some representative methods we used.

First anonymization method is a method that adds random noise to values of SAs (sensitive attributes) of the original data. For example, in the case of example of history data T (Table IV), we add random noise to attributes like date, fare, and location.

Second method is a method that replaces values of SAs with the average. For example, in the case of example of T, we replace four values of attribute of fare, $-194, -194, 2000, -150$,

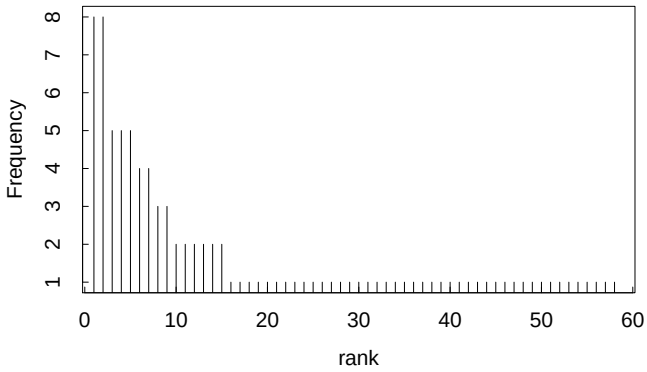

Fig. 8: Frequency of appearance of fares for goods

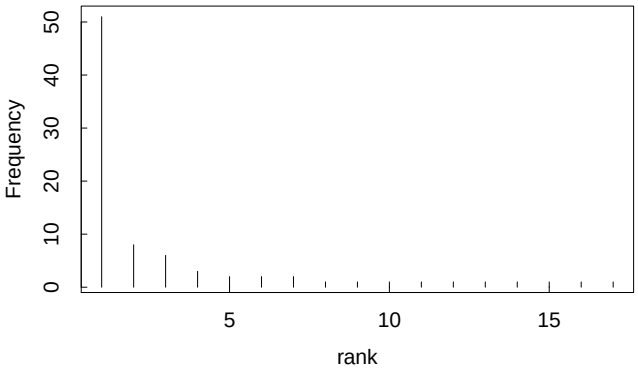

Fig. 9: Frequency of appearance of amount of deposit

with the average 365.5.

C. Privacy metrics

We propose six privacy metrics for the re-identification threat. Metric $S2$ gives the accuracy of identifying individuals from given anonymized data based on the sums of fares for customers. Metric $S3$ gives the accuracy of identifying individuals from the anonymized data based on the number of records for each usage.

We note that many metrics are defined for the same attributes of payment card data. For example, 5 of the 12 metrics for privacy are defined for fare attributes. We describe algorithms for $S2$ and $S3$ as follows. The privacy of anonymized data is evaluated based on the re-identification ratio, defined as the fraction of users identified correctly from the number of users in the anonymized data.

Algorithm $S2$

1) Input: original data T, anonymized data X, number of users of T n, and number of users of X m.
2) Compute the sum of fares for each user t_i in T and denote the set of sums by $f(t_1), ..., f(t_n)$.
3) Compute the sum of fares for user x_1 in X and denote this value by $f(x_1)$.
4) Find the nearest value, say t_i, to the $f(x_1)$ from $f(t_1), ..., f(t_n)$, and identify t_i as the anonymized user x_1. If there is a tie, choose a random t_i from among the tied users.
5) Repeat Steps 3 and 4 for $x_2, ..., x_m$.

Algorithm $S3$

1) Input: original data T, anonymized data X, number of users of T n, and number of users of X m.
2) Compute the number of records for each usage for each user t_i in T and denote these values by vectors $u(t_1), ..., u(t_n)$. In this case, $u(t_i)$ is a five-dimensional vector, whose elements are the number of uses for traffic, purchases, deposit, bus charges, and other usages.
3) Compute the numbers of records for each usage for user x_1 of X and denote this value by vector $u(x_1)$.
4) Find the nearest vector, say t_i, to the $u(x_1)$ from $u(t_1), ..., u(t_n)$ in terms of Euclidean distance, and identify t_i as the anonymized user x_1. If there is a tie, choose a random t_i from among the tied users.
5) Repeat Steps 3 and 4 for $x_2, ..., x_m$.

D. Utility metrics

We propose 12 metrics for the utility of the anonymized data. Metric $U1$ evaluates the distance between the original data and the anonymized data in terms of the mean absolute error in sums of fares for users. Metric $U10$ evaluates the distance between the original data and the anonymized data in terms of the mean absolute error in sums of fares for usages and users. The table shows the relationship between these metrics and the attributes stored in the payment card. Note that several metrics (5 of the 12) are defined in terms of the fare attribute. We describe algorithms for $U1$ and $U10$ as follows.

Algorithm $U1$

1) Input: original data T, anonymized data X, number of users of T n, number of users of X m, and a mapping p from T onto X
2) Let $f(t)$ be a vector of sums of fares for users. It is an n-dimensional vector $(f(t_1), f(t_2), ..., f(t_n))$, where $t_1, ..., t_n$ are the users in T and $f(t_i)$ is the sum of fares for the i-th user, based on p.
3) Let $f(x)$ be a vector of the sum of fares for user x. $f(x)$ is an m-dimensional vector $(f(p(t_1)), f(p(t_2)), ..., f(p(t_n)))$.
4) Compute the mean absolute error between $f(t)$ and $f(x)$ as an evaluated value. If $n \neq m$, add elements with null values to the smaller vector to give a vector with matching dimensionality. For example, if $n > m$, we add $n - m$ "0" elements to $f(x)$ and compute the mean absolute error between $f(t)$ and $f(x)$.

Algorithm $U10$

1) Input: original data T, anonymized data X
2) Compute the number of records for each usage in T as vector U_T. In this case, U_T is a five-dimensional vector (sum of numbers of records for all users in T for traffic, purchases, deposit, bus charges, and other usages.
3) Compute the number of records for usage of X as a vector U_X.
4) Compute the mean absolute error between U_T and U_X as an evaluated value.

E. Evaluation

1) Experiment overview: In this subsection, we examine the privacy and utility of anonymizing payment card data. We developed a web-based platform on *Linux* to evaluate anonymized data automatically. We used the system and metrics described in Subsections 4.1 and 4.2 to evaluate the re-identification risk in payment card data. Figure 10 outlines the system configuration of our evaluation platform.

The platform evaluates anonymized data in terms of various metrics and outputs the results and the rankings for submitted data sets based on selected metrics for evaluation. The results of evaluations are accumulated in an *SQL* database and will be used for further analysis. Users of the platform can upload

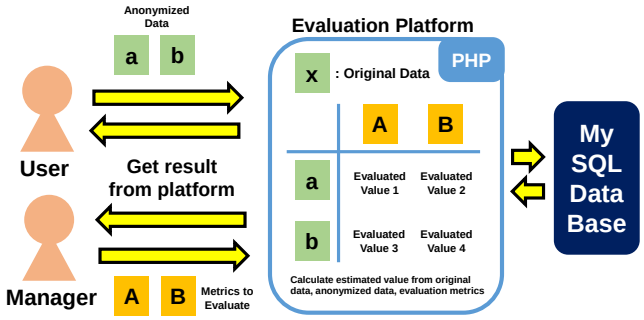

Fig. 10: System configuration

arbitrary metrics or data as a drag-and-drop operation. Figures 11 and 12 show an upload and the results of evaluation using the platform. We plan to extend this platform to evaluate not only payment card data but also other data.

2) Experimental results: Our experiments were conducted in August, 2016. Data comprising 47 anonymized payment card records that were processed with anonymization methods that mentioned in section of Anonymization methods were submitted to the evaluation platform by students of Meiji University. Figure 13 shows the distribution of the submitted anonymized data. The X-axis and Y-axis indicate the utility rank and the privacy rank, respectively.

The overall rank is given in terms of the mean rank (utility ranking + privacy ranking)/2, and the straight line in Figure 13 shows the same evaluation boundary as the original data. For our platform, the data below this boundary are evaluated as better than the original data, and the data above this line are less useful. In this case, the top right ranked data (1st–8th) are evaluated as better than the original data, the 9th ranked data are evaluated as very similar to the original data, and the remaining data (34 of the 47 items) are evaluated as less useful.

Figure 14 shows the evaluation of metrics for the privacy of anonymized data. Figure 15 shows the maximum values of these metrics. Most of the anonymized data (44 of the 47 items) were identified using more than 45% of the metrics. Therefore, we can claim that the re-identification risk in data that contain histories of traffic and purchases is high.

Figure 16 shows the relationship between results for $S2$ and $U1$, which involve processing fare attributes. Data 1 starts as the original data, having an evaluated value for $U1$ of 0 (no difference between the original data and the anonymized data) and an evaluated value for $S2$ of 1 (all users were completely identified). As more fare attributes are added, the evaluated value for $U1$ increases (Data 1 loses its utility), and the evaluated value for $S2$ decreases (Data 1 becomes more secure).

V. Related works

There are two representative methods to evaluate the privacy level of data, k-anonymity [1] and differential privacy [2]. k-anonymity, was defined by Sweeny in 2006, evaluates privacy level of data whether the data has at least k indistinguishable

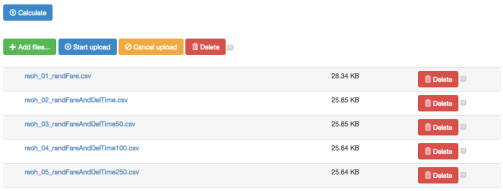

Fig. 11: Screen of evaluation platform (upload)

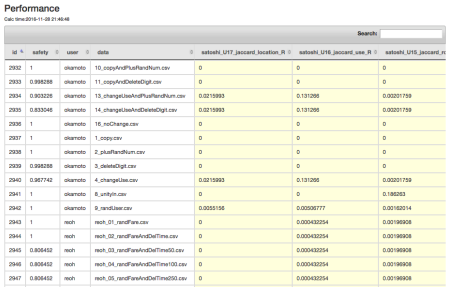

Fig. 12: Screen of evaluation platform (Result of evaluation)

records in terms of quasi-identifiers. Differential privacy, was defined by Dwork in 2006, evaluates privacy level of data whether the possibility to restore personal data from difference of analysis result of the data is high.

SO Technical Specification ISO/TS 25237 [4] defines *anonymization* as "a process that removes the association between the identifying data and the data subject." The ISO definition classifies anonymization techniques into masking and de-identification, and has been considered favorably [5]. Many anonymization algorithms have been proposed to preserve privacy, while aiming to retain the utility of the data that have been *anonymized*. That is, the data are made less specific so that a particular individual cannot be identified. Anonymization algorithms employ various operations, including *suppression* of attributes or records, *generalization* of values, replacing values with *pseudonyms*, *perturbation* with random noise, sampling, rounding, swapping, top/bottom coding, and micro-aggregation [3], [6].

Domingo [8] proposes a model for a maximum-knowledge attacker who knows both the original dataset and the anonymized dataset. The attacker can use all of the attributes to estimate the best possible linkages. Koot et al. proposed

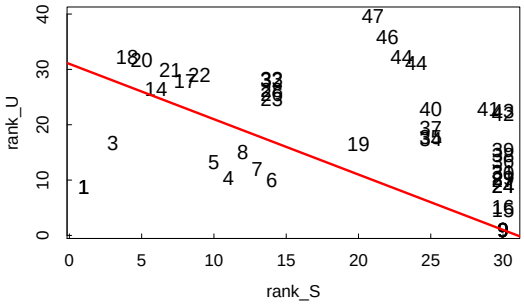

Fig. 13: Results of experiments

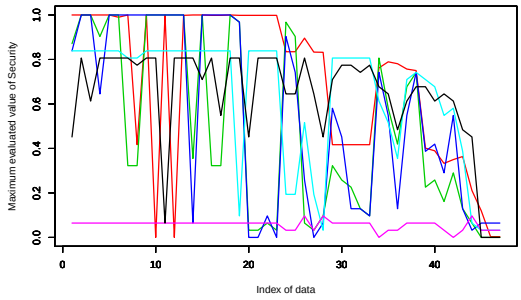

Fig. 14: The evaluations of privacy metrics

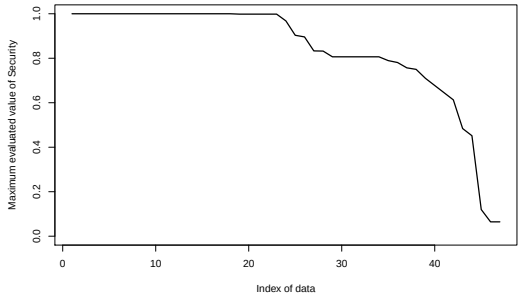

Fig. 15: The maximum evaluated value for privacy metrics

a method to quantify anonymity via an approximation of the uniqueness probability using a measure of the Kullback-Leibler distance in [9]

Monreale et al. proposed a framework for anonymization of semantic trajectories data, called c-safety in [10]. Based on the framework, Basu et al. presented an empirical risk model for privacy based on k-anonymous data release in [11]. Their experiment using car trajectory data gathered in Italian cities of Pisa and Florence allows the empirical evaluation of the protection of anonymization of real-world data.

In 2017, Torra gave a general introduction on data privacy studies in [16]. Zhizhou and Lai propose a new definition of δ-privacy model that requires that no adversary could improve more than δ privacy degree [17].

VI. Conclusions

In this paper, we have proposed metrics involving entropy to evaluate the re-identification risk of data that contain histories from multiple domains. We conducted experiments using real payment card data collected from 31 subjects. From the results, we found that that there is a correlation between histories of traffic and purchases acquired via a payment card. This means that the re-identification risk increases greatly when more than one type of history is available. We have proposed six privacy metrics and 12 utility metrics for payment card data and demonstrated evaluations of the re-identification risk given by these metrics. The results show that most (44 of 47 items) of the anonymized data were identified correctly with more than 45% accuracy for any of the privacy metrics, and we note that the re-identification risk for payment card data is very high.

Discussion of the metrics for evaluating privacy and utility of data is presented in *PWSCUP* [15]. In future work, we plan to collect and analyze larger-scale data, proposals for more-precise indexes to evaluate data, and more-practical anonymization methods.

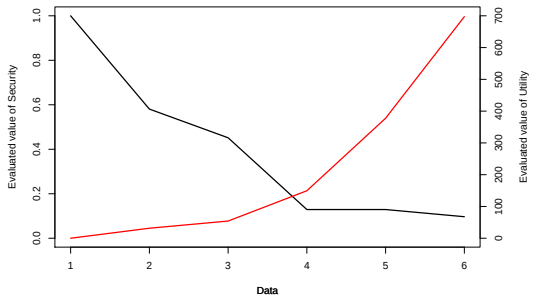

Fig. 16: The relationship between values for S2 and U1

References

[1] L. Sweeny, "k-anonymity: a model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570, 2006.
[2] C. Dwork, "Differential privacy", Proceedings of ICALP 2006, LNCS vol.4052, pp.1-12, 2006.
[3] Information Commissioner's Office (ICO), Anonymisation: managing data protection risk code of practice, 2012.
[4] "Health informatics – Pseudonymization", ISO Technical Specification ISO/TS 25237.
[5] Khaled El Emam, Luk Arbuckle, "Anonymizing Health Data Case Studies and Methods to Get You Started", *O'Reilly*, 2013.
[6] C.C. Aggarwal and P.S. Yu., "A General Survey of Privacy-Preserving Data Mining, Models and Algorithms", *Privacy-preserving data mining*, Springer, pp. 11-52, 2008.
[7] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111-133, 2001.
[8] Josep Domingo-Ferrer, Sara Ricci and Jordi Soria-Comas, "Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker", 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), *IEEE*, 2015.
[9] Koot, M. R., Mandjes, M., van't Noordende, G., and de Laat, C., "Efficient probabilistic estimation of quasi-identifier uniqueness", In Proceedings of ICT OPEN 2011, 14-15, pp. 119-126, 2011.
[10] A Monreale, R Trasarti, D Pedreschi, C Renso and V Bogorny, "C-safety: a framework for the anonymization of semantic trajectories", Transactions on Data Privacy, Vol. 4 (2), pp. 73-101, 2011.
[11] A. Basu, A. Monreale, R. Trasarti, J. C. Corena, F. Giannotti, D. Pedreschi, S. Kiyomoto, Y. Miyake and T. Yanagihara, "A risk model for privacy in trajectory data", Journal of Trust Management, 2:9, 2015.
[12] Daqing Chen, Sai Liang Sain, and Kun Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197–208, 2012.
[13] EAST JAPAN RAILWAY COMPANY, http://www.jreast.co.jp/e/, June 24, 2017.
[14] Money Forward, http://corp.moneyforward.com/, June 24, 2017.
[15] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, "What is the Best Anonymization Method? - a Study from the Data Anonymization Competition Pwscup 2015", Data Privacy Management Security Assurance (DPM2016), LNCS 9963, pp. 230 - 237, 2016.
[16] V. Torra, "Data Privacy: Foundations, New Developments and the Big Data Challenge", Studies in Big Data 28, Springer, 2017.
[17] Zhizhou Li, Ten H. Lai, δ-privacy: Bounding Privacy Leaks in Privacy Preserving Data Mining, DPM/CBT 2017, LNCS 10436, pp. 124142, Springer, 2017.