Electronic Preprint for Journal of Information Processing Vol.24 No.5

Regular Paper

Zipf Distribution Model for Quantifying Risk of Re-identification from Trajectory Data

Hiroaki Kikuchi[1,a)] Katsumi Takahashi[2]

Received: December 3, 2015, Accepted: June 2, 2016

Abstract: In this paper, we propose a new mathematical model for evaluating a given anonymized dataset that risks being re-identified. Many anonymization algorithms have been proposed in the area called privacy-preserving data publishing (PPDP), but, no anonymization algorithms are suitable for all scenarios because many factors, e.g., a requirement of accuracy, a domain of attributes, a size of dataset, and sensitivities of attributes, are involved. In order to address the issues of anonymization, we propose a new mathematical model based on the Zipf distribution. Our model is simple, but it fits well with the real distribution of trajectory data. We demonstrate the primary property of our model and we extend it to a more complex environment. Using our model, we define the theoretical bound for reidentification, which yields the appropriate optimal level for anonymization.

Keywords: anonymity, k-anonymity, re-identified risk, Zipf distribution

1. Introduction

At present, the volume of digital data is growing exponentially every year. Many business organizations try to collect our personal data so that they can share this data with partners and use data-mining algorithms to extract useful knowledge related to the behavior of customers and their preference for goods. However, many people are concerned about the leakage of personal data without their consent and violations of their privacy due to the publication of personal data.

Many anonymization algorithms have been proposed to preserve privacy in the area called PPDP. PPDP aims to retain the utility of data that have been *anonymized*, i.e., by making data less specific so that a particular individual cannot be identified. Anonymization algorithms employ various operations, including the *suppression* of attributes or records, *generalization* of values, replacing values with *pseudonyms*, *perturbation* with random noise, sampling, rounding, swapping, top/bottom coding, and microaggregation [1], [2].

It is not simple to anonymize data fully without the risk of reidentification. In particular, anonymization is affected by the following concerns.

(1) Motivated intruder.
It is difficult to predict the actions of an adversary. In [1], a *motivated intruder* is defined as an entity who may take a record from anonymized datasets and search for a match in publicly available information. In addition, it is not clear what information might be available to a motivated intruder.

(2) Lack of common metrics for quantifying privacy and utility.
It is well known that a publisher should be responsible for the risk of de-identification from their data. For example, as stated in Ref. [3], *"First, the company must take reasonable measures to ensure that the data is de-identified."* However, the risk is uncertain. The requirements and the utility vary according to the hypothesis employed by different algorithms. Thus, a multi-stakeholder process may help to obtain a consensus regarding measurement, but this requires a long time.
Although several measures for evaluating the risk of re-identification have been extensively proposed, e.g., Skinner and Elliot [4], Benitez et al. [5], there is no common measure that has been agreed for all.

(3) Lack of real datasets.
It is difficult to select the most appropriate algorithm given the data and an application because the results will vary with different datasets and parameters. Moreover, the original dataset is often not available in practice. Thus, instead of the original data, experimental open data are used to measure risk. For example, Ayala-Rivera et al. used the *Adult* census dataset from the UCI Machine Learning Repository [6] and the synthetically generated *Irish* dataset in Ref. [7]. Although the `sdcMicro` package [8] has been used by several researcher, it refers to a standard database and not to a location privacy.

Our Approach. Mathematical Model of Anonymized Data

In order to address the issues of anonymization, we propose a new mathematical model based on the Zipf distribution. Our model is simple but it fits well with the real distribution of trajectory data. We demonstrate the primary property of our model and

[1] Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
[2] NTT Secure Platform Laboratories, Musashino, Tokyo 180–8585, Japan
[a)] kikn@meiji.ac.jp

The primary version of this paper was published in the 13th Annual Conference on Security and Trust (PST), pp.14–21 (July 2015).

we extend it to a more complex environment. Using our model, we define the theoretical bound for reidentification, which yields the appropriate optimal level of anonymization.

Our Contributions.

Our first contribution is a proposal of a general mathematical model of data. Our model is based on the power law probability distribution, known as Zipf's law. Using the least-squared method, we can efficiently fit arbitrary quantities to our model with required accuracy. Especially, it fits well to general trajectory data, which is one of our targets to examine. Moreover, our model allows us to approximate a combination of multiple models because of its simplicity. Our mathematical model is simple enough to evaluate the risk to be re-identified without assuming any background knowledge of the intruder. Hence, the issue (1) of a motivated intruder is not necessary to be considered in our model.

Our second contribution is that we clarified the fundamental properties of our model. Based on the these properties, we prove the lower bound of the threshold for identifying an individual from anonymized records. The threshold plays an important role in the anonymization because it determines the fraction of records we need to suppress to satisfy k-anonymity. It is known that there is a tradeoff between the privacy degree and the utility of the data. Hence, we want to minimize the number of records being altered. Our model allows to clarify the least number of records being linked so that the preferable degree of anonymity is preserved without performing any experiment. The proposed quantities, such as the least number of records (will be defined in a later section), give the solutions to the problem (2) the common metrics for quantifying privacy and utility. The mathematical model may be acceptable for stakeholders because there is no subjective information.

Our third contribution is to demonstrate our proposed model based on the statistics provided from the Japan Railway Co. and clarified the risk of the anonymized data to be linked uniquely. We examine the least and the mean ranks in the set of stations that are unable to be identified uniquely. Our analysis reveals that most of the records can be identified with a very low degree of anonymity. Our demonstration shows that our methodology can be applied to an arbitrary use case without access to any confidential dataset. Therefore, the problem of a real dataset (3) is addressed in our scheme.

The problem that a few trips can identify a user is closely related to the problem of privacy in search log. Given multiple queries in search log, we could identify an individual with a statistic property of the database in the similar way. A similar problem may occur in many applications, e.g., the history of purchase in online shops, the list of books in a library, and so on. Hence, our model can be extended to a more general framework.

The rest of our paper is organized as follows. In Section 2, we define three threads in anonymized data, reidentification, distinguished, and identified, as the example of trajectory data. In Section 3, we propose a mathematical model of an anonymized dataset and study some fundamental properties. We also discuss the utility of anonymity as called the anonymity ratio. In Section 4.1, we examined the risk in the Japan Railway trajectory data. Finally, we conclude our study in Section 5.

2. Preliminary

2.1 Anonymization and Risk of Identification

Many models have been proposed to formally guarantee the privacy of big data, such as k-anonymity [9], [10], ℓ-diversity [11], t-closeness [12], and differential privacy [13].

To illustrate anonymization operations, let us consider the trajectory data example shown in **Fig. 1**, which is assumed to be owned by a railway company, where the attributes comprise the names of passengers, the day of use, the stations at which passengers board and alight, and the balance on prepaid RFID cards after being charged. The simplest way to anonymize data is *suppression*, which drops attributes such as the name, day, and balance because they can be exploited to identify particular individuals. If a record contains a significant distinguishing value, e.g., a rare station labeled "Nakano," then the whole record can be suppressed. The original table is anonymized by replacing names with random numbers called *pseudonym*, Occasionally, the pseudonyms may be refreshed again on a monthly or weekly basis.

The anonymized data may appear to be secure against an attacker who might try to identify individuals from the data. However, it is well known that the set of various attributes called *quasi-identifier* (QIDs) can be exploited to link the records in a table.

Figure 2 illustrates the threads present in anonymized data, where the sequence of stations is stored as pseudonyms. In this case, some threads are classified as follows:

(1) *reidentification*; particular individual names can be obtained from anonymized data for various reasons, such as matching with an auxiliary dataset to help identify a passenger.

(2) *distinguished*; some records are linked by QIDs. For example, pseudonym 3 distinguishes the first and the third records from others, thereby allowing all of the stations to be traced back to the pseudonym assigned to a specific individual. It should be noted that the records linked with pseudonyms

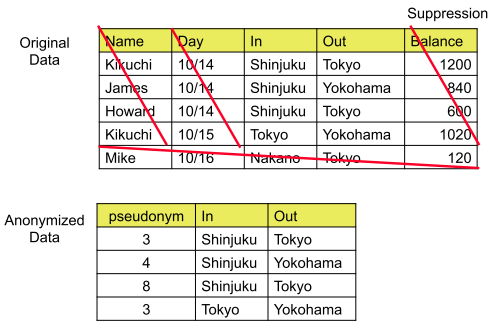

Fig. 1 Example of anonymized data.

Fig. 2 Threads related to reidentification.

3 and 4 have exactly the same values (stations), and thus the distinguished records don't always mean uniquely determined.

(3) *uniquely identified*; only one record is associated with the value and the individual can be identified uniquely. In the example, the record with pseudonym 6 can be uniquely identified based on the value "Nakano," which is associated with only one passenger. The stations "Tokyo" and "Shinjuku," are associated with at least two passengers so they cannot be identified uniquely.

Even if multiple records are linked by a pseudonym, some candidates will share the same values, such as the individuals assigned pseudonyms 3 and 4. Thus, we can say that they are not yet uniquely identified at that time. However, if their pseudonyms do not change over time, more stations will be linked with them and they can be uniquely identified. The likelihood of being uniquely identified increases with the duration of pseudonyms.

Even if an individual with pseudonyms either 3 or 4 is not uniquely identified, due to ℓ-diversity [11] there is a risk of an attribute value being inferred, i.e., an attacker, who knows the target went from Shinjuku to Tokyo, can infer that he went also from Tokyo to Yokohama. It is true that we should notice this threat, we leave the problem as one of the future study.

Algorithms such as k-anonymity ensure that every combination of published attributes and records is indistinguishable for no less than k instances. For example, by suppressing the unique records of pseudonym 6, the data satisfy 2-anonymity because two linked records have the QID, "Shinjuku"-"Tokyo"-"Yokohama."

Thus, we must address the following questions.

- How long can existing pseudonyms be used to maintain k-anonymity?
- How many records do we need to suppress to satisfy anonymity?
- How can we evaluate the degree of anonymity for an arbitrary dataset using different statistics?

2.2 Related Work

Some studies have been made for anonymization of trajectories. Monreale et al. proposed a framework for anonymization of semantic trajectories data, called c-safety in Ref. [21]. C-safety expresses the upper bound to the probability to infer that a given person has visited a sensitive place. Based on the framework, Basu et al. presented an empirical risk model for privacy based on k-anonymous data release in Ref. [20]. 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.

The gap between theoretical assumption and real-world data is also studied in Ref. [19]. Choi et al. compared the risk to be identified from a theoretical model and the real-world trajectories in Tokyo area in 2008. They claimed that the assumption of a uniform choice of next hop is too strong and proposed the revised theoretical model for trajectories data.

Garfinkel classifies various attacker models for computing the re-identification risk. According to Ref. [22], there are some scenarios including (1) a risk that a specific person in the dataset can be re-identified ("prosecutor scenario"), (2) a risk there exists at least one person in the dataset who can be re-identified ("journalist scenario"), (3) the percentage of identities in the dataset that can be correctly re-identified ("marketer scenario"), and (4) the distinguishability between an analysis performed on a dataset containing an individual and the same analysis performed on a dataset that does not contain the individual ("differential identifiability" scenario). In this classification, our study assumes (2) journalist scenario and (3) marketer scenario because we try to clarify the threshold for which no one can be re-identified in the sense of (2) and also compute the anonymity ratio that is a fraction of individuals to be re-identified, which corresponds to (3).

Classification of identifiers is one of the issues in anonymization. El Emam and Malin [23] have developed an 11-step process for performing the anonymization of a dataset on the classification of QIDs that includes step to determine direct identifiers evaluated by an expert. After masking direct identifiers, they suggest that the organization determines plausible adversaries and determines what minimal acceptable data utility. In our study, a sequence of stations is classified as a dynamic attribute and became available to an adversary. As Emam suggested, the classification of identifiers might be determined with the help of the organizer. We just note there are several attacker models and we should not determine just one of them.

3. Mathematical Model of an Anonymized Dataset

3.1 Fundamental Definition

We begin by defining a personal dataset characterized by parameters n and m.

Definition 3.1 Let n be a number of users. A *record* is a tuple of multiple attribute values for a user. A *dataset* is a set of m records for some users. Let D be a domain of (sensitive) attribute values and $d = |D|$ is the number of values in D.

A record belongs to a single user who performed an action at time t. A user may have multiple records in a dataset, so $m \geq n$ holds in general. Attributes are classified into two classes: *static* attributes, such as name, sex, marital status, and postal code; and *dynamic* attributes, such as location, money balance, blood pressure, heart rate, and name of disease. A set of previous attributes is known as *QID* if it links the records generated by a single user. Various properties have been studied to reduce the risk of reidentification based on an anonymized dataset, e.g., k-anonymity [9], [10] and ℓ-diversity [11]. Dynamic attributes are often referred as *sensitive attributes* (SAs) because they comprise critical information that the user may wish to hide. However, even if we suppress the QIDs from the dataset, the following theorem shows that collecting a very small number of records (SAs) can allow an individual to be identified.

Theorem 3.1 Given a dataset of n individuals where a SA is uniformly distributed with probability, $1/d$, all individuals can be uniquely identified from $s = \log_d n$ records.

Proof: The number of s-combinations of d-set is d^s, which equals n when all of the individuals are uniquely identified. Hence, we obtain the theorem by taking logarithms for both sides. □

For example, the population of Tokyo and its surroundings area in Japan comprises $n = 42{,}598{,}300$ individuals (Kanto area in 2012[*1]) and there are $d = 2{,}497$ stations[*2], and thus we find that $s = 2.25$ records are sufficient to uniquely identify all of the individuals in the Tokyo area. It should be noted that each individual has the same risk of being uniquely identified because we assume that there is a uniform distribution of station choices, and thus all cases with two records can be distinguished.

This number is surprisingly small. If we want to publish a trajectory dataset generated from smartcards logs as open data, the theorem implies that pseudonym IDs must be reassigned every three hops when traversing among stations. However, excessively frequent assignments of pseudonym IDs could degrade the correlations among the trajectories and the utility of the data would be lost.

Is the assumption of a uniform distribution of stations too strong?

The answer is no. In the following section, we show that the reidentification risk remains high even if the assumption of a uniform distribution is relaxed.

3.2 Mathematical Model of a Single-Station Record

To model the trajectory data, we use the following power law probability distribution, which is known as Zipf's law.

Definition 3.2 (**Zipf's Model**) Let $f(x)$ be the frequency of the item with the x-th rank. Then,

$$f(x) = \frac{a}{x^c}, \tag{1}$$

where a and c are constants.

The original Zipf's law states that the frequency of any word in a natural language is inversely proportional to its rank. This relationship applies to natural languages, but also in the physical and social sciences, such as the population ranks of cities, income rankings, and Web page rankings.

Fortunately, we found that the stations in the trajectory data were distributed according to an empirical power law and we present the data fitting results in a later section. According to statistics released by the railway company [14], the first station is Shinjuku with an average daily number of passengers $f(1) = 751{,}018$, followed by second station, Ikebukuro, with $f(2) = 550{,}350$, the third station, Tokyo, with $f(3) = 415{,}908$ etc. In the extended model, we still assume that passengers choose their destination independently, where they are distributed in $f(x)$, i.e., a destination is likely to be Shinjuku with a rate of $f(1)$ and Ikebukuro at a rate of $f(2)$. We ignore trivial cases where both the source and destination are the same station, but our hypothesis is sufficiently general to be applied to other examples. Before we describe the practical model, we consider the probability distribution of the Zipf model.

Suppose that the total number of passengers is denoted by N, which is obtained by evaluating the integral of $f(x)$ from 1 to the number of stations in a specific region, d, as follows;

$$N = \int_1^d a/x^c \, dx = \left[\frac{a}{1-c}x^{1-c}\right]_1^d = \frac{a}{1-c}(d^{1-c} - 1),$$

which allows us to define the probability function of our Zipf model.

Definition 3.3 In the dataset where the x-th item occurs $f(x)$ times, the probability of the x-th station being selected as the destination is $p(x) = f(x)/N$.

By fitting the open data [14] using the least-squares method, we obtain the constants $a = 8 \times 10^5$, and $c = 0.580$, and the probability $p(x) = 0.092/x^{0.632}$. The coefficient determination of the fitting is $R^2 = 0.9884$, which means the Zipf distribution fits with read datasets very well in high accuracy.

Table 1 Example Occurrence Probability ($m = 10$).

rank	value i	prob. p_i
1	Tokyo	4/10
2	Shinjuku	3/10
3	Yokohama	2/10 (= p^*)
4	Nakano	1/10

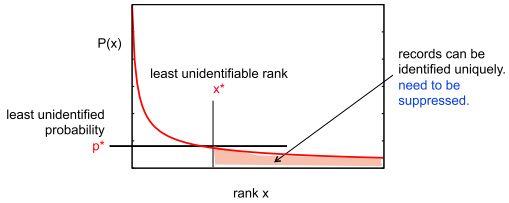

Fig. 3 Least unidentifiable rank and probability in a *long-tailed* distribution.

3.3 Risk of Records Being Linked

Some records are linkable with pseudonym IDs. If more attribute values are linked, it is more likely that the records will be unique. However, excessively frequent reassignment of pseudonym IDs could degrade the utility of the data. Thus, the length of linked records should be determined carefully based on a tradeoff between security and utility. To determine the least bound of the linked length, we define a threshold probability to ensure that an anonymized dataset is secure against reidentification.

Definition 3.4 A dataset comprises m records. Let p^* be the *least unidentifiable occurrence probability* defined as k/m. A record with a value that occurs with a probability less than p^* is uniquely identified in the dataset. The rank of the value is denoted by x^*, which is referred to as the *least unidentifiable rank*.

In this definition, the symbol k is the same one used in k-anonymity. Namely, in a dataset that satisfies k-anonymity, any record cannot be identified from its QID with at most probability $k/m(= p^*)$. Moreover, the least value of k is 2 because any dataset satisfies $k = 1$ without modifying. Hence, we define the default value of the least unidentified occurrence probability as $p^* = 2/m$.

For example, we consider a dataset with 10 records where the attribute values (station) occur with the probabilities given in **Table 1**. Station Nakano has a low frequency because only one passenger stops at the station. Thus, the records containing Nakano must be uniquely identified and they should be suppressed. The least unidentifiable occurrence probability p^* is $2/10$ and the rank of x^* is the third.

Figure 3 illustrates the least unidentifiable occurrence probability p^* in a distribution of frequencies in terms of the rank x.

[*1] http://en.wikipedia.org/wiki/Kanto_region
[*2] http://info.jmc.or.jp/ekiensen.html

The stations are sorted in order from the most frequent to the least based on the number of passengers boarding and alighting. The graph shows the fraction of passengers who board and alight at the station ranked at x, i.e., it shows the probability distribution $P(x)$ of x. The number of passengers boarding and alighting at stations is distributed with a "*long tail*," where a high-frequency population is followed by a low-frequency population that gradually tails off. Long-tail distributions are known to be common in many areas such as the occurrence of certain words in a natural language, the income distribution of a business, or the access counts of Web sites.

The most frequent stations are secure in terms of identification because many passengers have records for these stations, which prevents the identification of specific individuals. However, the records with rare stations located in the right-hand shaded area of the figure are likely to be identified and the names of the stations could be used as QIDs to trace their owners. Thus, these records need to be suppressed or generated using k-anonymity algorithms. The least unidentifiable rank x^* (and probability p^*) determines the threshold for stations that need to be suppressed. It should be noted that a dataset where no records occur with less than the least unidentifiable occurrence probability p^* satisfies 2-anonymity. In other words, the least unidentifiable occurrence probability p^* specifies a degree of anonymity such that $p^* = k/n$.

3.4 Zipf Model of Trajectory Data

We extend our single-station Zipf model to a model with multiple stations. First, we show that the probability function of our model satisfies a type of homomorphism in terms of a join.

Suppose that two records share a common pseudonym ID, but they have distinct stations ranked as x and y. The owner of a pseudonym boards a train at the x-th station and alights at the y-th station.

We assume that the sequence of stations can be regarded as a Markov chain, i.e., given the current station y, the probability distribution of next hop x depends only on the current station. The reason why we make assumption of Markov chains is that the choice of destination in a trip can be regarded as a random process and the statistical model is known to fit well to many real-world processes such as queuing theory in network, webpage rank, economics and finance.

In formal definition, a Markov chain of order N is a process satisfying

$$P(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2} = \cdots =, X_1 = x_1)$$
$$= P(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2} = \cdots =, X_N = x_N)$$

where N is a finite number of memory and X_i is a random variable taking value of station. For example, with order $N = 1$, a probability that a passenger at Shinjuku station chooses a destination Tokyo is given by $P(X_n = \text{Tokyo} | X_{n-1} = \text{Shinjuku})$. It does not depends on the previous station before time $n - 1$.

For the baseline analysis, we simply assume $N = 0$, that is, the conditional probability of selecting the x-th station given the current station $P(X_n = x | X_{n-1} = y)$ is equal to $P(X_n = x)$. Using the probability function $p()$ in the Zipf model, the probability of a path from the y-th station to the x-th station is given as the joint probability of $P(X_n = x)$ and $P(X_n = y)$, i.e.,

$$P(X_n = x, X_{n-1} = y) = P(X_n = x | X_{n-1} = y) P(X_{n-1} = y)$$
$$= p(y)p(x) = \frac{a}{x^c}\frac{a}{y^c} = \frac{a^2}{(xy)^c},$$

which belongs to a Zipf model of the form $p'(x') = a'/x'^c$ after replacing parameter a and x with $a' = a^2$ and $x' = xy$. Recursively, the join of multiple records follows the Zipf model. Therefore, this property allows us to identify the least unidentifiable occurrence probability in a trajectory containing several stations as follows.

Theorem 3.2 A trajectory that comprises s records $x_1, \ldots, x_s$, which are selected according to the Zipf model characterized by $f(x_i) = a/{x_i}^c$, has the least unidentifiable occurrence probability

$$x^* = x_1 \cdots x_s = (a^s n/2)^{1/c}. \quad (2)$$

Proof: From Definition 3.4, the least unidentifiable occurrence probability is more than $k/n = 2/n$. Hence, the joint probability of the trajectory of $x_1 \ldots, x_s$ needs to satisfy $a^s (x_1 a_2 \cdots a_x)^{-c} \geq 2/n$, which gives the theorem. □

Equation (2) implies that the corresponding least unidentifiable rank x^* increases exponentially relative to the length of the path, s. In order to ensure that the dataset remains unidentifiable, we need to reassign pseudonyms so the records cannot be linked within the limit. Alternatively, the records with minor stations that exceed the least rank can be dropped from the dataset.

An increase in the unidentifiable rank of the trajectory does not mean that a single rank increases as s increases. The rank of a trajectory is obtained by multiplying the ranks $x_1 x_2 \cdots x_s$ in the trajectory, so the average rank in the trajectory is the s-th root of the integrated rank x^*. For example, when $n = 42{,}598{,}300$ (population of Kanto area), the trajectory of $s = 3$ stations has a least unidentifiable rank of $x^* = 5{,}164 = 17^3$. Thus, the mean rank of x_1, x_2 and x_3 is the 17-th rank, Tamachi station, which is very common and more than 144,000 passengers stop there each day. By taking the s-th root of Eq. (2), we obtain the mean rank of the trajectory of s-stations as

$$\sqrt[s]{x^*} = a(n/2)^{1/cs}.$$

The mean unidentifiable rank of the trajectory decreases exponentially to s, as shown in **Fig. 4**. As s becomes longer, the mean

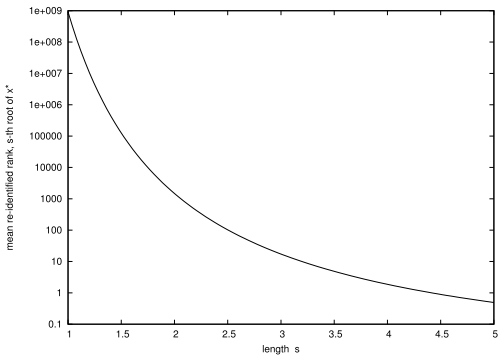

Fig. 4 Mean unidentifiable rank with regards to the length of linked records s.

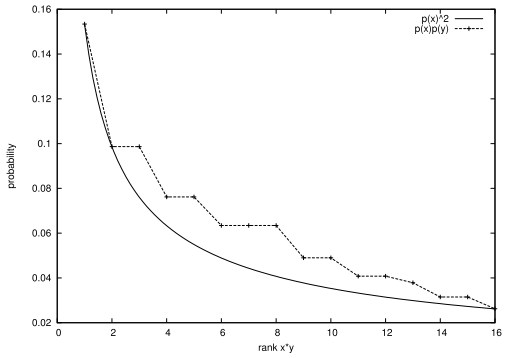

Fig. 5 Probability distribution for the trajectory $P(x, y)$ of $s = 2$ stations.

rank becomes shorter, and thus the number of records with minor stations that exceed the threshold increases, i.e., a longer s requires that most of the records are dropped from the dataset.

3.5 Quantifying the Utility of Anonymity

The main feature of data anonymization algorithms is that they usually modify the dataset by inserting fake records or suppressing critical records. However, it is well known that we lose more of the useful information if we suppress more records. Therefore, we should design an algorithm carefully based on a tradeoff between the security against risk of identification and the utility of the data. The utility function depends on the analysis applied and it is not easy to define a general formula. Hence, we use the following simple definition to quantify the loss of utility attributable to anonymization.

Definition 3.5 The *anonymity ratio* is the fraction of records suppressed to satisfy k-anonymity over all of the records in a dataset.

In the example shown in Table 1, the fourth record is suppressed out of m records and the anonymity ratio is $1/10$. In general, the anonymity ratio is obtained by evaluating the integral of the probability distribution function from the least unidentifiable rank x^* to the maximum rank d. Using the Zipf probability model for a single item, the anonymity ratio is

$$\int_{x^*}^{d} a/x^c \, dx = \frac{a}{1-c}(x^{*1-c} - d^{1-c}). \tag{3}$$

It is not trivial to extend the result obtained for a single item to the case of a trajectory with s records because there are many possible combinations of x and y such that $xy = z$ for a given threshold. For instance, there are two pairs $3 \times 2 = 2 \times 3 = 12$ for the trajectory of $z = 12$-th and three pairs for $z = 4$ because $z = 4 = 1 \times 4 = 4 \times 1 = 2 \times 2$. Hence, the probability distribution function for the trajectory does not have a closed form even if each element occurs in the Zipf model. **Figure 5** shows the probability distribution for the trajectory of $s = 2$ stations and $d = 4$, where the squared Zipf probability $p(x)$ is plotted for comparison. Note that these functions are shared at both ends because there is only one combination at $z = 1$ and $z = d^2$. The probability function of a trajectory is not continuous.

Instead of the closed form probability distribution of the trajectory, we prove the lower bound of the anonymity ratio using our Zipf model.

Theorem 3.3 The lower bound of the anonymity ratio for the trajectory of s linked records that occur in the Zipf model is

$$\frac{a^s}{1-c} d^{s(1-c)} - \frac{a^s}{1-c} x^*,$$

where x^* is the least unidentifiable rank and d is the size of the domain of SA.

Proof: (i) For any x, y, z such that $xy = z \le d$, $p(x)p(y) = a^2/(xy)^c \le p_2(z) = a^2/z^c$ holds.
(ii) Suppose that the inequality holds for s as $p(x_1) \cdots p(x_s) \le p_s(z) = a^s/z^c$. Then, for any x_{s+1}, the joint probability $p(x_1) \cdots p(x_s)p(x_{s+1}) \le p_{s+1}(z) = a^{s+1}/z^c$ also holds. From (i) and (ii), the inequality holds for any s. Hence, the integral of the probability of p_s from the least unidentifiable rank x^* to the maximum rank d^s

$$\int_{x^*}^{d^s} a^s/x^c \, dx = \frac{a^s}{1-c}(x^{*1-c} - d^{1-c})$$

gives the lower bound of the anonymity ratio for the dataset of at most s linked records. □

The exact solution can be obtain within a small s. In addition, we present the closed form of the anonymity ratio for particular $s = 2$ as follows.

$$\begin{aligned}
&\int_{x^*d}^{d}\int_{x^*/x}^{d} p(x)p(y)\,dx\,dy \\
&= \int_{x^*d}^{d}\int_{x^*x}^{d} \frac{a^2}{x^c y^c}\,dx\,dy \\
&= \int_{x^*d}^{d} \frac{a^2 d^{1-c}}{(1-c)x^c} - \frac{a^2 x^{*1-c}}{(1-c)x^c x^{1-c}}\,dx \\
&= \left[\frac{a^2 d^{1-c} x^{1-c}}{(1-c)^2} - \frac{a^2 x^{*1-c}}{(1-c)} \log x\right]_{x^*d}^{d} \\
&= \frac{a^2}{(1-c)^2}(d^{2(1-c)} - x^{*1-c}) - \frac{a^2 x^{*1-c}}{(1-c)} \log \frac{x^*}{d^2}
\end{aligned}$$

3.6 Extension of the Zipf Model to Data with Multiple Attributes

In Section 3.3, we studied the unidentifiable rank of a single attribute value, e.g., stations; however, we claim that our proposed scheme can also model data with multiple attribute, such as item purchases, amounts of payments, numbers of items, or the time available to use.

We consider attribute 1 and 2 with domains of size d_1 and d_2, and the probabilities of their values are approximated by the following Zipf models, $p_1(x_1) = a_1/x_1^{c_1}$, $p_2(x_2) = a_2/x_2^{c_2}$, respectively. Assuming that values occur independently, the joint probability of a record having x_1 and x_2 is given as a new Zipf model. Unfortunately, the closed formula for the exact joint probability is not trivial, but we give the lower bound as follows.

Theorem 3.4 The attributes have probabilities of $p_1(x_1) = a_1/x_1^{c_1}$ for $x_1 \in D_1$ and $p_2(x_2) = a_2/x_2^{c_2}$ for $x_2 \in D_2$, respectively. A record has both x_1 and x_2 with a probability $P(x_1, x_2)$ such that

$$P(x_1, x_2) = p_1(x_1)p_2(x_2) = a_{12}/x_{12}^{c_{12}},$$

where $a_{12} = a_1 a_2$, $d_1 = |D_1|$, $d_2 = |D_2|$, and

$$c_{12} = \frac{c_1 \log d_1 + c_2 \log d_2}{\log d_1 + \log d_2}.$$

Proof: Given the boundary conditions, i.e., $p_{12}(1) =$

Table 2 Least Unidentifiable Rank and Anonymity Ratio based on JR East Open Data [14].

length s	least unidentifiable rank x^*	mean rank $\sqrt[s]{x^*}$	domain size d^s	anonymity ratio
1	915,199,427	915,199,427	2,497	0
2	2,173,909	1,474	6,235,009	0.06556
3	5,164	17	1.557×10^{10}	0.9796
4	12	2	3.888×10^{13}	1.0
5	0	0	9.707×10^{16}	1.0

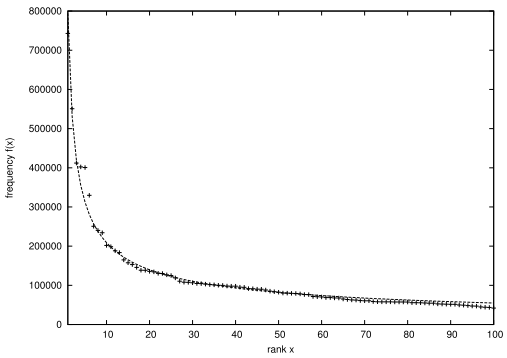

Fig. 6 Number of passengers at stations in JR East and the Zipf model $f(x)$.

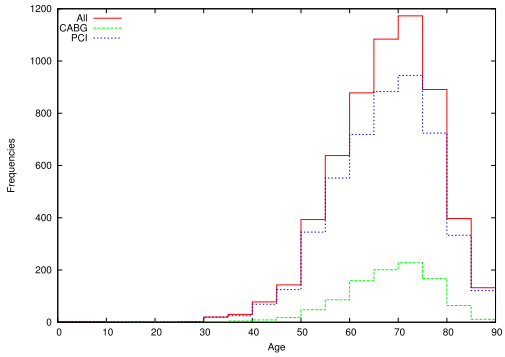

Fig. 7 The age distribution in two surgeries.

$p_1(1)p_2(1)$, and $p_{12}(d_1d_2) = p_1(d_1)p_2(d_1)$, we have the constants $a_{12} = a_1a_2$ and $d_1^{c_1}d_2^{c_2} = (d_1d_2)^{c_{12}}$. Similar to Theorem 3.3, we prove $p_1(x_1)p_2(x_2) \leq p_{12}(x_1x_2)$ for arbitrary $1 \leq x_1 \leq d_1$, and $1 \leq x_2 \leq d_2$. □

Note that this corresponds to Theorem 3.3 when $m_1 = m_2$, and $c_1 = c_2$. In the same manner, we can obtain the Zipf model for the combinations of multiple attributes and identify the least unidentifiable rank to quantity the risk of combined attributes being exploited to identify an individual uniquely.

4. Case Studies

4.1 Anonymity of Trajectory Data of the Japan Railway Co. (JR) Stations

JR East tried to sell trajectory data stored on its popular RFID-based fare card, Suica, without the consent of passengers in 2013, but they gave up because of excessive criticism. Thus, we studied the risk and utility of the anonymized data that JR East failed to sell.

First, using the least-squared method to fit the number of passengers at stations [14], we approximated a Zipf model $f(x)$ of the trajectory data as $a = 794{,}132 = 8 \times 10^5$, $c = 0.580$, as shown in **Fig. 6**. According to available open data, we observed similar behavior in the datasets of other railway companies, which demonstrated that ur Zipf model is a good generalization.

Next, using the Zipf model, we examined the least and the mean unidentifiable rank x^* for trajectories with $s = 1$ to 5 in **Table 2**. With a single station ($s = 1$), no records need to be suppressed. However, with $s = 2$, records with stations that exceeded the 1474-th rank could be uniquely identified and 6.5% records must be suppressed to satisfy 2-anonymity. The anonymity ratio reached 97.8% when $s = 3$, which means that most of the records need to be dropped and the utility of the anonymized data is lost.

Finally, we conclude that trajectory data need to be treated so the records cannot be linked to prevent combination becoming uniquely identifiable. The naive application of known anonymization algorithms could degrade the utility of open data because the complexity of the linked records becomes very high as the length of the trajectory increases, and thus many records might be suppressed.

4.2 Medical DPC Dataset

The DPC dataset, Disease, Procedure and Combination, covers medical records for more than 7 million patients in more than 1,000 hospitals [15].

With the international standard of disease, DPC data contains the followings; the hospital codes, the disease code, sex, age, ZIP code, the duration in hospital, the operation, the height, the weight, the degree of cancers, etc. The DPC dataset is used to study for hospital management and to provide a useful statistics in hospitals. Some of the statistical data is available online and used as open data for many purposes.

Figure 7 shows sample age distributions in the DPC dataset. The lapascope surgery (PCI) has odds ratio of 0.3774, which means that a lapascope surgery makes the probability of death to decrease by 0.3774 times of that who has an off-pump surgery (CABG). We can not find a significant difference between the two types of surgeries from the figure. We can approximate the age distribution in our model, where the most frequent age is about 70 years-old and the least is less than 40 years-old, which need to be suppressed to prevent the records from being identified.

5. Conclusions

In this study, to address the issues of anonymization, we proposed a new mathematical model based on the Zipf distribution. We demonstrated that our model is simple, but obtained a good fit with the actual distribution of the JR East trajectory data. We presented the primary property of our model and extended it to a more complex environment. Using our model, we defined the theoretical bound for reidentification, which yields the appropriate optimal level for anonymization.

Acknowledgments We thank members of the Technical Review Working Group, Cabinet Secretariat's IT Strategic Headquarters Research Society for Use and Circulation of Personal Data, for their discussion on the privacy threat in anonymized data.

References

[1] UK Information Commissioner's Office, Anonymisation: Managing Data Protection Risk Code of Practice (2012).
[2] Aggarwal, C.C. and Yu, P.S.: A General Survey of Privacy-Preserving Data Mining, Models and Algorithms, *Privacy-preserving data mining*, pp.11–52, Springer (2008).
[3] FTC Report, Protecting consumer privacy in an era of rapid change (2012), available from ⟨http://www.ftc.gov/sites/default/files/documents/reports/⟩.
[4] Skinner, C.J. and Elliot, M.J.: A measure of disclosure risk for microdata, *Journal of the Royal Statistical Society: Series B* (*statistical methodology*), Vol.64, No.4, pp.855–867 (2002).
[5] Benitez, K. and Malin, B.: Evaluating re-identification risks with respect to the HIPAA privacy rule, *Journal of the American Medical Informatics Association*, Vol.17, No.2, pp.169–177 (2010).
[6] UCI Machine Learning Repository, available from ⟨http://archive.ics.uci.edu/ml⟩.
[7] Ayala-Rivera, V., McDonagh, P., Cerqueus, T. and Murphy, L.: A Systematic Comparison and Evaluation of *k*-Anonymization Algorithms for Practitioners, *Trans. Data Privacy*, Vol.7, No.3, pp.337–370 (2014).
[8] Templ, M., Kowarik, A. and Meindl, B.: Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation, `sdcMicro` package in R (2015).
[9] Samarati, P.: Protecting respondents identities in microdata release, *IEEE Trans. Knowledge and Data Engineering*, Vol.13, No.6, pp.1010–1027 (2001).
[10] Sweeny, L.: *k*-anonymity, *Int. J. Uncertainty, Fuzziness & Knowledge-Based System*, Vol.10, pp.571–588 (2002).
[11] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: *L*-diversity: Privacy beyond k-anonymity, *ACM Trans. Knowl. Discov. Data 1*, Vol.1, Article 3 (2007).
[12] Li, N. and Li, T.: *t*-Closeness: Privacy beyond *k*-anonymity and *ℓ*-diversity, *Proc. IEEE 23rd Int'l Conf. on Data Engineering* (*ICDE'07*), pp.106–115, IEEE (2007).
[13] Dwork, C.: Differential privacy, *Automata, Languages and Programming Lecture Notes in Computer Science*, Vol.4052, pp.1–12 (2006).
[14] East Japan Railway Company, The numbers of passengers in 2012 (in Japanese) (2013), available from ⟨http://www.jreast.co.jp/passenger/⟩.
[15] Yasunaga, H., Horiguchi, H., Kuwabara, K., Matsuda, S., Fushimi, K. and Hashimoto, H.: Outcomes After Laparoscopic or Open Distal Gastrectomy for Early-Stage Gastric Cancer: A Propensity-Matched Analysis, *Annals of Surgery*, Vol.257, No.4, pp.640–646 (2012).
[16] Yao, C., Wang, X.S. and Jajodia, S.: Checking for *k*-anonymity violation by views, *Proc. 31st International Conference on Very Large Data Bases* (*VLDB '05*), VLDB Endowment, pp.910–921 (2005).
[17] Xiao, X. and Tao, Y.: Anatomy: Simple and effective privacy preservation, *Proc. 32nd International Conference on Very Large Data Bases* (*VLDB '06*), VLDB Endowment, pp.139–150 (2006).
[18] Poulis, G., Skiadopoulos, S. and Loukides, G. and Gkoulalas-Divanis, A.: Apriori-based algorithms for k^m-anonymizing trajectory data, *Transactions on Data Privacy*, Vol.7, No.2, pp.165–194 (2014).
[19] Choi, S., Hikita, T. and Yamaguchi, R.S.: A Comparison of Anonymizing between Real Log of Transit Ridership and Theoretical Model, *Computer Security Symposium 2015* (*CSS 2015*), 3B4-5, pp.1289–1296, IPSJ (2015).
[20] Basu, A., Monreale, A., Trasarti, R., Corena, J.C., Giannotti, F., Pedreschi, D., Kiyomoto, S., Miyake, Y. and Yanagihara, T.: A risk model for privacy in trajectory data, *Journal of Trust Management*, Vol.2, No.9 (2015).
[21] Monreale, A., Trasarti, R., Pedreschi, D., Renso, C. and Bogorny, V.: *C*-safety: A framework for the anonymization of semantic trajectories, *Trans. on Data Privacy*, Vol.4, No.2, pp.73–101 (2011).
[22] Garfinkel, S.L.: De-Identification of Personal Information, NISTIR 8053 (2015).
[23] Emam, K.E. and Malin, B.: Appendix B: Concepts and Methods for De-identifying Clinical Trial Data, *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. (2015).

Hiroaki Kikuchi was born in Japan. He received his B.E., M.E. and Ph.D. degrees from Meiji University in 1988, 1990 and 1994. After working in Fujitsu Laboratories Ltd. from 1990, in Tokai University from 1994, respectively, he joined Meiji University in 2013. He is currently a Professor at the Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University. He is a dean of Graduate School of Advanced Mathematical Sciences, Meiji University. He was a visiting researcher at the School of Computer Science, Carnegie Mellon University in 1997. His main research interests are fuzzy logic, cryptographic protocol, network security, and privacy-preserving data mining. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), the Japan Society for Fuzzy Theory and Systems (SOFT), IEEE and ACM. He is a fellow of the Information Processing Society of Japan (IPSJ).

Katsumi Takahashi was born in Japan. He received a B.S. in mathematics from Tokyo Institute of Technology and a Ph.D. in information science and technology from the University of Tokyo in 1988 and 2006. He joined NTT in 1988 and has studied information retrieval, data mining, security, and privacy. He is an executive research scientist of NTT Secure Platform Laboratories. He is a member of IEICE. He is a fellow of IPSJ.