**Regular Paper**

# Maintaining Information in Differential Privacy by Using Insensitive Relationships between Personal Attributes

Takayasu Yamaguchi[†‡]and Hiroshi Yoshiura[‡]

[†]Research Laboratories, NTT DOCOMO, Inc.
[‡]Department of Informatics, University of Electro-Communications
yamaguchitaka@nttdocomo.com, yoshiura@uec.ac.jp

***Abstract*** - Statistics generated from collections of personal data are used in both the public and private sectors, but there is a risk of the personal data being inferred from the statistics. To prevent such inference and protect the privacy of the individuals represented by the statistics, anonymization is used to modify the statistics. Among the methods for anonymization, those based on differential privacy are more promising because of their generality and rigid mathematical basis. Modification to achieve differential privacy, however, degrades the accuracy of the statistics. Because the degree of modification is proportional to the number of an individual's attributes in the statistics, the larger the number of attributes, the more degraded the statistics. On the other hand, the smaller the number, the less useful the statistics. We propose optimizing this trade-off by making use of the relationships between personal attributes. The proposed method uses only some of the attributes in the personal data while indirectly using the remaining ones by estimating their values from the values of the attributes used. The relationships between attributes are used for this estimation. However, if the relationships themselves are sensitive, i.e. they reflect personal data, additional anonymization is needed, which could degrade the effectiveness of the method. Thus, the key to effectiveness is identifying the relationships between personal attributes that are precise and insensitive. The effectiveness of our method was demonstrated by implementing it with a Laplace mechanism, a representative method for implementing differential privacy, and by evaluating the implemented system using the MovieLens 1M dataset.

***Keywords***: security, big data, data mining, anonymization, differential privacy

## 1 INTRODUCTION

Statistics generated from collections of personal data are used in both the public and private sectors, but there is a risk of the personal data being inferred from the statistics. To prevent such inference and protect the privacy of the individuals represented by the statistics, anonymization is used to modify the statistics. Among the methods for anonymization, those based on differential privacy are more promising because of their generality and rigid mathematical basis [1]-[4]. Modification to achieve differential privacy, however, degrades the accuracy of the statistics. Because the degree of modification is proportional to the number of an individual's attributes in the statistics, the larger the number of attributes, the more degraded the statistics. On the other hand, the smaller the number, the less useful the statistics.

We propose optimizing this trade-off by making use of the relationships between personal attributes. The proposed method uses only some of the attributes in the personal data while indirectly using the remaining ones by estimating their values from the values of the attributes used. The relationships between attributes are used for this estimation. For example, given the attributes of height, weight, and gender, height and gender are used and weight is estimated using the relationships between the three attributes, thus reducing the amount of modification by a third while maintaining the usefulness of the statistics. However, if the relationships themselves are sensitive, i.e. they reflect personal data, additional anonymization is needed, which could degrade the effectiveness of the method. Thus, the key to effectiveness is identifying the relationships between personal attributes that are precise and insensitive. The effectiveness of our method was demonstrated by implementing it with a Laplace mechanism, a representative method for implementing differential privacy, and by evaluating the implemented system using the MovieLens 1M dataset.

Section 2 describes related work. Section 3 describes our strategy for reducing the number of attributes, and Section 4 describes the implementation and evaluation. Section 5 summarizes the key points and mentions future work.

## 2 RELATED WORK

### 2.1 Anonymization Methods

Anonymization is applied either to records of personal information (called microdata) or to statistics calculated from microdata. The main purpose of anonymizing microdata is to prevent linking records to specific persons and that of anonymizing statistics is to prevent inferring the original microdata from which the statistics were derived. Representative methods for anonymizing microdata are methods based of k-anonymity [5], l-diversity [6], and t-closeness [7]. A set of techniques generically called statistical disclosure control are used for anonymizing statistics [8][9]. Methods based on differential privacy [1] and probabilistic k-anonymity [10][11] are used

ISSN1883-4566 © 2018 - Informatics Society and the authors. All rights reserved.

for both. Anonymization is implemented by modifying data, and methods for modification include perturbing, swapping, aggregating, and omitting data [12]. Thus, it is inevitable that anonymization degrades the original data and statistics. Two essential issues of anonymization are therefore security and data quality [13]. Security is related to how well the purpose of anonymization (i.e. preventing linkage and inference) can be achieved. Data quality is related to how well the information that can be extracted from the data or statistics is preserved. Because there is a trade-off between security and data quality, our goal is to improve security while maintaining the quality or to improve quality (i.e. minimize information degradation) while maintaining security.

## 2.2 Differential Privacy and Laplace Mechanism

Differential privacy is a criterion for achieving anonymization security that is particularly promising due to its generality and rigid mathematical basis. It is expressed as an inequality:

$$\begin{gathered} \forall D_1, \forall D_2 \in D, \forall S \subseteq Range(\mathcal{K}) \\ \{Pr[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S]\} \end{gathered} \tag{1}$$

where $\mathcal{K}$ is a mechanism for calculating and anonymizing statistics from microdata $D_x$, $D$ is a domain of $\mathcal{K}$, i.e. the set of microdata treated by $\mathcal{K}$, and $\epsilon$ is a security parameter. Thus equation (1) represents that the ratio between an item being included in the statistics produced from $D_1$ and the item included in the statistics produced from $D_2$ is limited by exponential ($\epsilon$) if $D_1$ and $D_2$ differ in only one record (i.e. differ for only one person). The smaller the $\epsilon$, the more strict the anonymization, i.e. the greater the modification of the statistics $\mathcal{K}(D_x)$ and the more degraded the statistics.

The Laplace mechanism, a representative mechanism for implementing differential privacy, adds Laplacian noise with a scale parameter ($\lambda$) to the statistics. The scale parameter is sometimes referred to as the diversity. The larger the $\lambda$, the greater the noise and the more degraded the statistics. The Laplace mechanism achieves differential privacy by satisfying an inequality:

$$\lambda \geq \frac{\Delta f}{\epsilon} \tag{2}$$

where $\epsilon$ is the security parameter mentioned above, and $\Delta f$ is the sensitivity of the statistical value against the microdata. It is the maximum change in the statistics when one record (i.e. one person) in the microdata is changed. The smaller the $\Delta f$, the better the quality of the anonymized statistics. This is reasonable considering the meaning of differential privacy mentioned above.

## 2.3 Issues Concerning Quality Maintenance

The loss of quality due to using the Laplace mechanism can be decreased by reducing $\Delta f$, which depends on the type of statistic. For example, $\Delta f$ is proportional to $1/n$ for the average value of the microdata, where $n$ is the number of records in the microdata. This is because the maximum change in the average value due to the change of one record is proportional to $1/n$. The $\Delta f$ is also proportional to $1/n$ for the variance. On the other hand, it is independent of $n$ for the maximum and minimum. Thus, average and variance suffer less degradation due to using the Laplace mechanism than maximum and minimum. One approach to maintaining quality is thus to reduce the target statistics to a combination of statistics that have good properties such as average and deviation [2] [14]. Though much improvement has been already done in this approach, the current differential privacy mechanisms are not practical for many applications in the trade-off between security and data quality.

By "sensitivity" we mean the maximum change in a statistic when any possible change in one record is considered in the context of any possible combination of other records [15]-[17]. Thus, another approach to reducing $\Delta f$ is to ignore rare cases [18]-[20] that make $\Delta f$ large. This approach, however, degrades anonymization security because anonymization does not work if rare cases occur.

The fewer the attributes of a record (i.e. of a person), the lower the sensitivity [21] [22]. For example, assume a record has three attributes: gender, height, and weight. The statistics to be calculated are the frequencies of tall men (men taller than 180 cm) and heavy men (men heavier than 80 kg). The sensitivity is 2 because a man of 190 cm and 90 kg could change into a woman of 150 cm and 50 kg[1]. If we consider only gender and height, the sensitivity is reduced to 1 because we can count only the frequency of tall men. Thus, a third approach is to consider a subset of the record attributes. However, straightforward methods in this direction simply use less information to calculate the statistics [23], which is less useful for such applications as marketing of sporting goods.

As shown in Table 1, increasing the number of attributes increases the richness of the information provided by the statistics, but it also increases the privacy risk. It results in the increased sensitivity ($\Delta f$) and more modification, which reduces the usefulness of the statistics. Reducing the number of attributes reduces the privacy risk and the amount of modification needed, but the information provided by the statistics becomes poorer, and the statistics become less useful. Thus, we need a way to generate statistics from fewer attributes that are information-rich.

Using primary component analysis (PCA) is one way to reduce the number of attributes while maintaining information quality [24]-[26]. However, PCA may reveal private informa-

Table 1: Technical requirements

| | Number of attributes | |
|---|---|---|
| | Many | Few |
| Information in statistics | Rich | Poor → Rich |
| Danger to privacy<br>Necessary modification | Large | Small |
| Usefulness | Bad | Bad → Good |

[1] This example is a little bit more complex than that described in the abstract.

tion in itself, e.g. how the values of the attributes are related.

# 3 STRATEGY FOR REDUCING NUMBER OF ATTRIBUTES

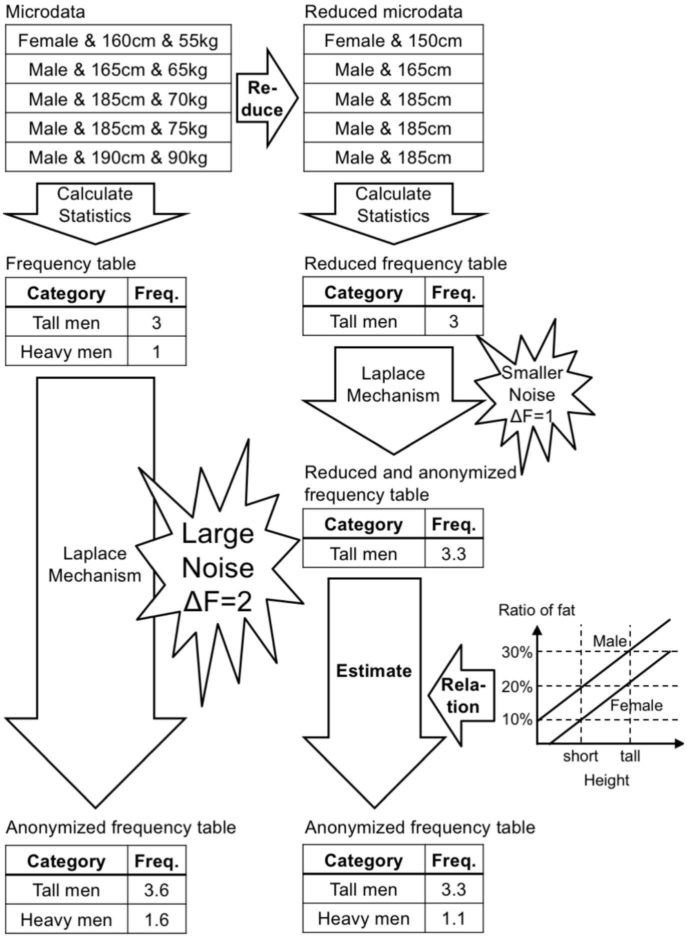
We maintain the quality of the anonymized statistics by using the third approach mentioned above; i.e. we use a subset of attributes in the microdata to calculate the statistics. We maintain the quality of information by using a subset of the attributes and estimating the values of the other attributes by using the relationships between the used attributes and the other attributes. We derive the relationships between the attributes from public information, thereby preventing the revelation of private information. If we cannot derive the relationships from public information only, we derive them by using statistics for the microdata such as the average and deviation, which are less sensitive than the target statistics.

Our strategy for anonymizing the statistics with less degradation comprises seven steps.

1. Calculate insensitive statistics, which suffer less from anonymization.
2. Anonymize insensitive statistics.
3. Use anonymized statistics to establish relationships between attributes.
4. Create statistics for subset of attributes without anonymization.
5. Anonymize these statistics with small noise.
6. Use relationships to estimate values of unused attributes.
7. Calculate target statistics from used attributes and values of unused attributes.

We explain our strategy by using the example shown in Fig. 1. The microdata are shown at the top-left. The target statistics are the frequencies of tall men and heavy men which are represented by a frequency table. A conventional method to obtain an anonymized frequency table is to generate a frequency table and anonymize it. However, the sensitivity in this case is 2 (as mentioned above), and may be too high to maintain the quality of the anonymized statistics. We thus use only values of gender and height in the microdata (at the upper-right in Fig. 1) and generate a frequency table from the reduced microdata. We anonymize this reduced frequency table, which reduces the sensitivity to 1.

We then estimate the frequency of heavy men by using the relationships between gender, height, and weight. Ideally, these relationships would be precisely obtained from public information such as statistics published by the Ministry of Health, Labour and Welfare. Otherwise, we can obtain the relationships as a combination of insensitive statistics such as average and deviation.



Frequency table

| Category | Freq. |
|---|---|
| Tall men | 3 |
| Heavy men | 1 |

Reduced frequency table

| Category | Freq. |
|---|---|
| Tall men | 3 |

Reduced and anonymized frequency table

| Category | Freq. |
|---|---|
| Tall men | 3.3 |

Anonymized frequency table

| Category | Freq. |
|---|---|
| Tall men | 3.6 |
| Heavy men | 1.6 |

Anonymized frequency table

| Category | Freq. |
|---|---|
| Tall men | 3.3 |
| Heavy men | 1.1 |

Figure 1: Example implementation of proposed strategy

# 4 EXAMPLE IMPLEMENTATION AND EVALUATION

Our use case is the construction of a recommendation system for movies. The system finds the top $m$ movie categories that are best suited for the age, gender, and occupation of the user. A post-processing system selects movies from those categories [27] [28].

## 4.1 Dataset

We used the MovieLens 1M dataset [29] as example microdata for this system. The dataset consists of 1,000,209 records, each of which is a user's rating of a movie. It contains data for 6,040 users and 3,952 movies. Table 2 shows a part of this dataset. Note that there are 18 basic movie categories such as Action, Adventure and Comedy, and there are $2^{18} - 19$ combinational categories such as "Action & Adventure," "Comedy & Romance," and "Action & Adventure & Sci-Fi."[2] Categories combining two basic categories, e.g. "Action & Adventure," are called second-order categories, those combining three are called third-order categories, and, in general, those combining $i$ basic categories are called $i$th-

[2] There are $2^{18} - 19$ combinations of basic categories except for the basic categories themselves; a null combination is excluded.

Table 2: Part of MovieLens 1M dataset

| User ID | Profile | Movie<br>(category) | Rating |
|---|---|---|---|
| 1 | Male, 18–24,<br>programmer | Waterworld,<br>(Action & Adventure) | 5 |
| 1 | Male, 18–24,<br>programmer | Beverly Hills Cop,<br>(Action & Comedy) | 4 |
| 2 | Female, 25–34,<br>writer | Sabrina,<br>(Comedy & Romance) | 3 |
| 3 | Male, 18–24,<br>programmer | Star Trek,<br>(Action & Adventure & Sci-Fi) | 5 |
| 4 | Female, 25–34,<br>artist | Sound of Music,<br>(Musical) | 4 |

order categories. Thus, we have $2^{18} - 1$ basic and combinational categories. The user ratings range from 1 to 5, with 5 being the highest.

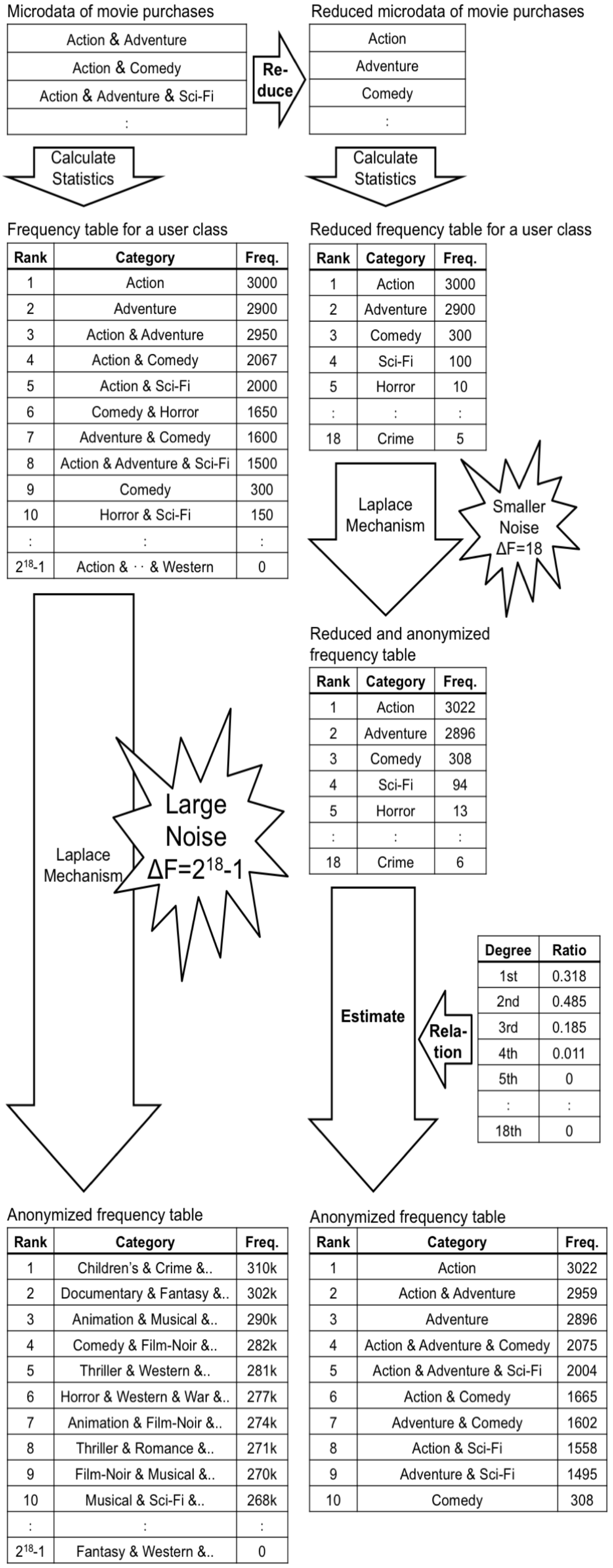
We took the rating data as sales records, i.e. we took that a user bought a kind of categories of movie if user rated it level 4 or 5. We assume that recommender can find which categories of movie are popular for user's profiles from the dataset as top $m$ rankings of categories. Since purchased movie categories are represented by combinations of 18 basic categories, these features could change in $2^{18} - 1$ patterns and reveal privacy and, thus, should be modified.

## 4.2 System Design

The microdata of movie purchases are shown at the top-left in Fig. 2. These microdata are linked to user profiles and are used to generate a frequency table for each user class (e.g. "Female & 25-34 & writer"). They were anonymized, as shown at the bottom-left. The top ten categories in the anonymized table were recommended by the system with respect to user class. In the anonymization, however, the sensitivity was $2^{18} - 1$ because one person could purchase all $2^{18} - 1$ categories or purchase none of them.

To reduce the sensitivity, we used only 18 basic categories to generate the frequency table shown at the middle-right of Fig. 2 and anonymize it. The sensitivity was thereby reduced to 18. The top ten categories were recommended from the anonymized frequency table. However, this frequency table is information-poor because it contains only basic categories. We call this method the reduced-attributes method.

To make the reduced-attributes method information-rich, we propose using the relationships between the basic categories and the higher-order categories, which are shown in the table below the reduced and anonymized frequency table in Fig. 2. This table represents the rates of the $i$th-order categories appearing in the top ten; i.e. the rates of the first, second, and third-order categories are $3 : 5 : 2$. Thus, the three most frequent basic categories are included in the top ten list. Similarly, the five most frequent second-order categories are included as well as the two most frequent third-order categories. The anonymized frequency table (shown at bottom-right in Fig. 2is generated from the frequency table created by the reduced-attributes method and the table of relationships.



Microdata of movie purchases

Action & Adventure
Action & Comedy
Action & Adventure & Sci-Fi
:

Reduced microdata of movie purchases

Action
Adventure
Comedy
:

Frequency table for a user class

| Rank | Category | Freq. |
|---|---|---|
| 1 | Action | 3000 |
| 2 | Adventure | 2900 |
| 3 | Action & Adventure | 2950 |
| 4 | Action & Comedy | 2067 |
| 5 | Action & Sci-Fi | 2000 |
| 6 | Comedy & Horror | 1650 |
| 7 | Adventure & Comedy | 1600 |
| 8 | Action & Adventure & Sci-Fi | 1500 |
| 9 | Comedy | 300 |
| 10 | Horror & Sci-Fi | 150 |
| : | : | : |
| $2^{18}$-1 | Action & ·· & Western | 0 |

Reduced frequency table for a user class

| Rank | Category | Freq. |
|---|---|---|
| 1 | Action | 3000 |
| 2 | Adventure | 2900 |
| 3 | Comedy | 300 |
| 4 | Sci-Fi | 100 |
| 5 | Horror | 10 |
| : | : | : |
| 18 | Crime | 5 |

Reduced and anonymized frequency table

| Rank | Category | Freq. |
|---|---|---|
| 1 | Action | 3022 |
| 2 | Adventure | 2896 |
| 3 | Comedy | 308 |
| 4 | Sci-Fi | 94 |
| 5 | Horror | 13 |
| : | : | : |
| 18 | Crime | 6 |

| Degree | Ratio |
|---|---|
| 1st | 0.318 |
| 2nd | 0.485 |
| 3rd | 0.185 |
| 4th | 0.011 |
| 5th | 0 |
| : | : |
| 18th | 0 |

Anonymized frequency table

| Rank | Category | Freq. |
|---|---|---|
| 1 | Children's & Crime &.. | 310k |
| 2 | Documentary & Fantasy &.. | 302k |
| 3 | Animation & Musical &.. | 290k |
| 4 | Comedy & Film-Noir &.. | 282k |
| 5 | Thriller & Western &.. | 281k |
| 6 | Horror & Western & War &.. | 277k |
| 7 | Animation & Film-Noir &.. | 274k |
| 8 | Thriller & Romance &.. | 271k |
| 9 | Film-Noir & Musical &.. | 270k |
| 10 | Musical & Sci-Fi &.. | 268k |
| : | : | : |
| $2^{18}$-1 | Fantasy & Western &.. | 0 |

Anonymized frequency table

| Rank | Category | Freq. |
|---|---|---|
| 1 | Action | 3022 |
| 2 | Action & Adventure | 2959 |
| 3 | Adventure | 2896 |
| 4 | Action & Adventure & Comedy | 2075 |
| 5 | Action & Adventure & Sci-Fi | 2004 |
| 6 | Action & Comedy | 1665 |
| 7 | Adventure & Comedy | 1602 |
| 8 | Action & Sci-Fi | 1558 |
| 9 | Adventure & Sci-Fi | 1495 |
| 10 | Comedy | 308 |

Figure 2: Overview of example implementation

To estimate the frequencies of the second- and third-order categories, we assume that the more frequent the basic categories, the more frequent the second-order category consisting of these basic categories. We thus estimate the frequency of a second-order category by using the sum of frequencies of the basic element categories, e.g. frequency of "Action & Adventure" is (frequency of Action + frequency of Adventure) / 2. We estimate the frequency of third-order categories similarly.

The relationships between the attributes reflects the original microdata and thus should be anonymized. We use the security parameter ($\epsilon$) for both anonymization of the frequency table and anonymization of the relationships by using $r\epsilon$ and $(1-r)\epsilon$, respectively. In our implementation, we set $r = 0.1$.

Since the proposed method uses only 18 basic categories to generate the frequency table, the sensitivity of the frequency table is 18. The sensitivity of each value in the relationship table is also 18 because we have only 18 kinds of categories, e.g. basic through the 18th category.

## 4.3 Evaluation

We first used our recommendation system without anonymization and obtained the top ten recommendations for each user class. We used these recommendations as correct recommendations. We then used the system with anonymization. Three anonymization methods were used. The first one was a conventional method that added Laplacian noise with a sensitivity of $2^{18} - 1$. The second one was a reduced-attributes method that used up to $i$th movie categories and ignored other categories. The second one was evaluated with $i$ = 1, 2, and 3. The third one was the proposed method.

Users were classified by gender and 7 age classes (i.e. 14 classes) in the first experiment, by occupation (21 classes) in the second experiment, and by gender, age, and occupation (294 classes) in the third experiment. For each experiment, two values of $\epsilon$ were used: 2.0 for weak anonymization, 1.0 for strong anonymization.

The recommendation accuracy (in percent) was measured in terms of the number of recommendations included in both the correct recommendations and the recommendations provided by the system. It was represented as precision at $m$ (P@$m$) [30] and measured for each method and each parameter value (order $i$ and security parameter $\epsilon$).

Table 3: Results for using gender and age (P@10)

| Required Security Level | Without Anonymization | Conventional Method | Reduced-Attributes Method (Order) | | | Proposed Method |
|---|---|---|---|---|---|---|
| | | | 1st | 2nd | 3rd | |
| None | 100% | 100% | 28% | 78% | 99% | 48% |
| Weak | NG | 0% | 28% | 51% | 1% | 48% |
| Strong | NG | 0% | 28% | 34% | 0% | 48% |

Table 4: Results for using gender and age (P@20)

| Required Security Level | Without Anonymization | Conventional Method | Reduced-Attributes Method (Order) | | | Proposed Method |
|---|---|---|---|---|---|---|
| | | | 1st | 2nd | 3rd | |
| None | 100% | 100% | 28% | 78% | 99% | 51% |
| Weak | NG | 0% | 28% | 67% | 4% | 51% |
| Strong | NG | 0% | 28% | 43% | 1% | 51% |

Table 5: Results for using occupation (P@10)

| Required Security Level | Without Anonymization | Conventional Method | Reduced-Attributes Method (Order) | | | Proposed Method |
|---|---|---|---|---|---|---|
| | | | 1st | 2nd | 3rd | |
| None | 100% | 100% | 30% | 82% | 100% | 50% |
| Weak | NG | 0% | 30% | 37% | 2% | 49% |
| Strong | NG | 0% | 30% | 18% | 1% | 46% |

Table 6: Results for using occupation (P@20)

| Required Security Level | Without Anonymization | Conventional Method | Reduced-Attributes Method (Order) | | | Proposed Method |
|---|---|---|---|---|---|---|
| | | | 1st | 2nd | 3rd | |
| None | 100% | 100% | 30% | 82% | 100% | 50% |
| Weak | NG | 0% | 30% | 54% | 4% | 50% |
| Strong | NG | 0% | 30% | 37% | 3% | 50% |

Table 7: Results for using gender, age, and occupation (P@10)

| Required Security Level | Without Anonymization | Conventional Method | Reduced-Attributes Method (Order) | | | Proposed Method |
|---|---|---|---|---|---|---|
| | | | 1st | 2nd | 3rd | |
| None | 100% | 100% | 28% | 77% | 93% | 36% |
| Weak | NG | 0% | 26% | 10% | 1% | 36% |
| Strong | NG | 0% | 24% | 8% | 1% | 27% |

Table 8: Results for using gender, age, and occupation (P@20)

| Required Security Level | Without Anonymization | Conventional Method | Reduced-Attributes Method (Order) | | | Proposed Method |
|---|---|---|---|---|---|---|
| | | | 1st | 2nd | 3rd | |
| None | 100% | 100% | 29% | 78% | 96% | 40% |
| Weak | NG | 0% | 28% | 18% | 2% | 40% |
| Strong | NG | 0% | 27% | 15% | 2% | 34% |

Table 9: Detailed results for using gender and age (P@10, $\epsilon = 2.0$)

| Gender | Age | Reduced-Attributes Method (Order) | | | Pro-posed Method |
|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | |
| Female | –18 | 30% | 0% | 0% | 50% |
| | 18–24 | 30% | 20% | 0% | 50% |
| | 25–34 | 30% | 40% | 10% | 50% |
| | 35–44 | 30% | 40% | 0% | 50% |
| | 45–49 | 30% | 30% | 0% | 60% |
| | 50–55 | 30% | 0% | 0% | 50% |
| | 56– | 20% | 0% | 0% | 40% |
| Male | –18 | 30% | 0% | 0% | 40% |
| | 18–24 | 30% | 60% | 0% | 50% |
| | 25–34 | 30% | 70% | 0% | 50% |
| | 35–44 | 20% | 60% | 0% | 40% |
| | 45–49 | 30% | 30% | 0% | 50% |
| | 50–55 | 30% | 40% | 0% | 50% |
| | 56– | 30% | 40% | 0% | 50% |
| Weighted Average | | 28% | 51% | 1% | 48% |

Table 10: Detailed results for using gender and age (P@10, $\epsilon = 1.0$)

| Gender | Age | Reduced-Attributes Method (Order) | | | Pro-posed Method |
|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | |
| Female | –18 | 30% | 0% | 0% | 50% |
| | 18–24 | 30% | 10% | 0% | 50% |
| | 25–34 | 30% | 20% | 0% | 50% |
| | 35–44 | 30% | 10% | 0% | 50% |
| | 45–49 | 30% | 10% | 0% | 60% |
| | 50–55 | 30% | 0% | 0% | 50% |
| | 56– | 20% | 0% | 0% | 30% |
| Male | –18 | 30% | 0% | 0% | 20% |
| | 18–24 | 30% | 40% | 0% | 50% |
| | 25–34 | 30% | 60% | 0% | 50% |
| | 35–44 | 20% | 30% | 0% | 40% |
| | 45–49 | 30% | 20% | 0% | 50% |
| | 50–55 | 30% | 10% | 0% | 50% |
| | 56– | 30% | 30% | 0% | 50% |
| Weighted Average | | 28% | 34% | 0% | 48% |

Table 11: Detailed results for using occupation (P@10, $\epsilon = 2.0$)

| Job | Reduced-Attributes Method (Order) | | | Pro-posed Method |
|---|---|---|---|---|
| | 1st | 2nd | 3rd | |
| academic / educator | 30% | 40% | 0% | 50% |
| artist | 30% | 30% | 0% | 50% |
| clerical / admin | 30% | 10% | 0% | 50% |
| college / grad student | 30% | 70% | 0% | 50% |
| customer service | 30% | 0% | 0% | 50% |
| doctor / health care | 30% | 10% | 0% | 50% |
| executive / managerial | 30% | 70% | 10% | 50% |
| farmer | 20% | 10% | 0% | 20% |
| homemaker | 20% | 10% | 0% | 50% |
| K–12 student | 30% | 10% | 0% | 50% |
| lawyer | 30% | 10% | 0% | 50% |
| programmer | 30% | 40% | 10% | 50% |
| retiree | 30% | 30% | 0% | 50% |
| sales / marketing | 30% | 10% | 0% | 50% |
| scientist | 30% | 10% | 0% | 50% |
| self-employed | 30% | 30% | 0% | 50% |
| technician / engineer | 30% | 20% | 0% | 50% |
| tradesman / craftsman | 30% | 0% | 0% | 40% |
| unemployed | 20% | 10% | 0% | 20% |
| writer | 30% | 20% | 0% | 40% |
| other / not specified | 0% | 50% | 0% | 50% |
| Weighted Average | 30% | 37% | 2% | 49% |

Table 12: Detailed results for using occupation (P@10, $\epsilon = 1.0$)

| Job | Reduced-Attributes Method (Order) | | | Pro-posed Method |
|---|---|---|---|---|
| | 1st | 2nd | 3rd | |
| academic / educator | 30% | 20% | 0% | 50% |
| artist | 30% | 20% | 0% | 50% |
| clerical / admin | 30% | 0% | 0% | 50% |
| college / grad student | 30% | 60% | 0% | 50% |
| customer service | 30% | 0% | 0% | 50% |
| doctor / health care | 30% | 10% | 0% | 40% |
| executive / managerial | 30% | 20% | 0% | 50% |
| farmer | 20% | 10% | 0% | 20% |
| homemaker | 20% | 0% | 0% | 40% |
| K–12 student | 30% | 0% | 0% | 50% |
| lawyer | 30% | 10% | 0% | 40% |
| programmer | 30% | 20% | 10% | 50% |
| retiree | 30% | 20% | 0% | 50% |
| sales / marketing | 30% | 0% | 0% | 50% |
| scientist | 30% | 0% | 0% | 40% |
| self-employed | 30% | 20% | 0% | 40% |
| technician / engineer | 30% | 0% | 0% | 50% |
| tradesman / craftsman | 30% | 0% | 0% | 30% |
| unemployed | 10% | 10% | 0% | 10% |
| writer | 30% | 0% | 0% | 30% |
| other / not specified | 30% | 20% | 0% | 50% |
| Weighted Average | 30% | 18% | 1% | 46% |

## 4.4 Results

Table 3 (P@10) shows the results of the first experiment when $m$ equalled ten. The reduced-attributes method used up to the first-, second-, and third-order movie categories. The conventional anonymization method was the best without anonymization but could not produce any correct recommendations after anonymization. The accuracy of the reduced-attributes method was lower than that of the conventional method without anonymization because it used only some attributes. It had better accuracy than the conventional method after anonymization. The reduced-attributes method was the best with weak anonymization ($\epsilon = 2.0$). The proposed method was better than the reduced-attributes method for most parameter values but worse than the second-order reduced-attributes method with weak anonymization. The advantage of the proposed method over the reduced-attributes method was larger when anony-mization was strong ($\epsilon = 1.0$).

Table 4 (P@20) shows the results of the first experiment when $m$ equaled twenty. The results in all cases were better than those when $m$ equaled ten (Table 3). The conventional anonymization method was the best without anonymization but the worst with anonymization. The reduced-attributes method was the best with weak anonymization ($\epsilon = 2.0$). The proposed method was the best with strong anonymization ($\epsilon = 1.0$)

Tables 5, 6, 7, and 8 show the results of the second and third experiments. Again, the conventional method could not produce any correct recommendations after anonymization. The proposed method was better than the reduced-attributes method for the case using strong anonymization ($\epsilon = 1.0$). It was better than the reduced-attributes method for the case using weak anonymization ($\epsilon = 2.0$) except for the result for using occupation (P@20) (Table 6).

Table 9 shows detailed results for the first experiment with $\epsilon = 2.0$. Accuracy is shown for each user class, anonymization method, and parameter value. The results for the conventional method are omitted because all values were 0%; i.e. none of the recommendations produced was correct. The proposed method was more accurate than the first- and third-order applications of the reduced-attributes method and partly less accurate than the second-order application. Its accuracy was stable across user classes while that of the reduced-attributes method greatly depended on the user class.

Table 10 shows detailed results for the first experiment with $\epsilon = 1.0$. The advantage of the proposed method over the reduced-attributes method was larger than that with $\epsilon = 2.0$ (Table 9).

Tables 11 and 12 show detailed results for the second experiment with $\epsilon = 2.0$ and $\epsilon = 1.0$, respectively. As in the other experiments, the advantage of the proposed method over the reduced-attributes method was larger when $\epsilon = 1.0$ than when $\epsilon = 2.0$.

## 4.5 Discussion

The conventional method was the best without anonymization but the worst with anonymization. Comparing Tables 3, 4, 5, 6, 7, and 8, we see that the proposed method outperformed the reduced-attributes method when strong security was required. This is because the proposed method is degraded less than the reduced-attributes method when strong security is required.

Table 13: Detailed results for using gender and age with number of records (P@10, $\epsilon = 2.0$)

| Gender | Age | Number of Records | Reduced-Attributes Method (Order) 1st | 2nd | 3rd | Proposed Method |
|---|---|---|---|---|---|---|
| Female | –18 | 5,337 | 30% | 0% | 0% | 50% |
| | 18–24 | 23,910 | 30% | 20% | 0% | 50% |
| | 25–34 | 53,537 | 30% | 40% | 10% | 50% |
| | 35–44 | 29,785 | 30% | 40% | 0% | 50% |
| | 45–49 | 14,677 | 30% | 30% | 0% | 60% |
| | 50–55 | 11,835 | 30% | 0% | 0% | 50% |
| | 56– | 6,498 | 20% | 0% | 0% | 40% |
| Male | –18 | 10,273 | 30% | 0% | 0% | 40% |
| | 18–24 | **76,889** | 30% | **60%** | 0% | **50%** |
| | 25–34 | **169,017** | 30% | **70%** | 0% | **50%** |
| | 35–44 | **86,908** | 20% | **60%** | 0% | **40%** |
| | 45–49 | 34,799 | 30% | 30% | 0% | 50% |
| | 50–55 | 33,249 | 30% | 40% | 0% | 50% |
| | 56– | 18,567 | 30% | 40% | 0% | 50% |
| Weighted Average | | | 28% | 51% | 1% | 48% |

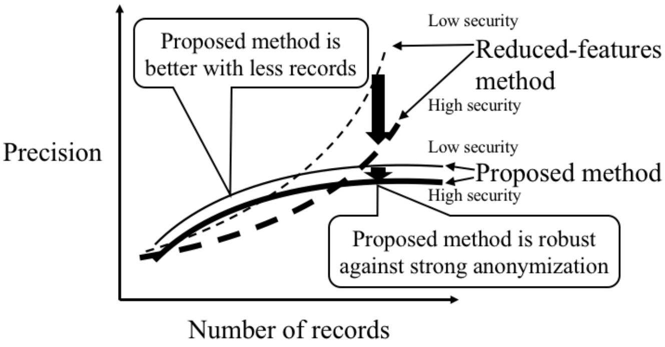


Figure 3: Features of proposed and reduced-attributes methods

Table 13 extends Table 9 by showing the number of records and focusing on the case in which the proposed method performed worse than the reduced-attributes method. From this table, we can see that the proposed method works well when there is a small number of records.

The features of the proposed and the reduced-attributes methods are summarized in Fig. 3.

## 5 CONCLUSION

The more personal features used in statistics, the strong anonymization needed, making statistics useless, while less personal features makes statistics information-poor. We have proposed a strategy for implementing differential privacy to reduce Laplacian noise and maintain the quality of anonymized statistics. A subset of attributes in the microdata is used to generate statistics, reducing noise to add on the statistics with reduced attributes, and restoring information in the statistics by using relationships between the selected attributes and the other attributes. Our method uses less features to make information-poor statistics and uses knowledge to make it information-rich. Our method has been implemented for differential privacy and evaluated with MovieLens 1M dataset, demonstrating its advantage when security requirement is strong and the number of records is small.

The dataset used in this paper has data categories containing a small number of basic categories, which does not hold generally. Future work thus includes extending the proposed method to cope with different kinds of datasets that have different kinds of categories or no categories.

## REFERENCES

[1] C. Dwork, Differential Privacy, ICALP, Vol. 4052, pp. 1–12 (2006).

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, ”Calibrating Noise to Sensitivity in Private Data Analysis,” TCC, pp. 265–284 (2006).

[3] C. Dwork and A. Roth, The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science Vol. 9, No. 3–4, pp. 211–407 (2014).

[4] K. Nissim, S. Raskhodnikova, and A. Smith, ”Smooth Sensitivity and Sampling in Private Data Analysis,” STOC, pp. 75–84 (2007).

[5] L. Sweeney, ”K-anonymity: A Model for Protecting Privacy,” IJUFKS, Vol. 10, No. 5, pp. 557–570 (2002).

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, ”l-Diversity: Privacy Beyond k-Anonymity,” TKDD, Vol. 1, No. 1 (2007).

[7] L. Ninghui, L. Tiancheng, and V. Suresh, ”t-Closeness; Privacy Beyond k-Anonymity and l-Diversity,” ICDE, pp. 106–115 (2007).

[8] C. Skinner, ”Statistical Disclosure Control for Survey Data,” Handbook of Statistics, Vol. 29, pp. 381–396 (2009).

[9] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf, Statistical Disclosure Control, A John Wiley & Sons (2012).

[10] J. Soria-Comas and J. Domingo-Ferrer, ”Probabilistic k-Anonymity through Microaggregation and Data Swapping,” ICFS, pp. 1–8 (2012).

[11] D. Ikarashi, R. Kikuchi, K. Chida, and K. Takahashi, ”Probabilistic k-Anonymity through Microaggregation and Data Swapping,” IWSEC (2015).

[12] C. C. Aggarwal and P. S. Yu, ”Privacy-Preserving Data Mining Models and Algorithms,” Vol. 34, Springer (2008).

[13] D. Kifer and B. Lin, ”Towards an Axiomatization of Statistical Privacy and Utility,” PODS, pp. 147–158 (2010).

[14] C. Dwork and A. Smith, ”Differential Privacy for Statistics: What we Know and What we Want to Learn,” JPC, Vol. 1, No. 2, pp. 135–154 (2009)

[15] C. Dwork, ”Differential Privacy in New Settings,” SODA, pp. 174-183 (2010).

[16] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, ”Differential Privacy under Continual Observation,” STOC, pp. 715–724 (2010).

[17] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias, ”Differentially Private Event Sequences over Infinite Streams,” VLDB Endowment, pp. 1155–1166 (2014).

[18] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, ”Our Data, Ourselves: Privacy via Distributed Noise Generation,” EUROCRYPT, Vol. 4004, pp. 486-503 (2006).

[19] S. P. Kasiviswanathan and A. D. Smith, ”A Note on Differential Privacy: Defining Resistance to Arbitrary Side Information,” IACR (2008).

[20] N. Li, W. Qardaji, and D. Su, ”On Sampling, Anonymization, and Differential Privacy or, K-anonymization Meets Differential Privacy,” ASIACCS, pp. 32–33 (2012).

[21] G. Cormode, M. Procopiuc, D. Srivastava, and T. T. L. Tran ”Differentially Private Publication of Sparse Data,” ICDT (2011).

[22] N. Shlomo, L. Antal, and M. Elliot, ”Measuring Disclosure Risk and Data Utility for Flexible Table Generators,” JOS, Vol. 31, No. 2, pp. 305-324 (2015).

[23] F. McSherry and I. Mironov, ”Differentially Private Recommender Systems, Building Privacy into the Net,” SIGKDD, pp. 627–636 (2009).

[24] S. Zhou, K. Ligett, and L. Wasserman, ”Differential Privacy with Compression,” ISIT, pp. 2718–2722 (2009).

[25] K. Chaudhuri, A. D. Sarwate, and K. Sinha, ”A Near-Optimal Algorithm for Differentially-Private Principal Components,” JMLR, (2013).

[26] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, ”Analyze Gauss: Optimal Bounds for Privacy-preserving Principal Component Analysis,” STOC, pp. 11–20 (2014).

[27] J. B. Schafer, J. A. Konstan, and J. Riedl, ”E-commerce Recommendation Applications,” DMKD, Vol. 5, pp. 115-153 (2001).

[28] D. K. Agarwal and B.-C. Chen, ”Statistical Methods for Recommender Systems,” Cambridge Univ. Press (2016).

[29] GroupLens, ”MovieLens 1M Dataset,” available from <http://grouplens.org/datasets/movielens/1m> (2003).

[30] Nick Craswell, Precision at n, Encyclopedia of Database Systems, pp. .2127-2128 (2009).

(Received October 20, 2017)
(Revised April 26, 2018)

**Takayasu Yamaguchi** received his B.E. and M.E. degrees from the University of Electro-Communications in 1999 and 2001. He is currently a Senior Research Engineer at Research Laboratories, NTT DOCOMO, Inc. and also a student of doctoral program in Department of Informatics, the University of Electro-Communications. His research interests, these days, lie in the area of privacy preserving data mining, statistical machine learning, and big data applications. He is a member of IPSJ.

**Hiroshi Yoshiura** received his B.Sc. and D.Sc. degrees from the University of Tokyo, Japan in 1981 and 1997. He is currently a Professor in Department of Informatics, the University of Electro-Com-munications. Before joining UEC, he had been at Systems Development Laboratory, Hitachi, Ltd. He has been engaged in research on information security and privacy. He is a member of IEEE, IEICE, IPSJ, JSAI, and JSSM.