Towards Practical *k*-Anonymization: Correlation-based Construction of Generalization Hierarchy

Tomoaki Mimoto, Anirban Basu and Shinsaku Kiyomoto
KDDI R&D Laboratories Inc, 2-1-15 Ohara, Fujimino, Saitama, 356-8502, Japan

Keywords: Privacy, Anonymization, *k*-anonymity, Generalization Hierarchy.

Abstract: The privacy of individuals included in the datasets must be preserved when sensitive datasets are published. Anonymization algorithms such as *k*-anonymization have been proposed in order to reduce the risk of individuals in the dataset being identified. *k*-anonymization is the most common technique of modifying attribute values in a dataset until at least *k* identical records are generated. There are many algorithms that can be used to achieve *k*-anonymity. However, existing algorithms have the problem of information loss due to a tradeoff between data quality and anonymity. In this paper, we propose a novel method of constructing a generalization hierarchy for *k* anonymization algorithms. Our method analyses the correlation between attributes and generates an optimal hierarchy according to the correlation. The effect of the proposed scheme has been verified using the actual data: the average of *k* of the datasets is 83.14, and it is around $1/3$ of the value obtained by conventional methods.

1 INTRODUCTION

When personal data has great potential for building an efficient and sustainable society, there are issues of privacy because it includes a lot of sensitive data such as location information, purchasing history and medical history. Breaches of privacy breach have been a major concern for users of personalized services, both online web services and offline real services. O2O (Online to Offline) is a new direction for commercial services; however, privacy concerns have become greater due to the expansion of service collaborations. Users become very concerned when they are diverted to services they were unaware of having any relationship with. In fact, some research results (Guha et al., 2010; Korolova, 2010) have suggested that Internet ads personalized with private data can leak users' private information.

Anonymization methods (Iwuchukwu and Naughton, 2007; LeFevre et al., 2006; Byun et al., 2007; He et al., 2012; Lin and Wei, 2008) are required in order to preserve privacy as well as to maintain the original utility of the data. This problem related to a trade-off between anonymity and utility has been the main issue anonymization research has had to deal with and various anonymization algorithms have been proposed. However, no universal criterion for anonymization under various circumstances has been provided so far.

Datasets generally consist of records including some attributes that may identify individuals when they are combined with other attributes such as age, sex and address. *k*-Anonymization(Sweeney, 2002b) is a major anonymization approach that use a conversion algorithm from an original dataset to an anonymized dataset that satisfies at least *k* records have identical attributes. There exist some conversion techniques for achieving *k*-anonymity: data addition, noise addition and generalization(Sweeney, 2002a) as well. Generalization is a method which replaces individual attribute values with a broader category (e.g., age:26 → 26-30 or 21-30). Obviously, *k*-anonymity is achieved by strengthening the degree of generalization, even though this conversion leads to information loss.

A generalization hierarchy is a configuration of data used in the generalization process and it defines the broader categories. A method to construct an effective generalization hierarchy under taking correlations between attributes in account has been proposed in this paper. Furthermore, we present a privacy risk assessment tool in order to evaluate the efficiency and *k*-anonymity of a dataset. Our experimental results using an actual health examination shows the efficiency of the proposed method.

The rest of the paper is organized as follows: related work is presented in section 2. Technical terms are defined in section 3. Section 4 introduces the al-

Mimoto, T., Basu, A. and Kiyomoto, S.
Towards Practical k-Anonymization: Correlation-based Construction of Generalization Hierarchy.
DOI: 10.5220/0005963804110418
In *Proceedings of the 13th International Joint Conference on e-Business and Telecommunications (ICETE 2016) - Volume 4: SECRYPT*, pages 411-418
ISBN: 978-989-758-196-0

gorithm we devised implemented as a tool. Furthermore, we clarify one issue of generalization which cannot be solved by existing tools and propose a new concept in order to solve it. Section 5 presents some experimental results using our tool in order to verify the effects of our concept. Finally, we conclude this paper in section 6.

2 RELATED WORK

There are two major approaches to avoiding leaks of private information from public databases: perturbative methods and non-perturbative methods. Deletion of the outlier records or cells and generalization are common non-perturbation methods. As perturbation methods, many techniques also have been proposed. Noise addition, data swapping(Fienberg and McIntyre, 2004) and microaggregation(Willenborg and de Waal, 2001) are widely known. Moreover, there are also many safety indexes of a dataset.

Differential privacy(Dwork, 2006; Dwork, 2008) is a notion of privacy for perturbative methods that is based on the statistical distance between two database tables differing by at most one element. The basic idea is that, regardless of background knowledge, an adversary with access to the dataset draws the same conclusions, whether or not a person's data is included in the dataset. That is, a person's data has an insignificant effect on the processing of a query. Differential privacy is mainly studied in relation to perturbation methods(Winkler, 2004; Dwork et al., 2006a; Dwork et al., 2006b) in an interactive setting, although it is applicable to certain generalization methods(Kiyomoto and Martin, 2010).

Samarati and Sweeney(Samarati and Sweeney, 1998a; Samarati, 2001; Sweeney, 2002a) proposed a primary definition of privacy that is applicable to generalization methods. A dataset is said to have k-anonymity if each record is indistinguishable from at least $k-1$ other records with respect to certain identifying attributes called quasi-identifiers(Dalenius, 1986). Clearly any generalization algorithm that converts a dataset into one with k-anonymity involves a loss of information in that dataset.

Minimizing this information loss thus presents a challenging problem in the design of generalization algorithms. The optimization problem is referred to as the k-anonymity problem. Meyerson reported that optimal generalization in this regard is an NP-hard problem(Meyerson and Williams, 2004). Aggarwal *etal*. proved that finding an optimal table including more than three attributes is NP-hard(Aggarwal et al., 2005). Nonetheless, k-anonymity has been widely studied because of its conceptual simplicity(Al-Fedaghi, 2005; Machanavajjhala et al., 2006; Machanavajjhala et al., 2007; Wong et al., 2006; Truta and Vinay, 2006; Sun et al., 2008).

For example, some techniques based on space division(Iwuchukwu and Naughton, 2007; LeFevre et al., 2006) and on clustering(Byun et al., 2007; He et al., 2012; Lin and Wei, 2008) have been proposed to achieve k-anonymity. In space division techniques, the records are represented as points on a multidimensional space, and the space is divided so that all of the space has more than k records. kd-tree(Freidman et al., 2009) or R-tree(Guttman, 1984) are usually used as the methods to divide the space. These techniques are fast, but the distance between the points is not considered, so it is possible that points some distance away from each other are placed in the same space, and that leads to information loss. In techniques based on clustering, the distance between the records is considered but all groups must include more than k records. Therefore, the cluster area is spread, and there is also the possibility of information loss.

3 PRELIMINARY

In this section, the notations and definitions are introduced.

3.1 Notation

The i-th record represents the t_i and j-th attribute as a_j. Dataset D consists of multiple records $T = \{t_1,...,t_i,...,t_n\}$ and attributes $A = \{a_1,...,a_j,...,a_m\}$. Moreover, the value of a_j of t_i represents $t_i[a_j]$, all values of a_j represent $T[a_j]$ and all values of t_i represent $t_i[A]$. So now we can describe D as follows:

$$\begin{aligned} D &= \{t_1[a_1],...,t_1[a_m],t_2[a_1],...,t_n[a_m]\} \\ &= \{T[a_1],...,T[a_j],...,T[a_m]\} \\ &= \{t_1[A],...,t_i[A],...,t_n[A]\} \end{aligned}$$

When you generalize $t_i[a_j]$ to $t'_i[a_j]$ in the condition of g, we express $t_i[a_j] \xrightarrow{g} t'_i[a_j]$ and the relationship between $t_i[a_j]$ and $t'_i[a_j]$ represents $t_i[a_j] \leq t'_i[a_j]$. When $t'_i[A] = \{t'_i[a_1],...,t'_i[a_l] | \forall t'_i[a_j] : (t'_i[a_j] = t_i[a_j]) \vee (t_i[a_j] \xrightarrow{g} t'_i[a_j])\}$ we express $t_i[A] \leq t'_i[A]$ in a similar manner.

3.2 Definition

The notion of a quasi-identifier is defined as follows.

Definition 1 (Quasi-identifier) (Samarati, 2001; Samarati and Sweeney, 1998b): Let $T[a_1,...,a_m]$ be a table. A quasi-identifier of D is a set of attributes $\{a_1,...,a_l\} \subseteq A$ the release of which must be controlled.

One example of a quasi-identifier is the date of birth. It is not possible to identify a person by date of birth alone but you may be able to do so by combining it with other attributes like sex.

We define *k*-anonymity for quasi-identifiers as follows.

Definition 2 (*k*-anonymity) (Samarati, 2001; Samarati and Sweeney, 1998b): Let $T[a_1,...,a_m]$ be a table and QI be the quasi-identifers associated with it. T is said to satisfy k-anonymity iff for each quasi-identifier $qi \in QI$, each sequence of values in $T[qi]$ appears at least with k occurrences in $T[QI]$.

4 PRIVACY RISK ASSESSMENT TOOL

In this section, we explain the algorithms of the risk assessment tool, which we designed and implemented on the basis of the previous research(Basu et al., 2015).

4.1 Overview of the Tool

This tool makes it possible to evaluate *k*-anonymity of a dataset by simulating an adversary (**Algorithm 3**). In order to simulate an adversary, we have developed an iterative process to evaluate the re-identification probabilities of records based on the knowledge of a subset of the attributes. Data generalization, the deletion of records or cells, and construction of a secure dataset are also performed (**Algorithm 1**). When we use this tool, we input a dataset D, the conditions G, and background attributes of adversary $ATTR$. In algorithm 1, a secure dataset D' is made by D and G. The privacy risk in a k-anonymous dataset D' is proportional to the probability that a record can be uniquely identified in the anonymized dataset when the adversary has knowledge of a certain set of attributes from another dataset considered as background information. This statement can be reduced to the fact that if there is any record in the anonymized dataset that is unique for a certain set of attributes then that record risks being re-identified if the adversary happens to know the values of those attributes. Thus, a data structure of D' is constructed in algorithm 2 and an attack simulation is performed in algorithm 3.

4.2 Basic Function

We explain the details of algorithms of the risk assessment tool.

4.2.1 Construction of the Dataset

We give the dataset D and the conditions of generalization G to `Construction()`. We can delete records or cells, and generalize attributes, and it is possible to establish detailed settings, for instance deleting $\forall t_i[A]$ such that $(t_i[QI_1] = X) \wedge (t_i[QI_2] = Y) \vee (\neg t_i[QI_3] = Z)$.

Algorithm 1: `Construction(`D, G`)`: generalization in attribute values.

```
Require: The $k$-anonymized dataset $D$, the conditions of generalization $G$
for $\forall t_p[A](p = 1,...,n) \in D$ do
    if $t_p[A]$ meets the conditions of $G$ then
        $t_p[A] \xrightarrow{G} t'_p[A]$.
        $D' \leftarrow t'_p[A]$.
    else
        $D' \leftarrow t_p[A]$.
    end if
end for
return $D'$
```

4.2.2 Simulation of Adversary

We supply the dataset D and the attributes $ATTR = \{attr_1,...,attr_l\}$ to be used for assessment to `indexRepeats()`. `indexRepeats()` uses HashMap and preserves the same number of attributes combinations as $t_p \in D$ by using $K_p = t_p[attr_1]|...|t_p[attr_l]$ as a key. We can achieve this function by using a hash table, but we adopt the radix tree because then we do not have to consider collision handling and the time for search and insertion is $O(K_p)$, which is relatively small.

In `attackSimulation()`, we use `indexRepeats(`D,$ATTR$`)` and perform a statistical risk analysis. It is possible to quantitatively determine the risk (=*k*-anonymity) in the actual attack by assuming the attacker's background knowledge.

4.3 Extension

We can construct a secure dataset and evaluate the *k*-anonymity by using this tool. In algorithm 1, however, the condition of G is decided heuristically and it may be not appropriate input. On the other hand, we evaluated several datasets and found that correlations

Algorithm 2: `indexRepeats(`D,$ATTR$`)`: indexing repeats in attribute values.

```
Require: The dataset $D$, the set of attributes $ATTR = \{attr_1, ..., attr_l\}$
Repeat detector $RT \leftarrow$ empty radix tree.
for $\forall t_p[ATTR](p = 1, ..., n) \in D$ do
    $K_p \leftarrow t_p[attr_1]||...||t_p[attr_l]$.
    if $K_p \in RT$ then
        $RT$.put($K_p, 1 + RT$.get($K_p$)).
    else
        $RT$.put($K_p, 1$).
    end if
end for
return $RT$
```

Algorithm 3: `attackSimulation(`D,$ATTR$`)`.

```
Require: The $k$-anonymized dataset $D$, an arbitrary set of attributes $ATTR$
$RT \leftarrow$ indexRepeats($D$,$ATTR$).
for $\forall r_p \in D$ do
    $n \leftarrow RT$. get ($K_p$).
    $\Pr(re\_id|K_p) \leftarrow \frac{1}{n}$.
end for
return Cumulative Distribution of $\Pr(re\_id|K_p)$ for all $ATTR$.
```

existed among the attributes. We propose an extended algorithm which can help to decide G based on this fact.

There is a correlation among the attributes of the actual dataset, such as between age and height. In the case of Table 1, we can use bottom coding and generalize the height under 150 cm. But there is a correlation between height and age, and regarding Table 2, we can use bottom coding and generalize height of the data where age is over 15 years and height is under 160 cm. The height of the data can also be delimited more finely where age is over 15 years and height is around 170 cm. Additionally, we can use top coding for the height of the data where age is under 12 years and height is over 160 cm. We now propose a generalization method taking correlation into account as in this example and assess the effects by experiment.

The method is as follows. First we select an attribute for which we want to build a generalization hierarchy, then check the correlation with other attributes. We can select various methods in `correlation()` algorithm, which is based on correlation coefficient, functional dependency and general statistical information, for example. We next construct a generalized hierarchy of $attr_f$ based on $attr_g$, which has the strongest correlation with $attr_f$. The construction rules are various, and in this paper, we divide the domain as each domain has over 2000 records. After that we adjust the domain size by using `adustment()`. We explain `adustment()` algorithm in the next section with an example. Finally, we delete some records to fulfill the required conditions, $k = 5$ for example, and we can generate a high utility value dataset. Algorithm 4 shows our algorithm and takes as input the dataset D, attributes $ATTR$ and a focused attribute $attr_f$. The algorithm examines the correlation among the attributes and outputs a generalization hierarchy of $attr_f$ and conditions of generalization G. In this way, a more appropriate dataset can be constructed.

Table 1: Height.

height(cm)	population
140	1
145	5
150	13
155	27
160	17
165	23
170	38
175	25

Table 2: Age and height.

height	age		
(cm)	12	15	20
140	0	1	1
145	4	1	0
150	11	0	2
155	23	2	2
160	9	7	1
165	2	9	12
170	1	20	17
175	0	10	15

5 EXPERIMENT

We conducted the following experiment in order to examine the difference between a dataset which is generalized using our method and a dataset which is simply generalized. Our object was to construct an appropriate 2-anonymized dataset in both experiments. We used part of some actual medical examination data in this experiment. The dataset includes $|T| = 20708$ and $|A| = 91$ as data, and we assume $QI = \{birth, sex, height, weight, medicalhistory, smoking, alcoholconsumption\}$. The elements of each attribute are described in Table 3.

Algorithm 4: `Hierarchy Construction`.

```
Require: The $k$-anonymized dataset $D$, an arbitrary set of attributes $ATTR = \{attr_1, ..., attr_l\}$ and a focused attribute $attr_f$.
for $\forall attr_q \in ATTR \setminus \{attr_f\}$ do
    correlation($attr_f, attr_q$): Calculate the correlations between $attr_f$ and $attr_q$.
end for
Construct a generalization hierarchy of $attr_f$ on the basis of $attr_g$ (Let the strongest correlations with $attr_f$ be $attr_g$).
for $\forall attr_q \in ATTR \setminus \{attr_f, attr_g\}$ do
    adjustment(A generalization hierarchy of $attr_f$, the correlations between $attr_f$ and $attr_q$):Adjust the range of $attr_f$.
end for
return A generalization hierarchy of $attr_f$ and the conditions of generalization $G$.
```

Table 3: Data for experiments.

QI	elements
Birth (year)	1940, 1941, ..., 1996
Sex	Male, Female
Height (kg)	128, 129, ..., 196
Weight (cm)	30, 31, ..., 165
Medical history	Yes, No
Smoking	Yes, No
Alcohol Consumption	Often, Sometimes, Seldom

5.1 Evaluation Method

Birth, height and weight are attributes for which we can construct generalization hierarchies, and our approach can be applied to them. In this experiment, we focused on height and examined the correlation between height and other attributes. First, we selected an attribute whose correlation with height was the strongest from the attributes of sex, medical history, smoking and alcohol consumption because they can have only two or three values. We next divided the range of height so that each domain has 2000 records, which is 10% of all records, and then we checked the correlations between height and the others. We decided that the domains of other attributes also have 2000 records. Finally, we checked the correlations between height and birth and between height and weight, and divided the domains. For comparison, we constructed a dataset where the domain of height was divided into three groups, while the domains of birth and weight were divided into two groups. After we constructed these datasets, we used the privacy risk assessment tool and evaluated them.

5.2 Experimental Results

We checked the correlation between height and other attributes, and we decided $attr_g$=sex in figure 1. The figure shows that the deeper the color is, the greater the population is in the domain. In consideration of this correlation, we divided height as follows: if sex is male, -162(cm), 163-164, 165-166, 167, 168, ..., 176, 177-178, 179-, if sex is female, -155(cm), 156-157, 158-160, 161-164, 165-. The dataset we use has a higher proportion of males, and the height of many records is around 170(cm), but when we focus on the correlation between height and medical history in figure 2, medical history is distributed uniformly. So we divided height as follows: if the answer to whether or not there is a medical history is yes, -157(cm), 158-162, 163-166, 167-170, 171-174, 175-, otherwise -155(cm), 156-157, 158-160, 161-162, 163-164, ..., 175-176, 177-179, 180-. We next adjusted the boundary as shown in figure 3. The solid line means the combination of plural domains and the dotted line means the decomposition of the domain. We decided that the final domains (the domains of $sex = male \wedge medicalhistory = yes$) are combined when the total number of domains (the number of domains of $sex = male$ and of $medicalhistory = yes$) is less than 5 and otherwise the final domains can be divided. This means, for example, there are 2000 records of $medicalhistory = yes$ in the domain of $height$ = 171-174 and the proportion of males is high, so we can divide the domain into 171-172 and 173-174 only if the records have sex=$male$ and $medicalhistory$=yes. It may seem that when the number of attributes is large, this step must be repeated again and again, but we divided the domain by referring to the attribute which had the strongest correlation.

We checked the correlation between height and smoking and between height and alcohol consumption, and divided the domains of height in the same way, and constructed the generalization hierarchy for

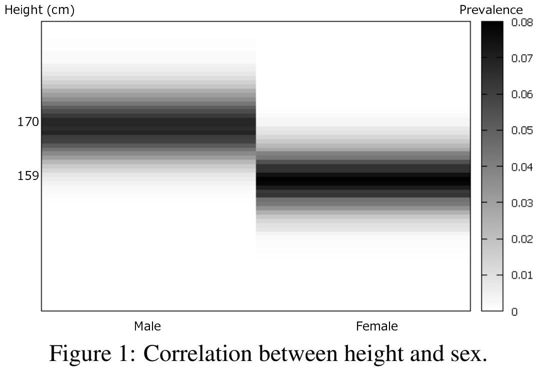

Figure 1: Correlation between height and sex.

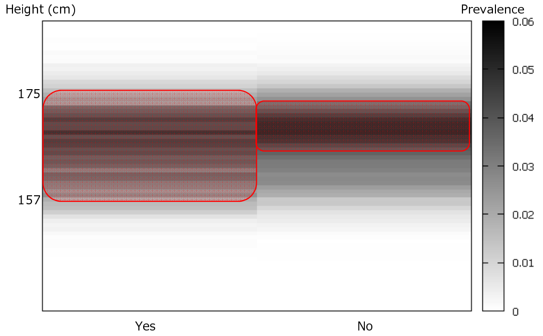

Figure 2: Correlation between height and medical history.

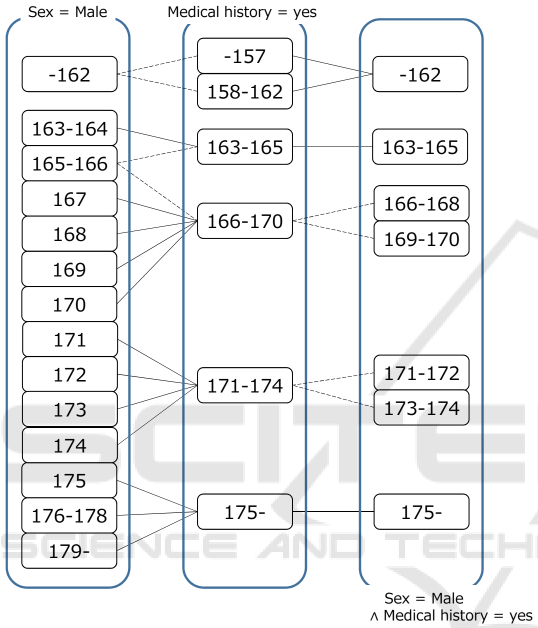

Figure 3: Construction of hierarchy.

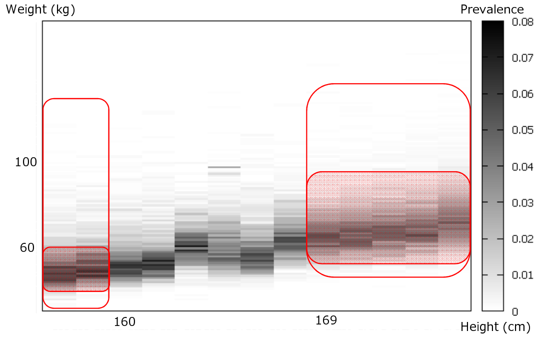

Figure 4: The correlation between height and weight.

height. When constructing the hierarchy, care must be taken so as not to generate conflict. When the domain 100-110 and 111-120 is created, the parent must be 100-120, for example. We next checked the correlation between height and weight. We can say from the figure 4 that the weights of tall people tend to be distributed uniformly, while on the other hand, the weight range for short people is wide but most of their weights are almost the same. So we weakened from -155(cm), 156-157 to -157 in order to increase the parameter. Finally, we divided the birth value simply into three parts because there is no relationship between height and birth. The generalization hierarchy that we finally constructed is shown in figure 5. This is the output of algorithm 4.

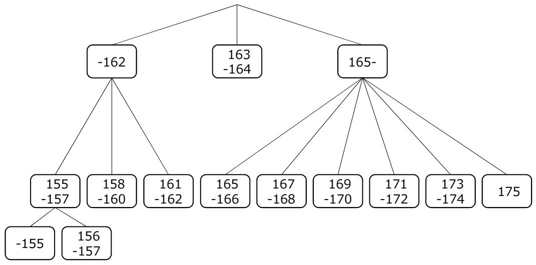

Figure 5: Generalization hierarchy.

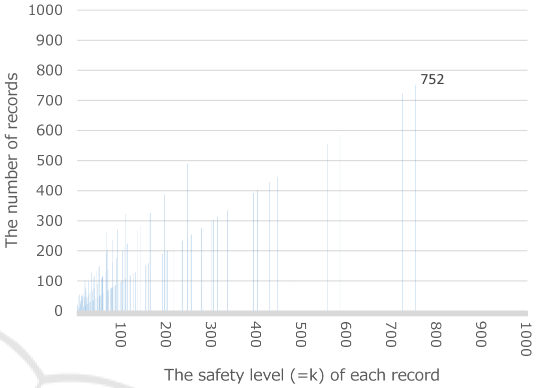

Figure 6: The result using the original method.

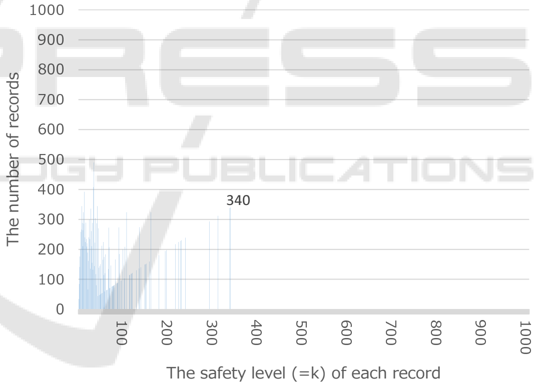

Figure 7: The result using our method.

The assessment results are below. Figure 6 is the result without considering correlation and figure 7 is the result where correlation is taken into account.

The former result shows that the risk values for many records are $k > 200$ and the maximum is $k = 752$, and this means the amount of information loss is too large. Notice that the purpose of this experiment is to construct a 2-anonymized dataset. On the other hand, the latter result shows that the risk values for many records are $k < 100$ and the maximum is $k = 340$. This means that we have constructed a dataset the risks of which records are distributed uniformly. Additionally, we decided on a focus attribute (height in this experiment) first, and we can analyze this attribute specifically.

6 CONCLUSION

In this paper, we proposed a method for constructing a generalization hierarchy based on an analysis of correlations between attribute values and analyzed the effect of the method using an actual medical examination dataset. We conclude that our method is an effective way to generate more practical *k* anonymized datasets.

ACKNOWLEDGEMENTS

This work was supported by CREST, JST.

REFERENCES

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. (2005). Anonymizing tables. In *Proc. of ICDT 2005, LNCS*, volume 3363, pages 246–258.

Al-Fedaghi, S. S. (2005). Balanced *k*-anonymity. In *Proc. of WASET*, volume 6, pages 179–182.

Basu, A., Nakamura, T., Hidano, S., and Kiyomoto, S. (2015). k-anonymity: risks and the reality, accepted for publication. In *IEEE International Symposium on Recent Advances of Trust, Security and Privacy in Computing and Communications (RATSP, collocated with the IEEE TrustCom)*.

Byun, J.-W., Kamra, A., Bertino, E., and Li, N. (2007). Efficient k-anonymity using clustering technique. In *Proc. of the International Conference on Database Systems for Advanced Applications*, pages 188–200.

Dalenius, T. (1986). Finding a needle in a haystack —or identifying anonymous census record. In *Journal of Official Statistics*, volume 2(3), pages 329–336.

Dwork, C. (2006). Differential privacy. In *Proc. of ICALP 2006*, volume 4052, pages 1–12.

Dwork, C. (2008). Differential privacy: A survey of results. In *Proc. of TAMC 2008*, volume 4978, pages 1–19.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *Proc. of Eurocrypt 2006, LNCS*, volume 4004, pages 486–503.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Proc. of TCC 2006, LNCS*, volume 3876, pages 265–284.

Fienberg, S. E. and McIntyre, J. (2004). Data swapping: Variations on a theme by dalenius and reiss. In *Proc. of PSD 2004, LNCS*, volume 3050, pages 14–29.

Freidman, J. H., Bentley, J. L., and Finkel, R. A. (2009). An algorithm for finding best matches in logarithmic expected time. In *ACM Transactions on Mathematical Software*, volume 16 (5), pages 670–682.

Guha, S., Cheng, B., and Francis, P. (2010). Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, IMC '10, pages 81–87.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, volume 14, page 47.

He, X., Chen, H., Chen, Y., Dong, Y., Wang, P., and Huang, Z. (2012). Clustering-based k-anonymity. In *Advances in Knowledge Discovery and Data Mining SE*, volume 7301, pages 405–417. Springer-Verlag.

Iwuchukwu, T. and Naughton, J. F. (2007). K-anonymization as spatial indexing: Toward scarable and incremental anonymization. In *Proceeding of the 33rd International Conference on Very Large Data Bases, VLDB*, pages 746–757.

Kiyomoto, S. and Martin, K. M. (2010). Towards a common notion of privacy leakage on public database. In *Proc. of BWCCA 2010, to appear*, pages 186–191. IEEE.

Korolova, A. (2010). Privacy violations using microtargeted ads: A case study. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 474–482.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *Proc. of the 22nd International Conference on Data Engineering (ICDE '06)*, pages 25–35. IEEE.

Lin, J.-L. and Wei, M.-C. (2008). An efficient clustering method for k-anonymization. In *Proc. of the 2008 international workshop on Privacy and anonymity in information society (PAIS '08)*, pages 46–50. ACM.

Machanavajjhala, A., Gehrke, J., and Kifer, D. (2006). *l*-diversity: Privacy beyond *k*-anonymity. In *Proc. of ICDE'06*, pages 24–35.

Machanavajjhala, A., Gehrke, J., and Kifer, D. (2007). *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In *Proc. of ICDE'07*, pages 106–115.

Meyerson, A. and Williams, R. (2004). On the complexity of optimal *k*-anonymity. In *Proc. of PODS 2004*, pages 223–228.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Trans. on Knowledge and Data Engineering*, 13(6):1010–1027.

Samarati, P. and Sweeney, L. (1998a). Generalizing data to provide anonymity when disclosing information. In *Proc. of PODS 1998*, page 188.

Samarati, P. and Sweeney, L. (1998b). *Protecting privacy when disclosing information: k-anonymity and* *its enforcement through generalization and suppression*. Technical Report SRI-CSL-98-04, SRI Computer Science Lab.

Sun, X., Wang, H., Li, J., Truta, T. M., and Li, P. (2008). (p^+, α)-sensitive *k*-anonymity: a new enhanced privacy protection model. In *Proc. of CIT'08*, pages 59–64.

Sweeney, L. (2002a). Achieving *k*-anonymity privacy protection using generalization and suppression. In *J. Uncertainty, Fuzziness, and Knowledge-Base Systems*, volume 10(5), pages 571–588.

Sweeney, L. (2002b). k-anonymity: a model for protecting privacy. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, volume 10 (5), pages 557–570.

Truta, T. M. and Vinay, B. (2006). Privacy protection: *p*-sensitive *k*-anonymity property. In *Proc. of ICDE'06*, pages 94–103.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, volume 155. LNS, Springer-Verlag.

Winkler, W. E. (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. In *Proc. of PSD 2004, LNCS*, volume 3050, pages 231–246.

Wong, R. C.-W., Li, J., Fu, A. W.-C., and Wang, K. (2006). (α, k)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing. In *Proc. of ACM SIGKDD'06*, pages 754–759.