Advancements in Privacy Preserving Data Mining

Lecture 11

Professor Ljiljana Brankovic
The University of Newcastle

Lecture Overview

- Privacy quantification
- Anonymisation and randomization methods
- Attacks on randomization

Dynamic Programming Algorithm for Calculating Confidential Attribute Equivocation H(ε)

Input: x[], p[], ε

Output: H(ε)

if ε==0

$$H(0) = \sum_{i=1}^{n} p(x_i) \log_2 \frac{1}{p(x_i)}$$

else

$$H(\varepsilon,0) = 0;$$

$$H(\varepsilon,1) = p_1 \lg \frac{1}{p_1}$$

Dynamic Programming Algorithm for Calculating Confidential Attribute Equivocation H(ε)

for i=2 to n

p=p_i

j=i-1

$$H(\varepsilon,i) = H(\varepsilon,i-1) + p_i \lg \frac{1}{p_i}$$

while ($x_i - x_j$ <= ε and j≠o) do

p=p+p_j

if H(ε, j-1) + p log 1/p < H(ε,j) then

H(ε,j) = H(ε, j-1) + p log 1/p

j=j-1

end

end

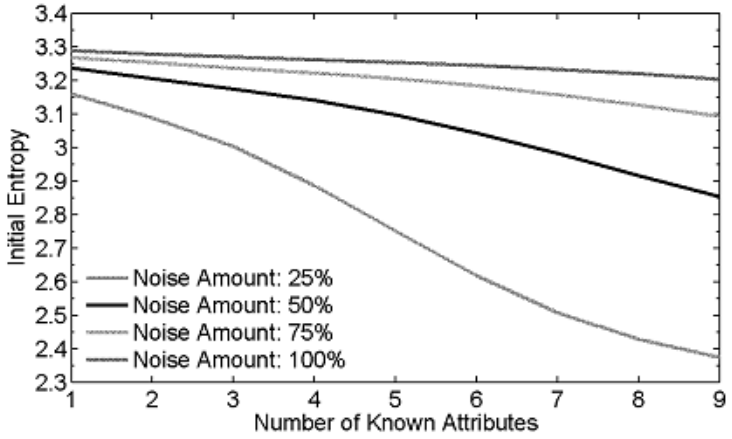
H(ε)=H(ε,n)

Comparative Study

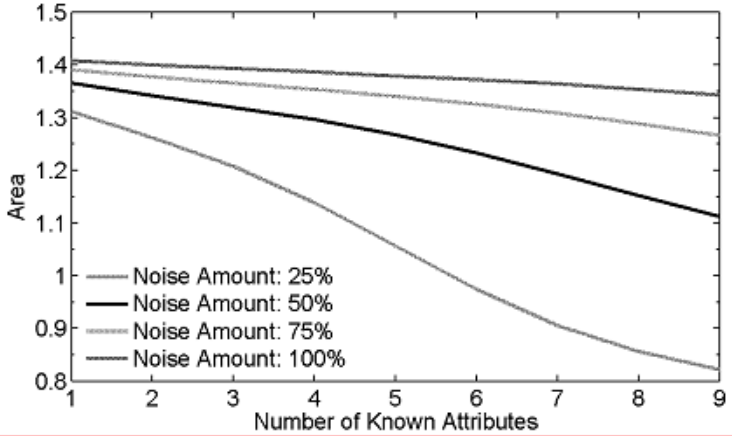
We use our privacy measure to compare the following disclosure control techniques:

- Noise Addition
- Query Restriction
- Sampling

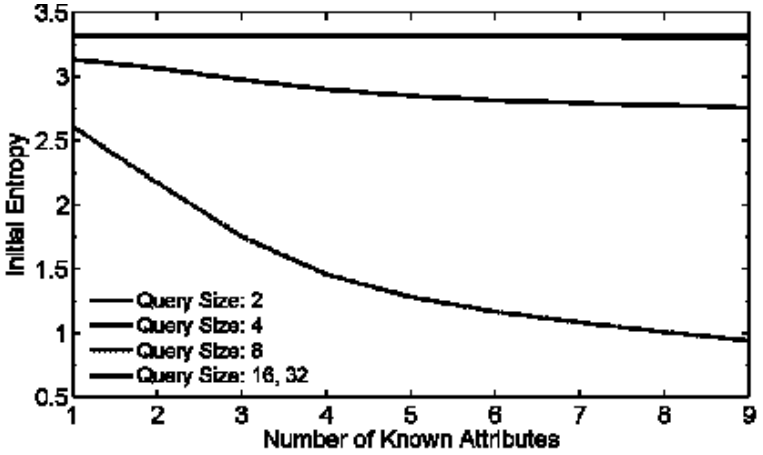
Noise Addition



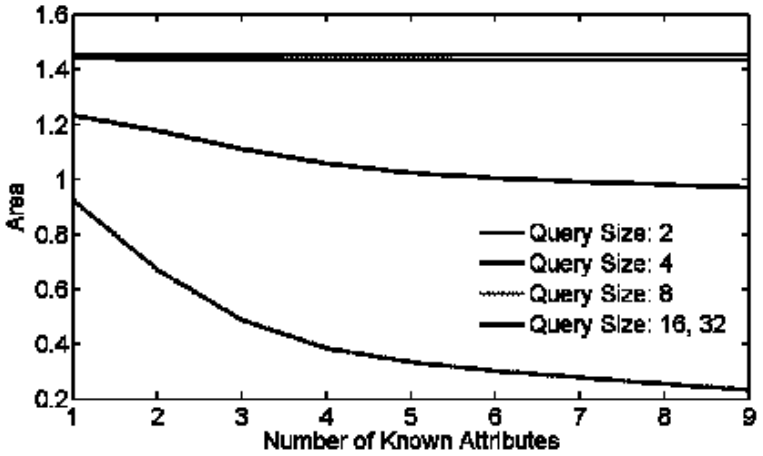
(a) H_0 vs Number of Known Attributes



Query Restriction

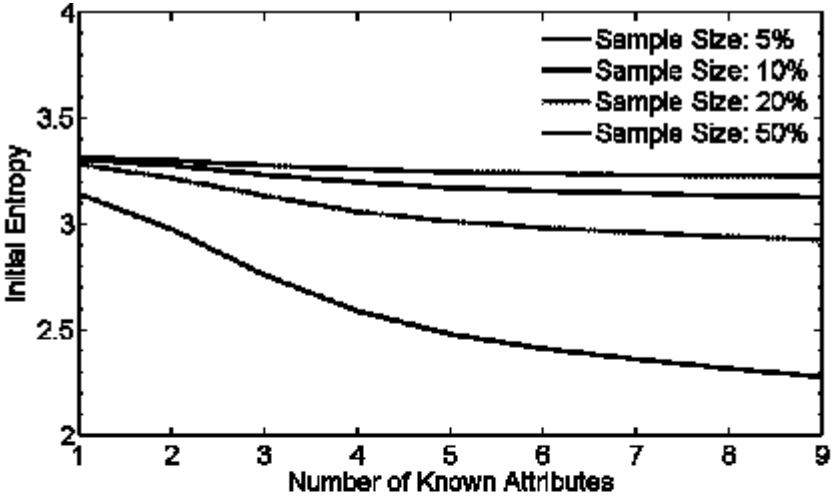


(a) $H_{\parallel}$ vs Number of Known Attributes

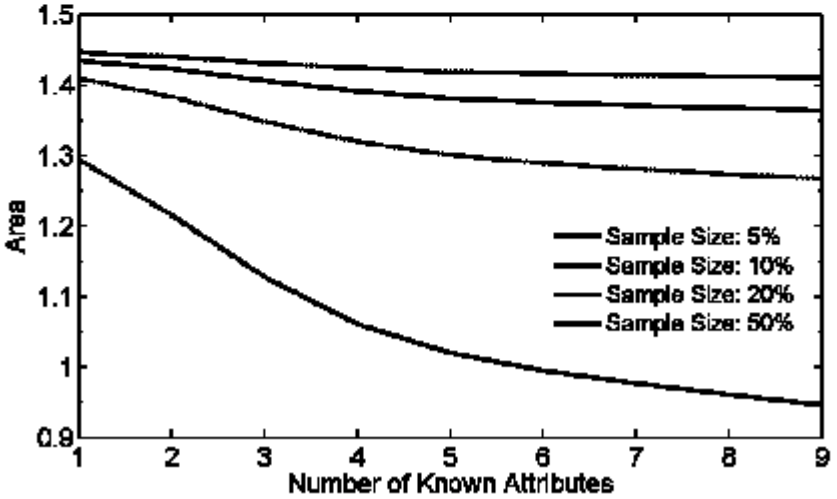


(b) Area vs Number of Known Attributes

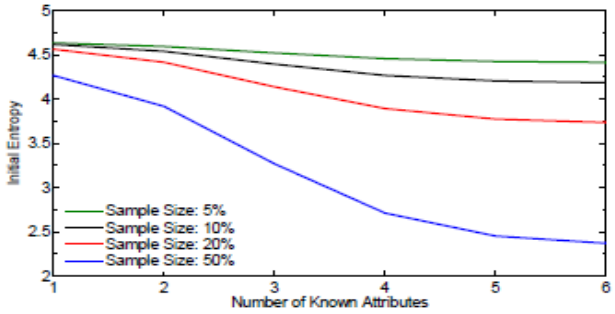
Sampling



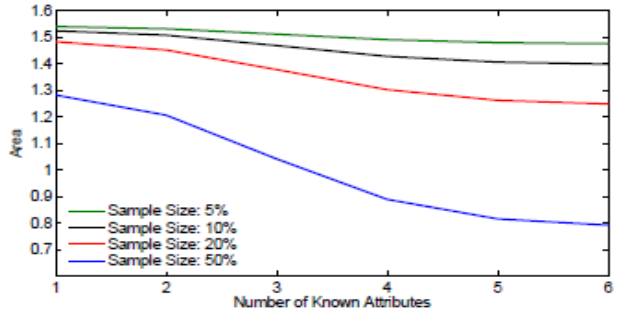
(a) H_0 vs Number of Known Attributes



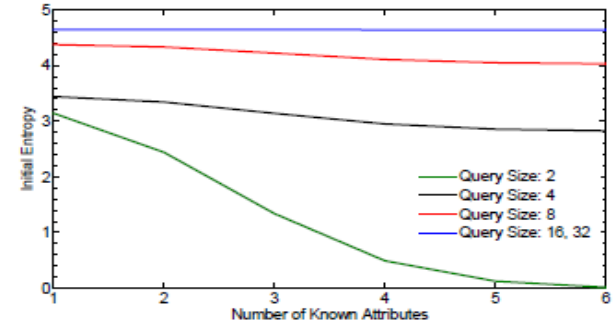
(b) Area vs Number of Known Attributes



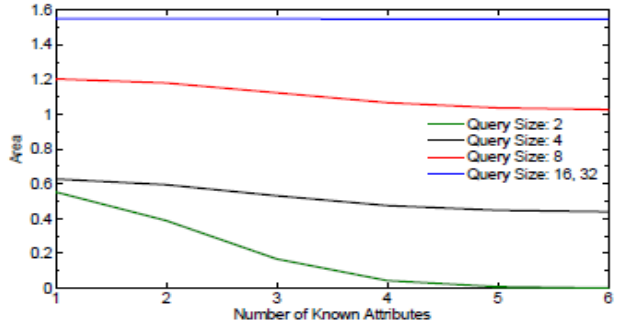
(a) Sampling: H_0 vs SK



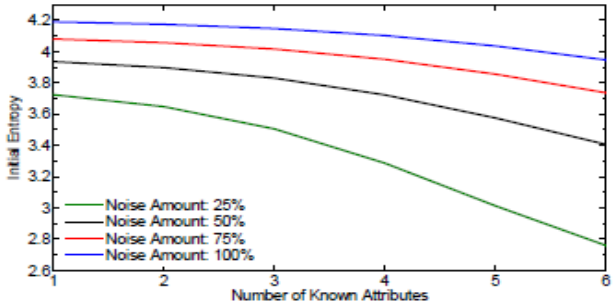
(b) Sampling: Area vs SK



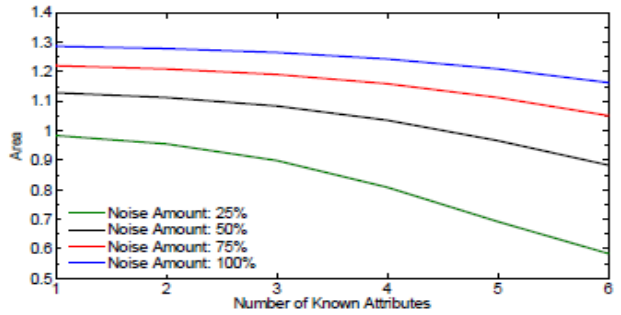
(c) QueryRestr.: H_0 vs SK



(d) QueryRestr.:Area vs SK



(e) NoiseAdd.: H_0 vs SK



(f) NoiseAdd.: Area vs SK

Data anonymisation

The raw data table typically does not satisfy specified privacy requirements and the table must be modified before being published.

The modification is done by applying a sequence of anonymization operations to the table.

Anonymization operations:

1. Generalization
2. Suppression
3. Anatomization
4. Permutation
5. Perturbation.

Data anonymisation

Generalization and suppression replace values of specific description, typically the QID attributes, with less specific description.

Anatomization and permutation de-associate the correlation between QID and sensitive attributes by grouping and shuffling sensitive values in a qid group.

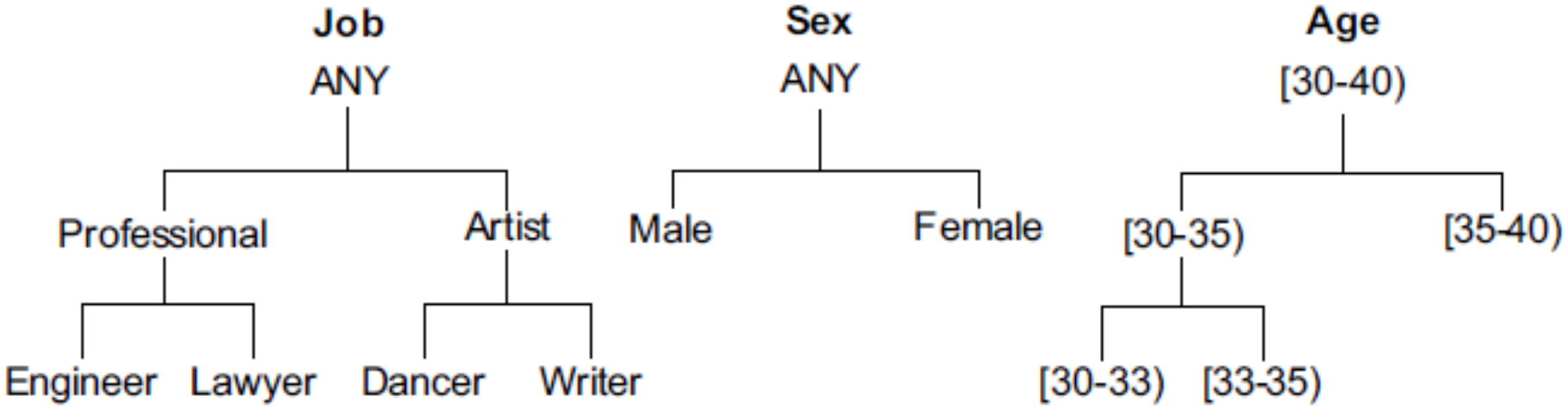
Perturbation distorts the data by adding noise, aggregating values, swapping values, or generating synthetic data based on some statistical properties of the original data.

Generalisation and Suppression

Each generalization or suppression operation hides some details in QID.

For a categorical attribute, a specific value can be replaced with a general value according to a given taxonomy.

In Figure below the parent node Professional is more general than the child nodes Engineer and Lawyer.



The root node, ANY Job, represents the most general value in Job.

Generalisation and Suppression

For a numerical attribute, exact values can be replaced with an interval that covers exact values.

If a taxonomy of intervals is given, the situation is similar to categorical attributes.

More often, however, no pre-determined taxonomy is given for a numerical attribute.

Generalisation and Suppression

Different classes of anonymization operations have different implications on privacy protection, data utility, and search space. However, they all result in a less precise but consistent representation of original data.

A **generalization** replaces some values with a parent value in the taxonomy of an attribute. The reverse operation of generalization is called **specialization**.

A **suppression** replaces some values with a special value, indicating that the replaced values are not disclosed. The reverse operation of suppression is called **disclosure**.

Generalisation Methods

We will consider the following 5 generalisation methods:

1. Full-domain generalization scheme
   [LeFevre et al., 2005; Samarati, 2001; Sweeney, 2002]

2. Subtree generalization scheme
   [Bayardo et al., 2005; Fung et al., 2005; Fung et al., 2007; Iyengar, 2002; LeFevre et al., 2005]

3. Sibling generalization scheme
   [LeFevre et al., 2005;].

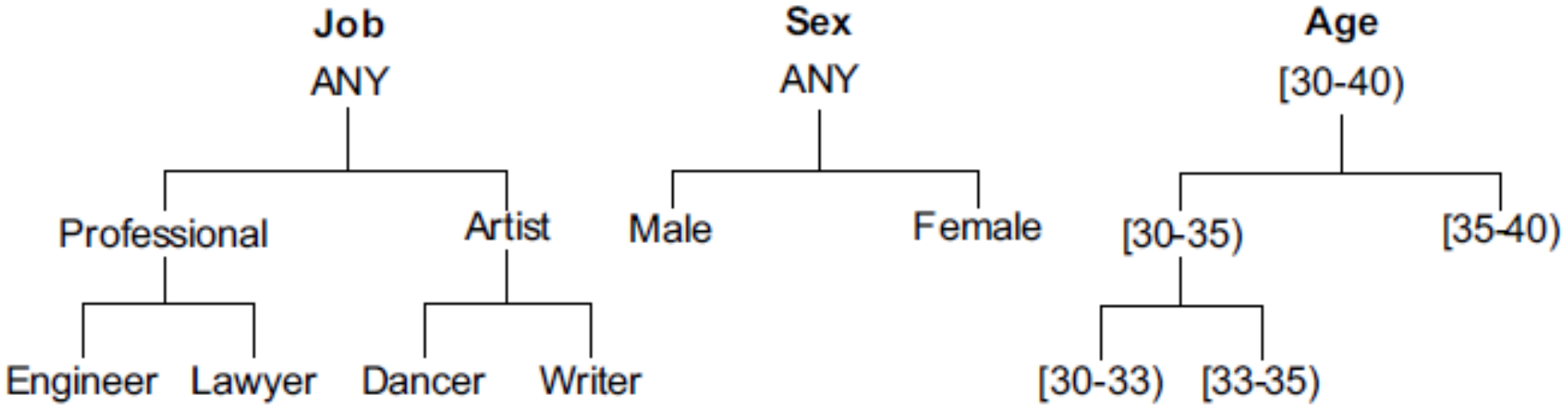
4. Cell generalization scheme
   [LeFevre et al., 2005; Wong et al., 2006; Xu et al, 2006]

5. Multidimensional generalization
   [LeFevre et al., 2006; LeFevre et al., Aug 2006]

Full-domain generalization

All values in an attribute are generalized to the same level of the taxonomy tree.

Example. In the figure below, if Lawyer and Engineer are generalized to Professional, then it also requires generalizing Dancer and Writer to Artist.

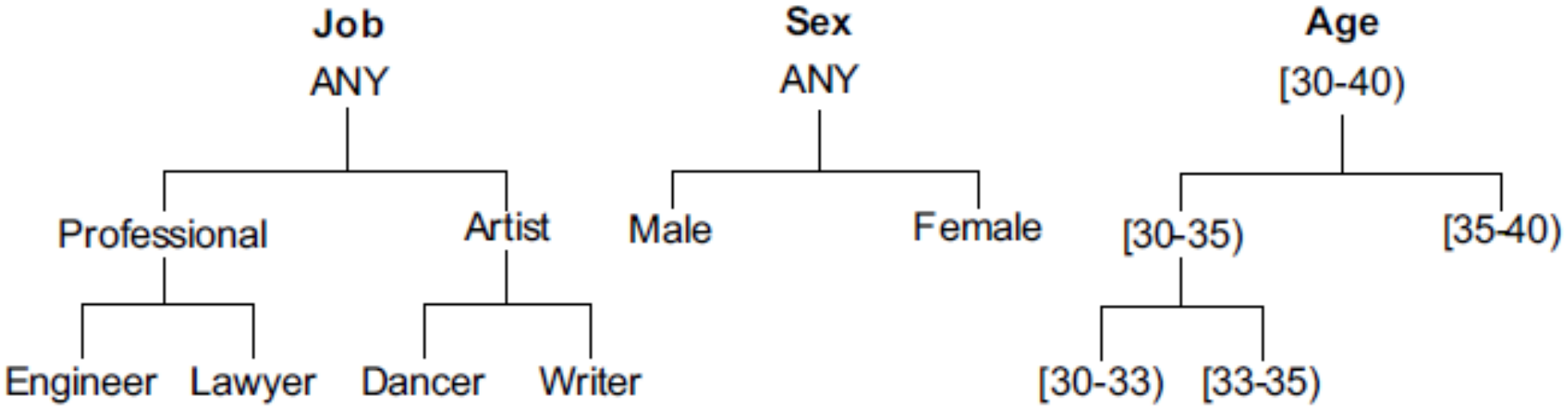


The search space for this scheme is much smaller than the search space for other schemes below, but the data distortion is the largest because of the same granularity level requirement on all paths of a taxonomy tree.

Subtree generalization

At a non-leaf node, either all child values or none are generalized.

Example. In the figure below, if Engineer is generalized to Professional, this scheme also requires the other child node, Lawyer, to be generalized to Professional, but Dancer and Writer, which are child nodes of Artist, can remain ungeneralized.

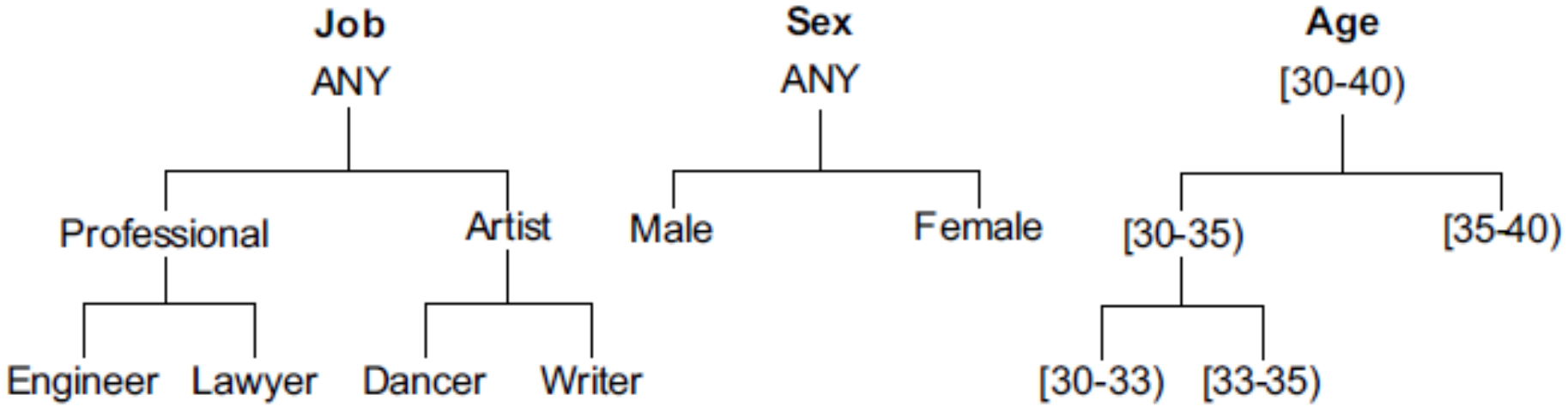


Intuitively, a generalized attribute has values that form a "cut" through its taxonomy tree, where a cut of a tree is a subset of values in the tree that contains exactly one value on each root-to-leaf path.

Sibling generalization

This scheme is similar to the subtree generalization, except that some siblings may remain ungeneralized. A parent value is then interpreted as representing all missing child values.

Example. In the figure below, if Engineer is generalized to Professional, and Lawyer remains ungeneralized, Professional is interpreted as all jobs covered by Professional except for Lawyer.



This scheme produces less distortion than subtree generalization schemes because it only needs to generalize the child nodes that violate the specified threshold.

Cell generalization

In all of the previous schemes, if a value is generalized, all its instances are generalized. Such schemes are called global recoding. In cell generalization, also known as local recoding, some instances of a value may remain ungeneralized while other instances are generalized.

Example. In Table 2.2. below, the Engineer in the first record is generalized to Professional, while the Engineer in the second record can remain ungeneralized.

Table 2.2: Original patient data

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

Cell generalization

Compared with global recoding schemes, this scheme is more flexible; therefore, it produces a smaller data distortion.

Nonetheless, it is important to note that the utility of data is adversely affected by this flexibility, which causes a data exploration problem: most standard data mining methods treat Engineer and Professional as two independent values, but, in fact, they are not.

For example, building a decision tree from such a generalized table may result in two branches, Professional → class2 and Engineer → class1. It is unclear which branch should be used to classify a new engineer.

Though very important, this aspect of data utility has been ignored by all works that employed the local recoding scheme.

Data produced by global recoding does not suffer from this data exploration problem.

Multidimensional generalization

Let D_i be the domain of an attribute A_i.

A single-dimensional generalization, such as full-domain generalization and subtree generalization, is defined by a function

$$f_i: D_{A_i} \rightarrow D'$$

for each attribute A_i in QID.

In contrast, a multidimensional generalization is defined by a single function

$$f_i: D_{A_1} \times \ldots \times D_{A_n} \rightarrow D'$$

which is used to generalize qid = < $v_1,\ldots,v_n$ > to qid' = < $u_1,\ldots,u_n$ > where for every v_i, either $v_i = u_i$ or v_i is a child node of u_i in the taxonomy of A_i.

Multidimensional generalization

This scheme flexibly allows two qid groups, even having the same value on some v_i and u_i, to be independently generalized into different parent groups.

Example.

< Engineer, Male > can be generalized to < Engineer, ANY_Sex> while

< Engineer, Female > can be generalized to < Professional, Female >.

The generalized table contains both Engineer and Professional.

Multidimensional generalization

This scheme produces less distortion than the full-domain and subtree generalization schemes because it needs to generalize only the qid groups that violate the specified threshold.

Note that in this multidimensional scheme all records in a qid are generalized to the same qid', but cell generalization does not have such constraint.

Both schemes suffer from the data exploration problem discussed previously.

Ercan Nergiz et al., 2007, further evaluate a family of clustering-based algorithms that even attempts to improve data utility by ignoring the restrictions of the given taxonomies.

Suppression Methods

There are also different suppression schemes.

1. Record suppression refers to suppressing an entire record.
   [Bayardo et al., 2005; Iyengar, 2002; LeFevre et al., 2005; Samarati, 2001]

2. Value suppression refers to suppressing every instance of a given value in a table.
   [Wang et al., 2007]

3. Cell suppression (or local suppression) refers to suppressing some instances of a given value in a table.
   [Cox, 1980; Meyerson et al, 2004]

Generalization and suppression

In summary, the choice of generalization and suppression operations has an implication on the search space of anonymous tables and data distortion.

The full-domain generalization has the smallest search space but the largest distortion, and the local recoding scheme has the largest search space but the least distortion.

Generalization and suppression

For a categorical attribute with a taxonomy tree H, the number of possible cuts in subtree generalization, denoted by C(H), is equal to

$$C(H_1) \times \cdots \times C(H_u)+1$$

Where H_1, ... ,H_u are the subtrees rooted at the children of the root of H, and 1 is for the trivial cut at the root of H.

The number of potential modified tables is equal to the product of such numbers for all the attributes in QID. T

he corresponding number is much larger if a local recoding scheme is adopted because any subset of values can be generalized while the rest remains ungeneralized for each attribute in QID.

Generalization and suppression

A table is ***minimally anonymous*** if it satisfies the given privacy requirement and its sequence of anonymization operations cannot be reduced without violating the requirement.

A table is ***optimally anonymous*** if it satisfies the given privacy requirement and contains most information according to the chosen information metric among all satisfying tables.

Generalization and suppression

Various works have shown that finding the optimal anonymization is NP-hard:

- Samarati, 2001, shows that the optimal k-anonymity by full-domain generalization is very costly.
- Meyerson et al, 2004, and Aggarwal et al., 2005, prove that the optimal k-anonymity by cell suppression, value suppression, and cell generalization is NP-hard.
- Wong et al., 2006, prove that the optimal (a,k)-anonymity by cell generalization is NP-hard.

In most cases, finding a minimally anonymous table is a reasonable solution and can be done efficiently.

Generalization and suppression

Various works have shown that finding the optimal anonymization is NP-hard:

- Samarati, 2001, shows that the optimal k-anonymity by full-domain generalization is very costly.
- Meyerson et al, 2004, and Aggarwal et al., 2005, prove that the optimal k-anonymity by cell suppression, value suppression, and cell generalization is NP-hard.
- Wong et al., 2006, prove that the optimal (a,k)-anonymity by cell generalization is NP-hard.

In most cases, finding a minimally anonymous table is a reasonable solution and can be done efficiently.

Attacks on Randomisation

A reconstructed distribution on the data can be used in order to reduce the privacy of the underlying data record.

The broad idea is that the correlation structure in the original data can be estimated fairly accurately (in larger data sets) even after noise addition.

Once the broad correlation structure in the data has been determined, one can then try to remove the noise in the data in such a way that it fits the aggregate correlation structure of the data. It has been shown that such techniques can reduce the privacy of the perturbation process significantly since the noise removal results in values which are fairly close to their original values.

Attacks on Randomisation

A second kind of adversarial attack is with the use of public information.

Consider a record X =(x1 ...xd), which is perturbed to Z =(z1 ...zd).

Then, since the distribution of the perturbations is known, we can try to use a maximum likelihood fit of the potential perturbation of Z to a public record.

References

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In Proc. of the 10th International Conference on Database Theory (ICDT), pages 246–258, Edinburgh, UK, January 2005.

[2] M. Alfaleyleh and L. Brankovic. "Quantifying Privacy: A Novel Entropy-Based Measure of Disclosure Risk",

[3] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), pages 217–228, Tokyo, Japan, 2005.

[4] L. H. Cox. Suppression methodology and statistical disclosure control. Journal of the American Statistical Association, 75(370):377–385, June 1980.

References

[5] M. Ercan Nergiz and C. Clifton. Thoughts on k-anonymization. Data & Knowledge Engineering, 63(3):622–645, December 2007.

[6] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), pages 205–216, Tokyo, Japan, April 2005.

[7] B. C. M. Fung, KeWang, and P. S. Yu. Anonymizing classification data for privacy preservation. IEEE Transactions on Knowledge and Data Engineering (TKDE), 19(5):711–725, May 2007.

[8] Fung, K. Wang, A. W.-C. Fu and P. S. Yu, *Introduction to Privacy-Preserving Data Publishing - Concepts and Techniques*, CRC Press, Tylor & Francis Group, 2011.

References

[9] V. S. Iyengar. Transforming data to satisfy privacy constraints. In Proc. of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 279–288, Edmonton, AB, Canada, July 2002.

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In Proc. of ACM International Conference on Management of Data (SIGMOD), pages 49–60, Baltimore, ML, 2005.

[11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE), Atlanta, GA, 2006.

[12] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In Proc. of the 23rd ACM SIGMOD-SIGACT-SIGARTPODS, pages 223–228, Paris, France, 2004.

References

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

[14] L. Sweeney. k-Anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, **10**(5):557–570, 2002.

[15] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu. Utility based anonymization using local recoding. In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, August 2006.

[16] K. Wang, B. C. M. Fung, and P. S. Yu. Handicapping attacker's confidence: An alternative to k-anonymization. Knowledge and Information Systems (KAIS), 11(3):345–368, April 2007.

References

[17] R. C. W. Wong, J. Li., A. W. C. Fu, and K. Wang. (a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 754–759, Philadelphia, PA, 2006.

[18] C.C. Aggarwal and P.S. Yu. A general survey of privacy-preserving Data Mining Models And Algorithms. *Privacy-preserving Data Mining: Models And Algorithms.*