On Enhancing Utility in *k*-Anonymization

Md Nurul Huda, Shigeki Yamada, and Noboru Sonehara

***Abstract*—*k*-anonymity is one of the most studied models of privacy preserving technology. It limits the linking confidence between specific sensitive information and a specific individual by hiding the identifications of each individual into at least *k*-1 others in the database. A *k*-anonymization algorithm is usually evaluated using information loss or data utility metrics. In this paper, we first propose a new quality metric, called the Efficiency metric. This metric overcomes the limitations of existing one dimensional metrics, representing either privacy measure or data utility measure, used in privacy preserving data sharing. We then present a new heuristic algorithm for *k*-anonymization that offers high data utility as well as a high level of privacy. Comparisons of experimental results of our algorithm with those of three other well-known algorithms for *k*-anonymity show that our algorithm performs the best both in terms of utility measure and privacy measure.**

***Index Terms*—Anonymization, information loss, privacy, utility.**

I. Introduction

Many service providers (such as hospitals) collect a large amount of data on individuals as a process of normal operations. In these data, there may exist correlations between attributes of quasi-identifiers (QID) (e.g., age, ZIP, Sex) and attributes of sensitive data (e.g., disease name). For example, a particular age-group might have the tendency of a certain disease. Finding such correlations between attributes is of particular interests to third parties such as researchers and data miners. However, sharing the collected data with third parties raises privacy concern even if the identifying attributes (e.g., name, social security number) are removed before data-sharing. Sensitive personal information may still be linked with individuals by using QIDs. Attributes of QIDs can be joined with external data sources having identifiers (e.g., public voter list) and hence sensitive information can be linked to individuals [1].

Privacy-preserving data mining [2] has been proposed as a paradigm of exercising data mining while protecting the privacy of individuals. One of many proposed approaches is *k*-anonymization, proposed in [1]. One method of *k*-anonymization suggests to modify the values in the quasi-identifier attributes by means of generalization so that if the data table is projected onto the subset of the quasi-identifier attributes, each record of the table becomes indistinguishable from at least (*k*-1) other records. Consequently, sensitive information in the released data cannot be linked to any specific individual with a confidence value of more than 1/*k*, hence the privacy of the individual is protected to some extent.

Due to the modification from a precise value in the original dataset to a more general value in the anonymous dataset, some information loss occurs in the generalization process. This reduces the utility of the anonymous data. The main challenge in *k*-anonymization is to achieve maximal utility (i.e., minimal loss of information). The problem of finding a *k*-anonymization with maximal utility was shown to be NP-hard [3-5]. Hence, the possible approaches are either heuristic algorithms [6-9], [10] or approximation algorithms with a guaranteed approximation factor [3-5]. Usually, algorithms of the former type outperform algorithms of the latter type.

Existing metrics for assessing quality of anonymization express either data utility (also sometimes expressed as information loss (IL) or cost or penalty) or data privacy [1], [5], [7], [9-12]. None represent the evaluation from both the data providers' and data users' point of view. Though, a lower limit of privacy is implied by "*k*" and data quality is expressed by the cost or information loss metric, those metrics do not represent a view of both the parties. Thus, an appropriate metric representing both is necessary.

We first propose the *Efficiency* metric for assessing the anonymous data from both the data provider's perspective and the data user's perspective. Then, we introduce a local-recoding based *k*-anonymization algorithm, called the *LowCost* algorithm that shows the best performance compared to existing well-known *k*-anonymization algorithms. The Efficiency metric overcomes the limitations of existing one dimensional metrics that represent either privacy or data utility measurement of the anonymous data. The LowCost algorithm extracts groups of records into an anonymous table based on the cardinalities of the attributes and cost of generalization of the groups. Extraction of records continues until the records remaining in the original table counts less than *k*, which are then suppressed. Experimental comparisons with existing similar algorithms show that the LowCost algorithm clearly outperforms others.

II. Terminologies

This section defines the data model and terminologies. Consider an initial data table (*T*) that holds information on individuals in *w* attributes $\{A_1,....,A_w\}$.

Quasi-Identifier: A quasi-identifier is a minimal set of attributes, QID = $\{A_1,....,A_d\} \subseteq \{A_1,....,A_w\}$ in the table *T* that can be joined with external information to re-identify individual records[1].

Let each individual be described by a quasi-identifier, and a sensitive attribute, A_{d+1}. Each of the attributes consists of several possible values: $A_j = \{a_{j,l} : 1 \le l \le m_j\}$, $1 \le j \le d+1$. The j^{th}

Manuscript received May 28, 2012; revised June 30, 2012.

M. N. Huda is with the Research Organization of Information and Systems (ROIS), Tokyo, Japan (e-mail: mnhuda@yahoo.com).

S. Yamada and N. Sonehara are with the National Institute of Informatics (NII), Tokyo, Japan (e-mail: shigeki@nii.ac.jp, sonehara@nii.ac.jp).

component of the record R_i (namely, the $(i,j)^{th}$ entry in the table T) will be denoted hereinafter as $A_{i,j}$.

Equivalence Class: An equivalence class E for T with respect to attributes $A_1,\ldots\ldots,A_d$ is the set of all records in T containing identical values $(a_1,\ldots\ldots,a_d)$ for $A_1,\ldots\ldots,A_d$.

k-anonymity: A table T with a quasi-identifier (QID) conforms to the k-anonymity, if and only if each unique record in the projection of T on QID occurs at least k times [1].

III. Related Work

In order to evaluate a k-anonymization algorithm, [5] uses a count of the number of suppressed entries in the generalized table as the generalization cost. Reference [1] assumes a monotone sequence of $h+1$ clusters, where the finest and the coarsest clustering in the sequence are the trivial ones. If a table entry is replaced with a subset from the i^{th} finest cluster that contains it (where $0\leq i\leq h$) the corresponding loss of information is i/h. The Loss Metric (LM) [9] calculates cost per entry depending upon the size of the generalized subset and $0\leq LM\leq 1$. The overall cost is the average cost per table element (cell):

$$LM(T,(T')) = \frac{1}{nd}.\sum_{i=1}^{i=n}\sum_{j=1}^{j=d}\frac{\left|\mathrm{R}_i'(j)\right|-1}{\left|A_j\right|-1}. \quad (1)$$

In [6],[7], the above two measures were combined – the LM measure for numerical attributes and the measure of [1] for categorical attributes. The Ambiguity Metric (AM) [9] is the average size of the Cartesian products of all generalized entries in each record in the table.

$$AM(T,g(T')) = \frac{1}{n}.\sum_{i=1}^{i=n}\prod_{j=1}^{j=d}\left|\mathrm{R}_i'(j)\right| \quad (2)$$

A drawback of the AM cost measure is that it also counts combinations of attribute values that do not appear in the original table. The Discernibility Metric (DM) [12] defines the cost of each generalized record R'_i as the number of records in the anonymous table that are indistinguishable from it. In k-anonymizations that are near-optimal, all clusters are of sizes close to k. So, all such anonymizations have approximately the same DM cost and this makes this measure less useful. The Classification Metric (CM) penalizes a record R_i either if its private value differs from the majority of the private values in its cluster, or if R_i is totally suppressed. An average is taken of all rows.

More accurate is the Normalized Certainty Penalty (NCP) [17]. For numerical attributes, the NCP of a cell on attribute A_j that falls in the equivalence class $E_{i,j}$ is defined as:

$$NCP_{A_{i,j}}^{E_{i,j}} = \frac{\max_{A_{i,j}}^{\mathrm{E}_{i,j}} - \min_{A_{i,j}}^{\mathrm{E}_{i,j}}}{\max_{A_j} - \min_{A_j}}. \quad (3)$$

In the case of categorical attributes, where no total order or distance function exists, NCP is defined with respect to the taxonomy tree of the attribute:

$$NCP_{A_{i,j}}^{E_{i,j}} = \begin{cases} 0, & |u| = 1 \\ \left|u_{i,j}\right| / \left|A_j\right|, & otherwise \end{cases} \quad (4)$$

where, u is the lowest common ancestor of all A_j values included in $E_{i,j}$, $|u_{i,j}|$ is the number of leaves (i.e., attribute values) in the subtree of u under which $A_{i,j}$ falls into and $|A_j|$ is the total number of distinct A_j values. NCP of the table is the average of NCPs over all attributes for all of the groups.

$$NCP(T') = \frac{1}{nd}.\sum_{i=1}^{i=n}\sum_{j=1}^{j=d}NCP_{A_{i,j}}^{E_{i,j}} \quad (5)$$

Two commonly employed techniques to preserve privacy are *generalization* and *suppression* [1]. Generalization is the act of replacing an element of the original table with a general value that includes that element (e.g., replacing city name with its state name). Suppression excludes some quasi-identifier attributes or entire records from the table. There are two main models of generalization. In *global recoding* [12], [13], [14], every occurrence of a unique value in the j^{th} column of the table is mapped to a single value that contains it. As a consequence, every single value $a \in A_j$ is always generalized in the same manner. On the other hand, in *local recoding* [4], [5], [7], [8], [12], [14], [19], each entry in the j^{th} column is generalized independently to different values which includes it. Hence, if the age 34, for example, appears in the table in several records, it may be left unchanged, or generalized to (30-39). Local recoding is more flexible and has the potential to achieve higher utility. We consider the case of local recoding that allows greater flexibility and, hence, enables achieving k-anonymity with possibly higher utility.

IV. Quality Metrics

In privacy preserving data sharing, the quality measure can have two components: privacy (P) and utility (U). Privacy measure indicates how difficult it is for a third party to correlate some specific sensitive information to the parties' identity and utility measure indicates how useful the anonymous data might be to the data users. Utility measure can be expressed a function of information loss (IL). The simplest function would be,

$$U = 1 - IL \qquad where, 0 \leq IL \leq 1 \quad (6)$$

A. Proposed Quality Metric

The threshold privacy of an entity in k-anonymity is defined by k i.e., the minimum group size of identical QID values. Different k-anonymization algorithms create different group sizes (with lower bounds of k) and in the same algorithm, different entities can get into groups of different sizes. Thus, different algorithms give different privacy levels. A cost or utility or information loss metric only reflects the degree of accuracy of the anonymous data compared to the original data. However, it does not imply the privacy level of the anonymous data. On the other hand, privacy metric does not represent the quality of the anonymous data. Thus, for fair assessment, we need a new metric that can express both

the utility and privacy level of the anonymous data. For this purpose, we propose our new metric, named *Efficiency* metric.

Let the privacy (P) metric of an entity in a group containing *GS* number of entities (having the same quasi-identifier value) be defined as:

$$P = 1 - \frac{1}{GS}. \tag{7}$$

In case of *k*-anonymity, the value of *GS* is bounded by $k \leq GS \leq n$, where *n* is the number of records, because the minimum group size must be at least equal to *k* and the maximum group size can be equal to *n*. We define *Efficiency*, ξ as:

$$\xi = (1 - IL) \times \left(1 - \frac{1}{GS}\right). \tag{8}$$

Any of the information loss measurement metrics can be used for IL. We choose NCP as it has been accepted by most of the researchers. Thus, Efficiency metric can be calculated as:

$$\xi = \left(1 - \frac{1}{nd} \cdot \sum_{i=1}^{i=n} \sum_{j=1}^{j=d} NCP_{A_{i,j}}^{E_{i,j}}\right) \times \left(1 - \frac{1}{|G|} \sum_{l=1}^{l=|G|} \frac{1}{GS_l}\right). \tag{9}$$

From (9), we can see that the Efficiency metric will return a value of zero if the group size of each of the records is equal to 1. Also, if all of the attribute values are generalized to the root of the generalization hierarchy i.e., $\max_{A_{i,j}}^{E_{i,j}} = \min_{A_{i,j}}^{E_{i,j}}$ $\forall i; 1 \leq i \leq n$ and $\forall j; 1 \leq j \leq d$ and *GS*=*n*, then the Efficiency metric will return a value of (1-1/*n*). Thus the Efficiency metric will get a value between 0 and (1-1/*n*).

V. Algorithms for *K*-Anonymity

Due to the poor performance and limitations of the provable approximation algorithms, heuristic algorithms are invoked [6]-[9], [16]. Among the algorithms with which we compare our algorithm, the Mondrian algorithm [16] is the existing best performing heuristic algorithm. It considers that the data points are sorted along each attribute or dimension. Starting from the whole dataset as a single group, it partitions across the dimension (like *kd*-trees) with the widest normalized range of values. The median along the dimension in the group is used as the dividing point, so approximately half the items fall in each new subgroup. When the normalized ranges of two dimensions are the same, Mondrian selects the first one and splits it into segments. A recursive branch halts when a group cannot be further divided due to the limitation of group size *k*.

Similar to the framework used in Mondrian, we also consider that each of the data items in our framework can be represented with a numerical value so that they become ordered. The data set is considered sorted along each attribute. First, we describe a simple Greedy algorithm for *k*-anonymization.

A. Greedy Algorithm

In the first step, the algorithm searches in each of the *d* attributes of the QID of the original table for the largest groups of records (one for each of the attributes) of size ≥ *k* that, if put into one equivalence class, suffers from minimum information loss on their respective attribute. The algorithm then takes the attribute-data-group (among *d* groups) for which the information loss is minimum on its respective attribute and inserts the attribute in a "WHERE" clause (e.g., "WHERE Sex = 'Male'") so that the "WHERE" clause returns the group of records. The records returned by the above "WHERE" clause becomes the data source and the remaining (*d*-1) attributes becomes the search space for the next step. The above operations of searching the largest group with the minimum information loss in the search space are repeated until each of the attributes of the QID is included in the "WHERE" clause or the group size has become equal to *k*. Finally, the group of records that results from the constructed "WHERE" clause is sent to the anonymous table after being put into an equivalence class (i.e., generalized), and is removed from the original table.

The operations described in the above paragraph are repeated until the number of remaining records in the original table becomes less than *k*. Those remaining records are suppressed. Note that local recoding is used here and thus the equivalence class is created dynamically based on the maximum values and minimum values in the group on each of the attributes.

By analyzing experimental results of the greedy algorithm we have developed the LowCost heuristics algorithm that performs better than our Greedy heuristics.

B. Low Cost Algorithm

In most of the cases, the largest group with the least information loss (if put into one equivalence class) on an attribute was found to be the attribute with the smallest cardinality. Thus unlike the Greedy algorithm, instead of searching in all of the attributes of the QID, the LowCost algorithm searches in the smallest cardinality attribute first for the largest groups of size ≥ *k* that, if put into one equivalence class, suffers from minimum information loss and includes that attribute in the "WHERE" clause. It then searches in the next higher cardinality attribute and similarly includes that attribute in the "WHERE" clause. The process of including an attribute in the "WHERE" clause continues until all of the attributes are included or the group size becomes *k*. The rest of the steps are similar to that of the Greedy algorithm.

For a lower computational cost we follow some additional heuristics. When the cardinality of an attribute in the source data is low $(n/|D_j|) >> k$, instead of searching in the whole dataset for the largest group on the attribute for the minimum cost, we check the group size of the largest group having a distinct value in the attribute. If the group size ≥*k* then a condition can be constructed with the distinct value. Otherwise, a sequential search consisting of at least *k* records on the attribute is necessary. The above heuristics saves much of the computation time, because if $(n/|D_j|) >> k$, there will be at least one value on the j^{th} attribute for which the number of records is ≥*k*. On the other hand, if $(n/|D_j|) << k$, we only need

to search in a few attributes, because as soon as the group size reaches down to *k,* we do not need to search in the remaining attributes. Fig. 1 presents a pseudo code of the LowCost algorithm

```
Low Cost (Input T) {
n=record_count(T); d=QID_size; A[]=QID;
while n>k do {
   condition=null;
   sort A[] on |Dj| Asc, where 1 ≤ j ≤ d;
   for  j=1 to d {
      Find Largest group Gi so that Gen_Cost_to_one_
      Eqv_Class(Gi,j) is minimum;
      condition = condition +  " and Aj between Gi,j(min_value)
      and Gi,j(max_value)";
      T="Select * from T where" + condition };
   T''=T''+ "Select * from T where" + condition;
   T = T- "Select * from T where" + condition; }
output (T'') }
```

Fig. 1. Pseudo code for the lowcost algorithm.

C. Complexity Analysis

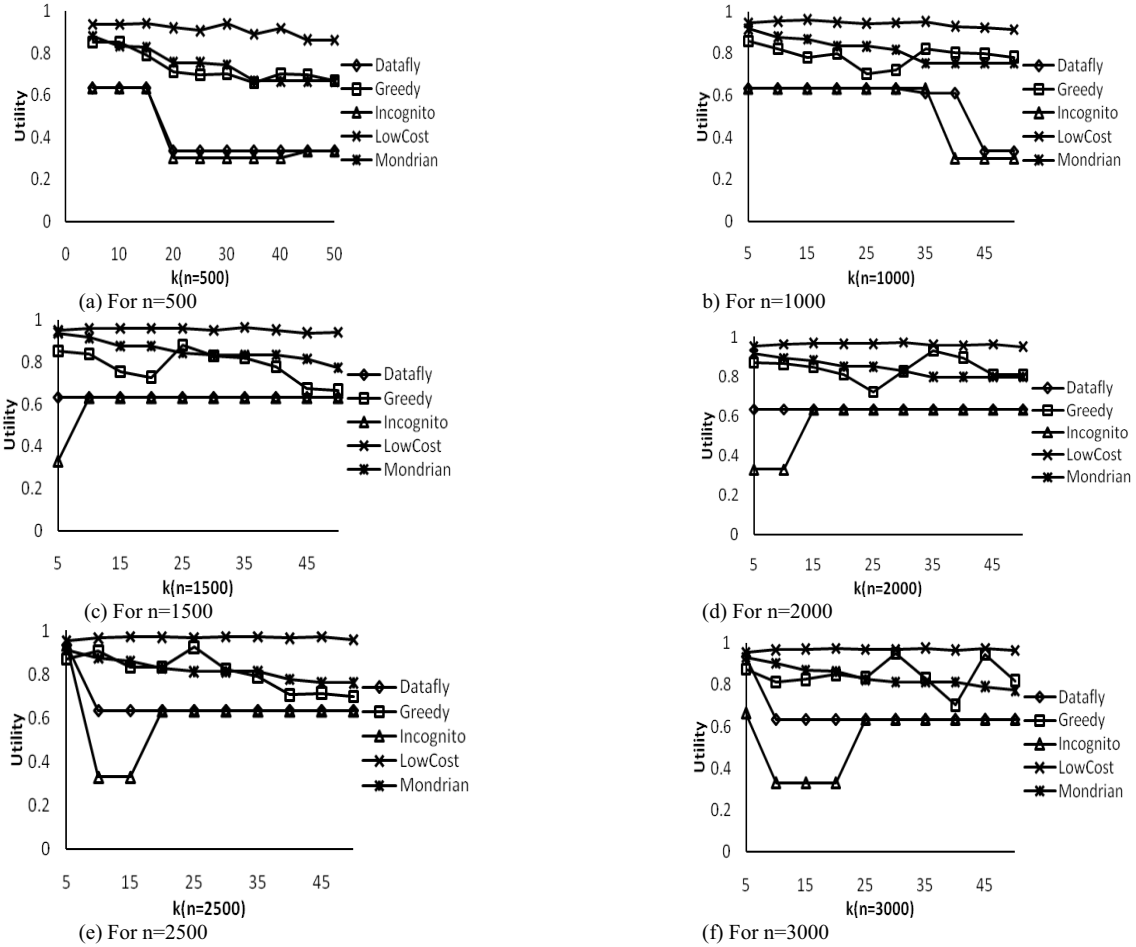
Data are considered sorted on each attribute of the QID. For extracting the first group of records from the original table in to the anonymous table, the maximum search complexity for searching in the first attribute would be *(n-k)*. Search space for the 2nd attribute depends upon the number of records selected by the clause that includes the 1st attribute. A maximum of *(n-k)* searches are necessary for the second attribute. In the worst case, a total of $|A|*(n-k)$ searches are necessary for extracting the first group of records. So, for the whole data set a total of $|A|*((n-k) + (n-2k) + (n-(n/k-1)*k)) = |A|*(n/k)*(n-(k*(k+1) /2))$ comparisons are necessary. Thus, the worst case complexity of the algorithm is of order $O(n^2)$. However, with real data, most of the attributes have a small cardinality of $n/|D_j| \geq k$ (e.g., For "Sex" attribute $|D_j|=2$, for "Age" attribute $|D_j| \approx 100$). The average group size for a specific value at an attribute A_i would be $n/|D_j|$. If $n/|D_j| \geq k$ then the searching cost on attribute A_i will be 1 (largest group at the top). So, average computation time with real data will be very low.

VI. Experimental Results

We took three attributes as the quasi-identifier: "Date of birth", "Sex" and "Zip code" with their domain cardinality of 3653(i.e., 10 years range), 2 (Male or Female) and 1000 (3 digits) respectively. Random data were generated for the attributes within the above mentioned cardinalities. We measured utility and Efficiency for Datafly [18], Greedy, Incognito [11], LowCost and Mondrian [16] algorithms.

A. Comparison in the Utility Metric

Fig. 2(a) to 2(f) compare the average utilities per data element ($A_{i,j}$) in the five algorithms for 500 records, 1000 records, 1500 records, 2000 records, 2500 records and 3000 records respectively with varying anonymity parameter *k.*



(a) For n=500 (b) For n=1000 (c) For n=1500 (d) For n=2000 (e) For n=2500 (f) For n=3000

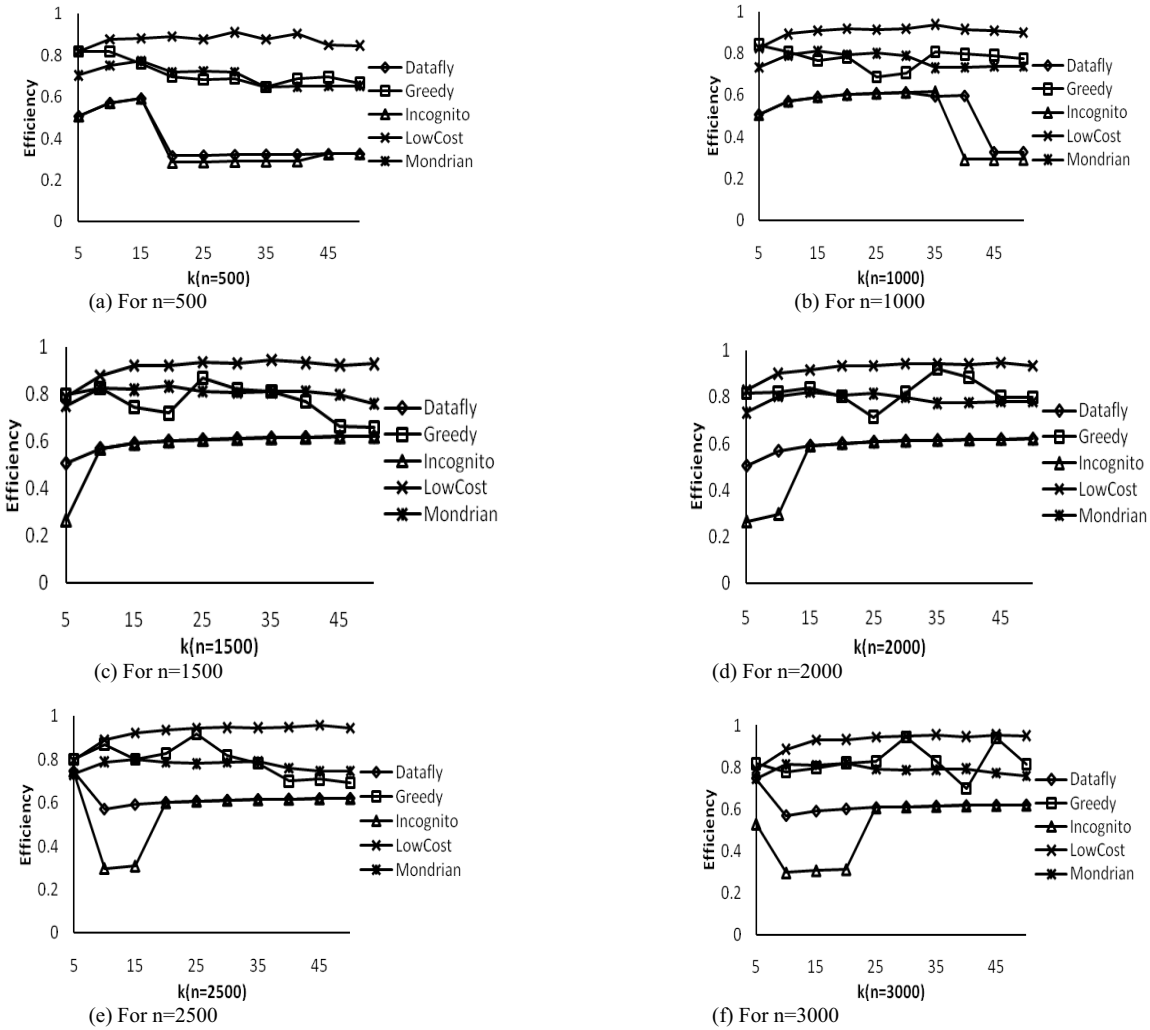
Fig. 2. Comparison of utility (U) measures in five algorithms for varying anonymization parameter *k*.

In general, the utilities offered by all of the algorithms increases with an increase in the value of *n*. Datafly and Incognito offer similar utility but are the worst among the algorithms and change to fewer levels with changes in the values of *k*. They perform poorly because of their use of global recoding for generalization.

The utility offered by Mondrian decreases with an increase in the value of *k* while it remains almost constant in LowCost algorithm with different values of *k*. Fig. 2 shows that the LowCost al gorithm performs the best, followed by Mondrian and the Greedy algorithm. The higher cost in Mondrian is due to the fact that it starts with the largest and most generalized group and then splits the group if the splitting satisfies *k*-anonymity. However, in our algorithm we start with the most specific value and find an equivalence class of size ≥*k*. Thus, the values are generalized less in our algorithm.

A. Comparison in the Efficiency Metric

Fig. 3 compares the average Efficiency (ξ) for the same settings. In general, ξ value in an algorithm does not change much for different values of *n*, though a smaller ξ value was observed in the LowCost algorithm for very small values of *n*. Here also, the ξ value in the Greedy algorithm was found to be relatively arbitrary for varying *k* and those for Datafly and Incognito changed to fewer levels with varying *k*. Efficiency values in Datafly and Incognito were similar and worse, followed by Mondrian and Greedy. The LowCost algorithm offers the best Efficiency among the five algorithms.



(a) For n=500 (b) For n=1000

(c) For n=1500 (d) For n=2000

(e) For n=2500 (f) For n=3000

Fig. 3. Comparison of Efficiency measures in five algorithms for varying anonymization parameter *k*.

Please note that the graphs represent utility and Efficiency measures of the anonymous data per cell. So, a small difference in the measurement per cell will make a huge difference on the total utility or Efficiency for the whole dataset when the number of records and/or the number of attributes is large.

VII. Conclusions

In this paper, we first propose the Efficiency metric ξ that represents both the utility and privacy of the anonymous data and can assess anonymization algorithms fairly. It gets a higher value with increased utility and also with increased privacy. Its value ranges from 0 to (1-1/*n*) where *n* is the number of records is the database. We also present our heuristic Greedy algorithm and LowCost algorithm. The LowCost algorithm appears to be the algorithm of choice for efficiently finding *k*-anonymization with high utility and high privacy as indicated by comparative experimental results with other well-known algorithms. While other algorithms' utilities are around 80% and below, our LowCost algorithm achieves above 95% utility. Also, while other algorithms'

Efficiencies remain below 80%, the Efficiency of LowCost reaches above 90%. In conclusion, the LowCost algorithm is more efficient than existing algorithms for k-anonymization.

REFERENCES

[1] L. Sweeney, "*k*-Anonymity: A Model for Protecting Privacy," *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, pp. 557–570, 2002.

[2] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *ACM Sigmod Record*, vol. 29, no. 2, pp. 439-450, 2000.

[3] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation algorithms for k-anonymity," *Journal of Privacy Technology*, 2005.

[4] A. Gionis and T. Tassa, "k-Anonymization with minimal loss of information," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 2, pp. 206-219, 2009.

[5] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," *Proc. of ACM Sigmod-Sigact-Sigart Symposium, Pods*, 2004, pp. 223-228.

[6] J. W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," *Proc. of Int. Conf. on Database systems for advanced applications (DASFAA)*, 2007, pp. 188-200.

[7] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A framework for efficient data anonymization under privacy and accuracy constraints," *ACM Trans. Database Systems*, vol. 34, no. 2, 2009.

[8] A. Gionis, A. Mazza, and T. Tassa, "k-Anonymization revisited," in *Proc. of IEEE Int. Conf. on Data Eng. (ICDE)* , 2008, pp. 744 – 753.

[9] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," *Journal of Data and Knowl. Eng*., pp. 622-645, 2007.

[10] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. of Int. Conf. on Knowl. discovery and data mining*, 2002, pp. 279 – 288.

[11] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain k-anonymity," *Proc. of SIGMOD*, 2005, pp. 49–60.

[12] R. C. W. Wong, J. Li, A. W. C. Fu, and K. Wang, "(α, *k*)-anonymity: An enhanced *k*-anonymity model for privacy preserving data publishing," *Proc. of Int. Conf. on Knowl Discovery and data mining*, 2006, pp. 754–759.

[13] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," *Proc. of IEEE Int. Conf. on Data Eng.(ICDE)*, 2005, pp. 217-228.

[14] H. Park and K. Shim, "Approximate algorithms for k-anonymity," in *Proc. of SIGMOD*, 2007, pp. 67-78.

[15] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *Proc. of SIGMOD*, 2006, pp. 217–228.

[16] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. of IEEE Int. Conf. on Data Eng. (ICDE)* , 2006, pp. 25-25.

[17] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu, "Utility-Based Anonymization Using Local Recoding," *Proc. of Int. Conf. on Knowl. discovery and data mining*, 2006, pp. 785–790.

[18] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[19] J. Goldberger and T. Tassa, "Efficient Anonymizations with Enhanced Utility," *Transactions on Data Privacy*; vol. 3, pp. 149–175, 2010.

Md Nurul Huda received his BSc degree in applied physics from the University of Dhaka, Bangladesh, in 1995, MSc degree in computer science from the same university in 1997, and PhD degree in the field of informatics from the Graduate University for Advanced Studies, Tokyo, Japan in 2007. He was a faculty at the Department of Computer Science and Engineering, University of Dhakafrom 1998 to 2010. He was awarded a JSPS Fellowship in 2007 and was a postdoctoral fellow at the National Institute of Informatics, Tokyo, Japan from 2007 to 2009. He is currently working as a researcher at the Research Organization of Information and Systems (ROIS), in Tokyo, Japan. His interests include privacy enhancing technologies, privacy preserving data sharing, usability against privacy, mobile and wireless communication and networks, disaster recover networks, delay tolerant network (DTN) routing and architecture, performance improvement in DTN and mobile ad hoc routing protocols. Dr. Huda is a member of the IPSJ (Information Processing Society of Japan) and the IEEE.

Shigeki Yamada was born in Sapporo, Hokkaido, Japan, on Sep. 16, 1949. He received his BS, MSc, and PhD degrees all in electronic engineering from Hokkaido University, Japan in 1972, 1974 and 1991 respectively. From 1972 to 1999, Dr. Shigeki was joining NTT (Nippon Telegraph and Telephone Corporation) laboratories and was involved in the research and development of digital switching systems, massively parallel processing systems and distributed network architectures. In 2000, he joined National Institute of Informatics (NII), Tokyo, Japan. Currently, he is Professor and Director of Principles of Informatics Research Division, NII. He has published more than 100 reviewed journal and international conference papers. His current research interests include future network architectures, Delay- and Disruption- Tolerant Networks (DTN), advanced network applications and privacy protection technologies. Prof. Yamada is a member of IEICE (Institute of Electronics, Information and Communication Engineers), IPSJ (Information Processing Society of Japan) and senior member of IEEE. He is Trustee of JPNIC (Japan Network Information Center). He received the Tutorial Paper Award from IEICE Communications Society in 2009.

Noboru Sonehara received his BE and ME degrees from Shinshu University, Japan in 1976 and 1978, respectively, and his Ph D in 1994. He has been a Professor in the information and society research division at the National Institute of Informatics since 2004. From 2001 to 2004, he was a project manager (Content Commerce Project) at NTT cyber solutions laboratories. He has been a Director of the Information and Society Research Division since 2006. His current research interests include ICT security, privacy, trust, risk, resilience, and e-authentication platforms. Prof. Sonehara is a member of IEICE (Institute of Electronics, Information and Communication Engineers), IPSJ (Information Processing Society of Japan) and the IEEE. He was awarded Fellow of the IEICE in 2010.