On the Optimal Selection of k in the k–Anonymity Problem

Rinku Dewri, Indrajit Ray, Indrakshi Ray and Darrell Whitley

Department of Computer Science, Colorado State University
Fort Collins, CO 80523, USA
{rinku,indrajit,iray,whitley}@cs.colostate.edu

***Abstract*— When disseminating data involving human subjects, researchers have to weigh in the requirements of privacy of the individuals involved in the data. A model widely used for enhancing individual privacy is k–anonymity, where an individual data record is rendered similar to $k-1$ other records in the data set by using generalization and/or suppression operations on the data attributes. The drawback of this model is that such transformations result in considerable loss of information that is proportional to the choice of k. Studies in this context have so far focused on minimizing the information loss for some given value of k. However, owing to the presence of outliers, a specified k value may or may not be obtainable. Further, an exhaustive analysis is required to determine a k value that fits the loss constraint specified by a data publisher. In this paper, we formulate a multi-objective optimization problem to illustrate that the decision on k can be much more informed than being a choice solely based on the privacy requirement. The optimization problem is intended to resolve the issue of data privacy when data suppression is not allowed in order to obtain a particular value of k. An evolutionary algorithm is employed here to provide this insight.**

I. Introduction

Public dissemination of personal data is often a requirement to facilitate various scientific studies, business processes, or legal procedures. Such data often involve sensitive information unsuitable for dissemination in a public manner. Data in such cases is usually made void of the presence of sensitive attributes and made anonymous before sharing. However, re-identifying these hidden attributes is not impossible when other publicly available information can be linked with the shared data. Thus, individuals who were anonymized previously can be de-anonymized leading to privacy violations.

To address such privacy concerns, Sweeney proposed the concept of k–anonymity that reduce the chances of a "linking attack" being successful [1]. Anonymization of data sets involve transforming the actual data set into a form unrecognizable in terms of the exact data values, by using data generalization and suppression. An unavoidable consequence of performing anonymization is a loss in the quality of the data set. Most studies till now have focused on algorithms to minimize the information loss for a predetermined value of k. However, from a data publisher's perspective, the question should be asked in the opposite direction – given an acceptable level of information loss, what value of k should be used to guarantee higher privacy. Further, certain k values may be unobtainable when data suppression is not allowed.

In this paper, we ask the following questions: When data suppression is not allowed, can one guarantee higher privacy levels for a larger number of the individuals represented in the data set? If an acceptable level of information loss is specified, can such a solution be generated? Our approach is primarily based on the formulation of a multi-objective optimization problem, the solutions to which provide a preliminary understanding of the trade-off present between the level of privacy and the quality of the anonymized data set. We employ a popular evolutionary algorithm to solve the multi-objective optimization problem relevant to this study.

II. Related Work

Several algorithms have been proposed to find effective k–anonymization. Sweeney's Datafly approach uses a heuristic method to generalize the attribute containing the most distinct sequence of values for a provided subset of quasi-identifiers [2]. Samarati's algorithm [3] can identify all k–minimal generalizations, out of which an optimal generalization can be chosen based on certain preference information provided by the data recipient. Iyengar proposes a flexible generalization scheme and uses a genetic algorithm to perform k–anonymization on the larger search space that resulted from it [4]. Although the method can maintain a good solution quality, it has been criticized for being a slow iterative process. In this context, Lunacek et al. introduces a new crossover operator that can be used with a genetic algorithm for constrained attribute generalization, and effectively show that Iyengar's approach can be made faster [5]. In order to obtain a guaranteed optimal solution, Bayardo and Agrawal propose a complete search method that iteratively constructs less generalized solutions starting from a completely generalized data set [6].

III. Preliminaries

A data set D can be visualized as a tabular representation of a multi-set of tuples $r_1, r_2, \ldots, r_{n_{row}}$ where n_{row} is the number of rows in the table. Each tuple (row) r_i comprises of n_{col} values $\langle c_1, c_2, \ldots, c_{n_{col}} \rangle$ where n_{col} is the number of columns in the table. The values in column j correspond to an *attribute* a_j, the domain of which is represented by the ordered set $\Sigma_j = \{\sigma_1, \sigma_2, \ldots, \sigma_{n_j}\}$.

A *generalization* G_j for an attribute a_j is a partitioning of the set Σ_j into ordered subsets $\langle \Sigma_{j_1}, \Sigma_{j_2}, \ldots, \Sigma_{j_K} \rangle$ which preserves the ordering in Σ_j, i.e. if σ_a appears before σ_b

in Σ_j then, for $\sigma_a \in \Sigma_{j_l}$ and $\sigma_b \in \Sigma_{j_m}$, $l \leq m$. Further, every element in Σ_j must appear in exactly one subset. The elements in the subsets maintain the same ordering as in Σ_j. Generalizations for categorical attributes are *constrained* to only those which respect a specified hierarchy.

Equivalent tuples after anonymization can be grouped together into equivalence classes. The *k–anonymity* problem is then defined as the problem of finding a set of generalizations for the attributes in D such that the equivalence classes induced by anonymizing D using the generalizations are all of size at least k.

IV. Problem Formulation

Consider the data value $v_{i,j}$ at row i and column j in the data set D. The general loss metric assigns a penalty to this data value based on the extent to which it gets generalized during anonymization. Let $g_{i,j} = G_j(v_{i,j})$ be the index of the subset to which $v_{i,j}$ belongs to in the generalization G_j, i.e. $v_{i,j} \in \Sigma_{j_{g_{i,j}}}$. The penalty for information loss associated with $v_{i,j}$ is then given as follows:

$$loss(v_{i,j}) = \frac{|\Sigma_{j_{g_{i,j}}}| - 1}{|\Sigma_j| - 1} \tag{1}$$

For categorical data, the loss for a cell is proportional to the number of leaf nodes rooted at an internal node (the generalized node) of the hierarchy tree. The loss attains a maximum value of one when the cell is suppressed ($G_j = \langle \Sigma_j \rangle$), or in other words, when the root of the tree is the generalized node. Subtracting one ensures that a non-generalized value incurs zero loss since the cardinality of the subset to which it belongs would be one. The generalization loss is then obtained as the total loss over all the data values in the data set.

$$GL = \sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}} loss(v_{i,j}) \tag{2}$$

The presence of outliers in a data set makes it difficult to find a suitable value of k when suppression of data is not allowed. In this formulation, we strictly adhere to the requirement that no tuple in the data set can be deleted. Intuitively, such a strict requirement makes the k–anonymity problem insensible to solve for a given k, as the optimization algorithm will be forced to overly generalize the data set in its effort to ensure k–anonymity. The outliers usually belong to very small equivalence classes and the only way to merge them into a bigger one is by having more generalization. This effectuate more loss in information which is often not acceptable to an user.

Although solving the k–anonymity problem is not possible in terms of its strict definition, it is worth noting that a generalization can still affect the distribution of the equivalence classes even when suppression is not allowed. An equivalence class E_k in this description group all tuples that are similar to exactly $k-1$ other tuples in the anonymized data set. An ideal generalization would then maintain an acceptable level of loss by keeping the number of rows in smaller equivalence classes (small k) relatively lower than in the bigger equivalence classes. Although this does not guarantee complete k–anonymity, the issue of privacy breach can be solved to a limited extent by reducing the probability that a randomly chosen row would belong to a small equivalence class.

With this motivation, we define the *weighted-k–anonymity* multi-objective problem to find generalizations that produce a high weighted-k value and low generalization loss. Each equivalence class E_k defines a k value, $k \leq n_{row}$, for its member tuples – every tuple in the equivalence class is same as exactly $k-1$ other tuples in the same class.

Note that this notion of an equivalence class is different from the one stated in the k–anonymity problem. Two rows in the original data set belong to the same equivalence class in the k–anonymity definition if the generalization transforms them into the same tuple. However, in this formulation, a row belongs to the equivalence class E_i if a generalization makes it i–anonymous.

The weighted-k for a particular generalization inducing the equivalence classes $E_1, E_2, \ldots, E_{n_{row}}$ on the anonymized data set is then obtained as follows:

$$k_{weighted} = \frac{\sum_{i=1}^{n_{row}} (i \cdot |E_i|)}{\sum_{i=1}^{n_{row}} |E_i|} \tag{3}$$

Note that, in most cases, not all equivalence classes with all possible k values will be generated. The weighted-k value provides a sufficiently good estimate of the distribution of the equivalence classes. A high weighted-k value implies that the size of the equivalence classes with higher k is relatively more than the size of the lower k ones. The multi-objective problem is then formulated as *finding the generalization that maximizes the weighted-k and minimizes the generalization loss for a given data set.*

V. Solution Methodology

We employ the NSGA-II algorithm [7] to solve the multi-objective problem defined in the previous section. Here we adopt the encoding suggested by Iyengar in [4]. Further, to be able to use Lunacek et al.'s crossover operator, the validity of the parent solutions must be guaranteed. Starting from the root node, a node randomly decides if it would allow its subtrees to be distinguishable. If it decides not to then all nodes in its subtrees are assigned the same identifier. Otherwise the root of each subtree receives an unique identifier. The decision is then translated to the root nodes of its subtrees and the process is repeated recursively. Once all leaf nodes are assigned an identifier, two adjacent leaf nodes in the imposed ordering are combined only if they have the same identifier. Since a parent node always make the decision if child nodes will be combined or not, all generalizations so produced will always be valid.

We applied our methodology to the "adult.data" benchmark dataset available from the UCI machine learning repository[1]. We prepare the dataset as described in [4], [6]. For NSGA-II, we set the population size as 200 and the maximum

[1] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/

number of iterations as 250. A single point crossover is used for unconstrained attributes while Lunacek et al.'s crossover operator is used for constrained attributes. Also, mutation is only performed on the unconstrained attributes. The remaining parameters of the algorithm are set as follows: crossover rate $= 0.9$, mutation rate $= 0.1$ with binary tournament selection.

VI. Results and Discussion

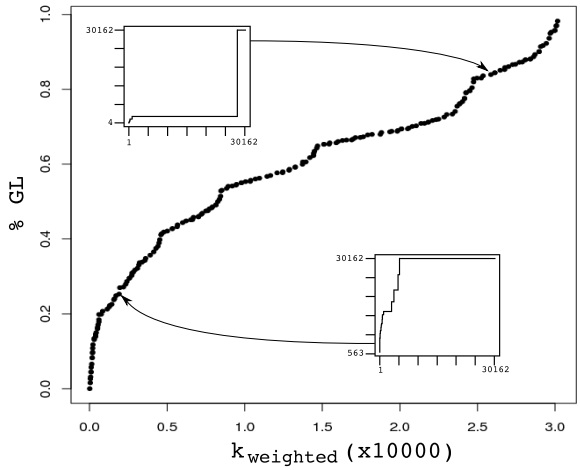

Fig. 1. Solutions found by NSGA-II. Inset figures show cumulative distribution of $|E_k|$ as k increases.

Fig. 1 shows the different trade-off solutions obtained by NSGA-II. A point in the plot corresponds to a solution that induces a particular distribution of k values on the anonymized data set. As expected, the generalization loss increases as the distribution of equivalence classes get more inclined towards higher k values. In the absence of suppression, a single k value is often hard to enforce for all tuples in the data set. Thus, a solution here results in different k values for different tuples. A higher k-weighted value signifies that most tuples have a high k value associated with them, in which case, the generalization loss is higher. A solution with low k-weighted value results in a generalization with low k values for its tuples.

The inset figures in the plot depict the cumulative distribution of the number of tuples belonging to equivalence classes (y-axis) with different k values (x-axis). The distributions of two extreme solutions corroborate the speculation that a higher generalization loss must be incurred to assure a greater level of privacy (higher k values) for a larger section of the dataset. Low generalization losses are only possible when most tuples belong to equivalence classes of lower k value.

However, it should be noted that the distributions for the two example solutions are not complementary in nature. For the solution with lower generalization loss, the distribution has a continuously increasing trend, implying that equivalence classes of different k values exist for the solution. The other solution shows an abrupt increase signifying that the tuples either belong to equivalence classes with very small k or ones with very large k. The sought balance in the distribution can therefore exist with an acceptable level of generalization loss.

At this point, a data publisher's question regarding a solution with a specified level of information loss can be answered from the plot. Further, the trade-off analysis enables one to suggest a better privacy level (higher k-weighted) which can perhaps be obtained by increasing the level of information loss by a small amount.

VII. Conclusions

In this paper, we present an analytical approach to demonstrate that the choice of the k value in the k–anonymity problem can be made in a much informed manner rather than arbitrarily as is currently done. For generalizations without suppression, an unique k may not be available. However, the analysis indicates that generalizations are possible that provide a higher level of privacy for a higher fraction of the dataset without compromising much on its information content. We believe that such an analysis not only reinstates the decision maker's confidence in its choice of a particular generalization, but also identifies ways of examining if the level of privacy requested by a human subject is achievable within the acceptable limits of perturbing data quality.

Future work in this front can involve the formulation of a series of multi-objective problems when suppression is allowed up to a given limit. Analysis of this nature need not be limited to k–anonymity alone and other data anonymization techniques can be explored as well. Further, there is the more difficult problem of obtaining generalizations that are optimal with respect to more than one privacy preserving technique.

Acknowledgment

This work is partially supported by AFOSR under contract number FA9550-07-1-0042. The views and conclusions contained in this document are of the authors and should not be interpreted as representing official policies, either expressed or implied, of AFOSR or other federal government agencies.

References

[1] L. Sweeney, "k–Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[2] 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.

[3] P. Samarati, "Protecting Respondents' Identities in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[4] V. S. Iyengar, "Transforming Data to Satisfy Privacy Constraints," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Alberta, Canada, 2002, pp. 279–288.

[5] M. Lunacek, D. Whitley, and I. Ray, "A Crossover Operator for the k-Anonymity Problem," in *GECCO 2006: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, Seattle, Washington, USA, 2006, pp. 1713–1720.

[6] R. J. Bayardo and R. Agrawal, "Data Privacy Through Optimal k-Anonymization," in *ICDE 2005: Proceedings of the 21st International Conference on Data Engineering*, Tokyo, Japan, 2005, pp. 217–228.

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA–II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.