A Multi-Objective Approach to Data Sharing with Privacy Constraints and Preference Based Objectives

Rinku Dewri, Darrell Whitley, Indrajit Ray and Indrakshi Ray
Department of Computer Science
Colorado State University
Fort Collins, CO 80523, USA
{rinku,whitley,indrajit,iray}@cs.colostate.edu

ABSTRACT

Public data sharing is utilized in a number of businesses to facilitate the exchange of information. Privacy constraints are usually enforced to prevent unwanted inference of information, specially when the shared data contain sensitive personal attributes. This, however, has an adverse effect on the utility of the data for statistical studies. Thus, a requirement while modifying the data is to minimize the information loss. Existing methods employ the notion of "minimal distortion" where the data is modified only to the extent necessary to satisfy the privacy constraint, thereby asserting that the information loss has been minimized. However, given the subjective nature of information loss, it is often difficult to justify this assertion. In this paper, we propose an evolutionary algorithm to explicitly minimize an *achievement function* given constraints on the privacy level of the transformed data. Privacy constraints specified in terms of anonymity models are modeled as additional objectives and an evolutionary multi-objective approach is proposed. We highlight the requirement to minimize any bias induced by the anonymity model and present a scalarization incorporating preferences in information loss and privacy bias as the achievement function.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*security, integrity, and protection*; I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search—*Heuristic methods*

General Terms

Experimentation

Keywords

Disclosure control, anonymization bias, constraint handling, multi-objective optimization

1. INTRODUCTION

Privacy violations emanating from the sharing of personal information collected at various public institutions have raised an important concern in recent few years. Much research in information assurance has therefore delved into the protection of respondent identity. The question to answer is how such information be modified so that the data is useful for statistical studies while protecting the respondents' identities.

Removing personally identifiable information such as name and social security number is not sufficient to ensure privacy. A recent study on the year 2000 census data of the U.S. population reveals that 53% of the individuals can be uniquely identified by their gender, city and date of birth; 63% if the ZIP code is known in addition [5]. Such attributes, called *quasi-identifiers*, can be linked with other publicly available information to establish the re-identification. A classic experiment demonstrating the possibility is presented by Sweeney [18] where she managed to obtain the medical records of the Governor of Massachusetts from a medical insurance data set, containing no explicit identifying information, and a voter's registration list.

To attend to such privacy concerns, Samarati and Sweeney proposed the concept of data *generalization* to be used to satisfy a property called k*–anonymity* [16, 17]. Generalization of data is performed by grouping together data attribute values into a more general one, for example replacing the age by an age range. A transformed data set of this nature is then said to be k–anonymous if each record in it is same as at least $k-1$ other records. This property implies that any record can be related to at least k underlying individuals.

A consequential drawback of performing generalization is a loss in integrity of the data set. A number of algorithms have therefore been proposed to generalize a data set to meet the k–anonymity property while resulting in minimum information loss [1, 4, 6, 7, 9, 10, 11, 15, 19]. The standard approach is to progressively generalize the data until it is k–anonymous. Such an approach cannot guarantee optimality if different attributes carry different levels of significance. For example, in a medical data set, attributes such as age and disease are more important than the ZIP code of the underlying patient. This opens up the possibility that a minimum information loss can be sustained even for higher values of k, thereby providing better privacy than specified. Searching for higher privacy generalizations is also fruitful if the data publisher can tolerate an information loss higher than the minimum possible. Existing optimization attempts do not embed such preference criteria.

Further, k–anonymity is only a minimalistic measure of the privacy level. The actual privacy levels of two individuals in a k–anonymous data set can be very different. For example, consider a 3–anonymous data set. If record A is same as 2 other records while record B is same as 9 other records, the privacy level of individual B is much higher (9/10) than that of individual A (2/3). This factor, which we call the *anonymization bias,* is induced by nature of the k–anonymity model since it only helps identify the worst case privacy level while minimizing the information loss. Given the subjective nature of information loss, we cannot ignore the possibility of a reciprocal relationship between privacy bias and information loss.

In this paper, we present an approach to obtain data generalizations satisfying the k–anonymity property given preference values on the information loss and privacy bias. An achievement scalarizing function is formulated using the preference values and subjected to a constrained minimization. We provide the necessary arguments to prove that a constrained minima of this function is Pareto-optimal with respect to loss and bias. The approach draws its power from the multi-objective treatment of the constrained single objective optimization problem. We show how the proposed evolutionary multi-objective approach helps resolve the issue of finding better privacy levels than specified (by the parameter k) in the presence of varying data attribute significance and data publisher preferences. We extend and compliment this work by investigating how factors such as population size, weight assignments and bias preference can affect solution quality.

The remainder of the paper is organized as follows. Section 2 reviews some of the existing optimization algorithms for k–anonymization. Section 3 provides a preliminary background on the problem along with metrics to quantify information loss and anonymization bias. Section 4 defines the achievement function and formulates a constrained optimization problem over it. Section 5 presents our solution methodology. Empirical results on a benchmark data set are presented in Section 6 along with arguments discussing the effectiveness of the proposed approach. Finally, Section 7 concludes the paper.

2. RELATED WORK

Several algorithms have been proposed to find effective k–anonymization. The *μ-argus* algorithm is based on the greedy generalization of infrequently occurring combinations of quasi-identifiers and suppresses outliers to meet the k–anonymity requirement [6]. The *Datafly* approach uses a heuristic method to first generalize the quasi-identifier containing the most number of distinct values [17]. Sequences of quasi-identifier values occurring less than k times are suppressed.

On the more theoretical side, Sweeney propose the *MinGen* algorithm [17] that exhaustively examines all potential generalizations to identify the optimal generalization that minimally satisfies the anonymity requirement. However, the approach is impractical even on modest sized data sets. Meyerson and Williams have recently proposed an approximation algorithm that achieves an anonymization with $\mathrm{O}(k \log k)$ of the optimal solution [12]..

A genetic algorithm based formulation is proposed by Iyengar to perform k–anonymization [7]. Iyengar's method requires that values in an attribute's domain be linearly ordered in some manner in order to enable a flexible representation of a generalization scheme. Bayardo and Agrawal propose a complete search method that iteratively constructs less generalized solutions starting from a completely generalized data set [1]. The algorithm starts with a fully generalized data set and systematically specializes it into one that is minimally k–anonymous. The idea of a *solution cut* is presented by Fung et al. in their approach to top down specialization [4]. A generalization is visualized as a "cut" through the taxonomy tree of each attribute. A cut of a tree is a subset of values in the tree that contains exactly one value on each root-to-leaf path. A solution cut is a cut that satisfies the anonymity requirement.

As mentioned earlier, these algorithms lack any exploration of solutions with higher privacy levels than specified. Moreover, they do not embed any preference criteria in their design. Anonymization bias is also not taken into account by any of these algorithms.

3. DISCLOSURE CONTROL

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 $\Sigma_j = \{\sigma_1, \sigma_2, \ldots, \sigma_{n_j}\}$. The ordering of elements in the set can be implicit by nature of the data. For example, if the attribute is age, the ordering can be done in increasing order of the values. Categorical data are usually associated with a taxonomy tree. The leaf nodes in this tree constitute the actual values that the attribute can take. The ordering for these values can be assigned based on the order in which the leaf nodes are reached in a preorder traversal of the tree [7].

A *generalization* G_j for an attribute a_j is a partitioning of the set Σ_j into ordered subsets $\langle \Sigma_{j_1}, \Sigma_{j_2}, \ldots, \Sigma_{j_P} \rangle$ 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. For the age attribute with the domain $\{10, \ldots, 90\}$, a possible generalization can be $\langle \{[10, 30]\}, \{(30, 50]\}, \{(50, 70]\}, \{(70, 90]\} \rangle$. For categorical attributes, a generalization is typically required to respect the taxonomy tree. However, an ill-defined taxonomy tree can heavily constrain the number of generalizations possible for such attributes and affect the overall information content in the anonymized data. Another strategy, and the one adopted in this study, is to take into account all possible generalizations for categorical attributes as well and perform the appropriate nomenclature of the groups in a post-optimization stage.

Given the generalizations $G_1, G_2, \ldots, G_{n_{col}}$, the data set D can be transformed to the *anonymized* data set D' by replacing each value v_{ij} at row i and column j in D by $G_j(v_{ij})$ where $G_j(v_{ij})$ gives the subset index to which v_{ij} belongs in the generalization G_j. Note that the number of partitions (or groups) of an attribute domain signifies the extent of generalization that will be performed for the attribute. If $P = 1$ then all values of the attribute will be transformed to the same subset index 1, in which case all information in that attribute is lost. On the other extreme, if $P = |\Sigma_j|$ for attribute a_j then every value will map to its own unique in-

dex (no generalization) and all information in the attribute will be maintained in the original form.

A consequence of performing generalization is the appearance of equivalent tuples. Two tuples in D are equivalent if their subset indices are equal in every column of D'. Such equivalent tuples can then be grouped together into equivalence classes. We associate a value ec_i to each tuple in D' signifying the size of the equivalence class to which it belongs to. *k–anonymity* is then defined as follows.

Definition 1. **(*k–anonymity*)** An anonymized data set D' is said to be k–anonymous if $\min(EC_{\mathsf{D}'}) \geq k$, where $EC_{\mathsf{D}'}$ is the vector $(ec_1, \ldots, ec_{n_{row}})$ for D'.

In other words, every tuple in a k–anonymous data set is same as at least $k-1$ other tuples. Higher the value of the parameter k, better is the privacy guarantee. We can say that the probability of privacy breach is at most $1/k$ in a k–anonymous data set. k–anonymity satisfies the *monotonicity* property, i.e. a k–anonymous data set is also $(k-1)$–anonymous. We shall thus refer to the parameter k in k–anonymity as k_{pref} and $\min(EC_{\mathsf{D}'})$ as the effective k resulting from the generalizations.

3.1 Normalized Weighted Penalty

Loss metrics assign some notion of penalty to each tuple whose data values get generalized, thereby reflecting the total information lost in the anonymization process. Consider the data value v_{ij} at row i and column j in the data set D. Let $g_{ij} = G_j(v_{ij})$ be the index of the subset to which v_{ij} belongs in the generalization G_j, i.e. $v_{ij} \in \Sigma_{j_{g_{ij}}}$. Further, let $(w_1, \ldots, w_{n_{col}})$ be a vector of weights where weight $0 \leq w_i \leq 1$ reflects the importance of the attribute a_i. The sum of weights is fixed at 1.0. The penalty for information loss associated with the value v_{ij} is then given as follows.

$$penalty(v_{ij}) = \frac{w_j(|\Sigma_{j_{g_{ij}}}| - 1)}{(|\Sigma_j| - 1)}$$

The loss is thus proportional to the size of the partition to which a data value belongs to. It attains a maximum value (equal to the weight of the attribute) when $P = 1$. Subtracting one ensures that a non-generalized value incurs zero loss since the cardinality of the subset to which it belongs would be one. An entire tuple can thus have a penalty of at most 1.0. The *normalized weighted penalty* in D' is then obtained as the fractional penalty over all tuples in the data set.

$$NWP(\mathsf{D}') = \frac{\sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}} penalty(v_{ij})}{n_{row}}$$

3.2 Normalized Equivalence Class Dispersion

The k–anonymity model is only representative of the worst case privacy measurement. As a result, it is possible that two anonymized versions of a data set, both satisfying k–anonymity, result in very different equivalence class sizes for the tuples. The privacy level of a tuple is directly related to its ec_i value – the higher the value, lower is the probability of privacy breach. Since the k–anonymity definition does not enforce any requirement on how ec_i values should be distributed, it is often possible that an anonymization is biased towards a set of tuples ($ec_i \gg k_{pref}$) while providing minimalistic privacy ($ec_i = k_{pref}$) for others. Our attempt here is to control the occurrence of such biased privacy within acceptable limits.

The value of ec_i for a tuple can range from 1 to the number of tuples in the data set, i.e. n_{row}. This range reflects the maximum bias that can be present in the anonymized data set. The *normalized equivalence class dispersion* measures the bias as the maximum dispersion present in the ec_i values relative to the maximum possible dispersion.

$$NECD(\mathsf{D}') = \frac{\max(EC_{\mathsf{D}'}) - \min(EC_{\mathsf{D}'})}{n_{row} - 1}$$

4. PREFERENCE BASED OPTIMIZATION

A typical constrained optimization problem explored in disclosure control is to find an anonymized version of a data set, or effectively a set of generalizations resulting in the anonymized version, that induce minimum information loss subject to the constraint that the anonymized data set is k–anonymous. Given the NP-hard nature of the problem [12], heuristic based approaches in this context progressively increase the amount of generalization for the attributes until the k–anonymity constraint is satisfied [1, 4, 19]. The anonymized data set at this point is assumed to incur minimum information loss. These approaches have two major drawbacks.

First, the information loss metric is assumed to have a monotonic relationship with the amount of generalization. In other words, as more generalization is performed (no matter for which attribute), the information loss increases. Only under this assumption can one claim that by performing generalization only to the extent necessary to satisfy the k–anonymity constraint, we shall also be minimizing the information loss. However, the assumption is not valid when all attributes do not carry the same significance. Hence, less important attributes may be generalized more while more important attributes may be generalized less without affecting the information content of the data set. This implies that the optimal solution need not necessarily have an effective k equal to k_{pref}, but perhaps much higher.

Second, these approaches do not provide the framework to explore the possibility of attaining higher effective k values without increasing the information loss. Owing to the monotonicity property, an effective k value higher than k_{pref} will also satisfy the anonymity constraint, but comes with the added advantage of better privacy. Hence, exploring solutions beyond the ones that strictly satisfy the constraint is desirable. Further, existing approaches do not take into account any preference specified on information loss. There are some successful attempts to obtain all possible k–anonymized versions of a data set [9, 15], out of which the optimal one can be chosen based on preference criteria. Nonetheless, the set of solutions obtained with such an approach still remains exponentially large, making the search for an optimal choice equally difficult to perform. The issue of privacy bias remains unexplored in all these attempts. The possibility of non-dominance characteristics between privacy bias and information loss makes the problem further difficult and interesting at the same time.

We start by introducing the notion of an *efficient solution.*

Definition 2. **(*Efficient Solution*)** Given an integer $1 \leq k_{pref} \leq n_{row}$, an anonymized data set D' is efficient if $\min(EC_{\mathsf{D}'}) \geq k_{pref}$ and there does not exist another anonymized data set D'' satisfying $\min(EC_{\mathsf{D}''}) \geq k_{pref}$ such that $NWP(\mathsf{D}'') < NWP(\mathsf{D}')$ and $NECD(\mathsf{D}'') < NECD(\mathsf{D}')$.

In other words, an efficient solution is a feasible non-dominated point in the objective space of NWP versus NECD. Preference information on NWP and NECD are embodied into the problem by using an *achievement function* [14]. The achievement function is a scalarization of NWP and NECD, dependent on a specified *preference point* $(NWP_{pref}, NECD_{pref})$. An appropriate achievement function is one which can be used to determine efficient solutions by performing a minimization of the scalar function. In this regard, we define the following function adapted from the commonly known *Chebyshev min-max* problem [13].

$$\mathbf{ach}(\mathsf{D}') = \max \begin{bmatrix} w(NWP(\mathsf{D}') + \epsilon), \\ (1-w)(NECD(\mathsf{D}') + \epsilon) \end{bmatrix}$$

$$\text{where } w = \frac{\frac{1}{NWP_{pref}+\epsilon}}{\frac{1}{NWP_{pref}+\epsilon} + \frac{1}{NECD_{pref}+\epsilon}}.$$

Here ϵ is a very small positive number. The *ideal point* in the NWP versus NECD objective space lies at $(0, 0)$ and thus $(-\epsilon, -\epsilon)$ is an *utopian point*. Minimization of **ach** results in a solution that provides the maximal overachievement beyond the preference point if it is feasibly attainable or otherwise minimal underachievement if the preference point is not attainable. The parameter w allows a directed search along the direction from the utopian point to the preference point, the desired solution being the point where the search hits the feasible region. Refer to [13] for details on how the parameter is typically used in a non-evolutionary framework.

Observation. *A minima* D' *of* **ach** *subject to the constraint* $\min(EC_{\mathsf{D}'}) \geq k_{pref}$ *is an efficient solution.*

Proof. Since D' is a minima of **ach** satisfying the constraint $\min(EC_{\mathsf{D}'}) \geq k_{pref}$, we have $\mathbf{ach}(\mathsf{D}') \leq \mathbf{ach}(\mathsf{D}'')$ for all D'' that satisfies $\min(EC_{\mathsf{D}''}) \geq k_{pref}$.

Let us assume that D' is not an efficient solution. Hence, there exists a D'' satisfying $\min(EC_{\mathsf{D}''}) \geq k_{pref}$ such that $NWP(\mathsf{D}'') < NWP(\mathsf{D}')$ and $NECD(\mathsf{D}'') < NECD(\mathsf{D}')$. Assuming that $NWP_{pref} \geq 0$ and $NECD_{pref} \geq 0$ (which is a valid assumption since the preference point should at best be the ideal point), we have $0 < w < 1$ and $0 < 1 - w < 1$. Given $\epsilon > 0$, we thus have the following two relations.

$$\begin{aligned} w(NWP(\mathsf{D}'') + \epsilon) &< w(NWP(\mathsf{D}') + \epsilon) \\ (1-w)(NECD(\mathsf{D}'') + \epsilon) &< (1-w)(NECD(\mathsf{D}') + \epsilon) \end{aligned}$$

Using the result $a < b, c < d \Rightarrow \max(a, c) < \max(b, d)$ on the above relations, we obtain $\mathbf{ach}(\mathsf{D}'') < \mathbf{ach}(\mathsf{D}')$ which is a contradiction. Therefore, D' must be an efficient solution. □

For the case when the minima of **ach** is not unique, we use a *preference deviation* metric to choose a solution amongst the multiple minima solutions.

$$pref_{dev}(\mathsf{D}') = NWP(\mathsf{D}') + NECD(\mathsf{D}') - NWP_{pref} - NECD_{pref}$$

The solution chosen from multiple minima points is the one with minimum $pref_{dev}$. This returns the solution providing the maximal overachievement or minimal underachievement in the total sum of NWP and NECD. With these components, we can now define our optimization problem in disclosure control as follows.

Optimization Problem (OP): *Given a data set* D, k_{pref}, NWP_{pref} *and* $NECD_{pref}$, *find the anonymized data set* D' *(or effectively the generalizations that induce it) that minimizes the achievement function* **ach** *subject to the constraint* $k_{pref} - \min(EC_{\mathsf{D}'}) \leq 0$.

5. A MULTI-OBJECTIVE APPROACH

The optimization problem at hand is a constrained single objective problem. In this section we propose an approach based on evolutionary multi-objective optimization to find a solution to the problem. The method involves transforming the constraint into a separate objective giving us a bi-objective vector optimization problem [2]. The multi-objective variant of OP is formulated as follows.

Multi-Objective Optimization Problem (MOOP): *Given a data set* D, k_{pref}, NWP_{pref} *and* $NECD_{pref}$, *find the anonymized data set* D' *(or effectively the generalizations that induce it) that minimizes the achievement function* $f_1(\mathsf{D}') : \mathbf{ach}(\mathsf{D}')$ *and the function* $f_2(\mathsf{D}') : k_{pref} - \min(EC_{\mathsf{D}'})$.

Solutions to the MOOP are characterized by the Pareto-dominance concept. Under such a characterization, an anonymized data set D' found by the solution methodology is a non-dominated solution to the MOOP if it cannot find another solution D'' such that

- $f_1(\mathsf{D}'') \leq f_1(\mathsf{D}')$ and $f_2(\mathsf{D}'') < f_2(\mathsf{D}')$, or
- $f_1(\mathsf{D}'') < f_1(\mathsf{D}')$ and $f_2(\mathsf{D}'') \leq f_2(\mathsf{D}')$.

A direct and positive consequence of using this formulation is the exposure of higher effective k solutions, if any. Note that a solution to OP only needs to satisfy the constraint $k_{pref} - \min(EC_{\mathsf{D}'}) \leq 0$. In the multi-objective formulation, the solutions undergo further filtering based on non-dominance – for two solutions with equal value of **ach**, the one with higher effective k (lower f_2) gets preference. Thus, if multiple solutions to OP exists at different effective k values, the multi-objective approach directs the search towards the one providing the highest level of privacy. In addition, the method exposes the trade-off characteristics between the level of privacy attainable and the **ach** function (effectively NWP and NECD). We shall use the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [3] to obtain solutions to the MOOP. Using a population based approach allows us to explore the non-dominated front using Pareto-dominance without involving external parameters such as weights on the objectives.

5.1 Solution representation

Before NSGA-II can be applied, a viable representation of a generalization has to be designed for the algorithm to work with. Here we adopt the encoding suggested by Iyengar [7]. Consider the numeric attribute age with values in the domain $[10, 90]$. Since this domain can have infinite values, the first task is to granularize the domain into a finite number of intervals. For example, a granularity level of 5 shall discretize the domain to $\{[10, 15], (15, 20], \ldots, (85, 90]\}$. Note that this is not the generalization applied on the age attribute. The discretized domain can then be numbered as $1 : [10, 15], 2 : (15, 20], \ldots, 16 : (85, 90]$. The discretized domain still maintains the same ordering as in the continuous domain. A binary string of 15 bits is now used to represent all possible generalizations for the attribute. The i^{th} bit in this string is 0 if the i^{th} and $(i+1)^{th}$ intervals are supposed to be combined, otherwise 1. The granularization step can

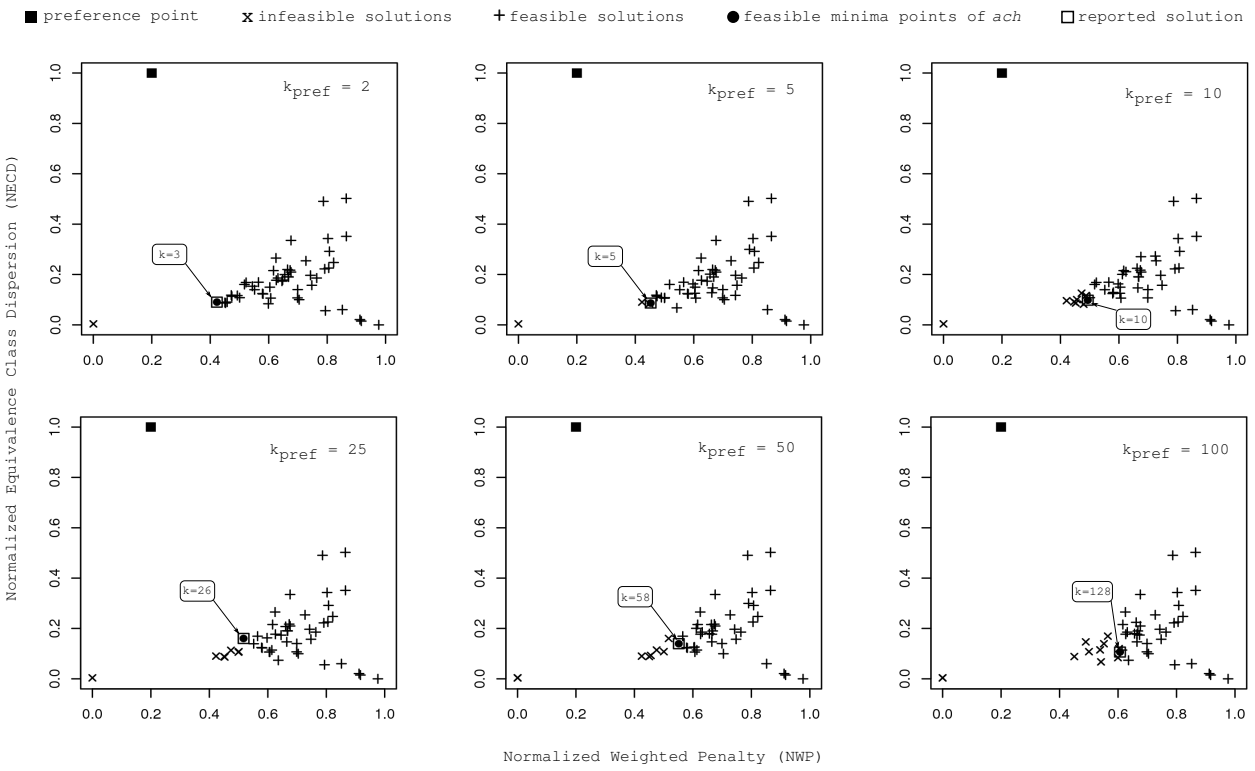

Figure 1: NWP and NECD values of non-dominated solutions returned by NSGA-II for different k_{pref}. The effective k value obtained is highlighted for the reported solution. An unique feasible minima of ach is obtained in all cases and the reported solution is an efficient point, w.r.t. the set of solutions returned by NSGA-II, in the NWP vs. NECD objective space.

be skipped for attributes with a small domain and a defined ordering of the values. The individual encodings for each attribute are concatenated to create the overall encoding of the generalizations for all attributes.

5.2 Selection operator

While most components of NSGA-II are retained in the original form, a modification is proposed for the selection procedure in order to direct solutions towards the feasible region of OP. NSGA-II employs a *crowded comparison operator* as part of its binary tournament selection scheme. This operator gives preference to solutions with lower ranks (in accordance with the non-dominated ranking scheme of the algorithm). For the case when the compared solutions are of the same rank, a *crowding distance* metric is used to choose the solution with least diversity. Our modification involves distinguishing between feasible ($f_2(\mathsf{D}') \leq 0$) and infeasible ($f_2(\mathsf{D}') > 0$) solutions of OP during the selection procedure. The procedure is outlined as follows for two solutions x and y.

1. If both x and y are feasible, select based on crowded comparison operator.
2. If x is feasible and y is not, or vice versa, select the feasible one.
3. If both x and y are infeasible:
   (a) select one with minimum f_2.
   (b) if f_2 is equal, select one with minimum f_1.
   (c) if f_1 is also equal, use crowding distance metric.
4. Any case unresolved by the above three cases is resolved by arbitrary selection.

Using this selection procedure, we can initially direct the search towards the feasible region of OP and thereafter concentrate on exploring the trade-off characteristics. Step 3 of the procedure uses a lexicographic approach to selection. Note that the space of possible solutions to the problem is increasingly dense as the effective k approaches 1, i.e. a relatively higher number of solutions are feasible for lower k_{pref}. Step 3b and 3c are particularly useful in such settings.

5.3 Solution to OP

Once the final non-dominated solution set $\mathcal{ND}$ to MOOP is obtained, the solution to OP is chosen as the point D' such that

$$\mathsf{D}' = \underset{\mathsf{D}'' \in \mathcal{ND}_f}{\operatorname{argmin}} f_1(\mathsf{D}'') \text{ where } \mathcal{ND}_f = \{\mathsf{D}_i \in \mathcal{ND} | f_2(\mathsf{D}_i) \leq 0\}.$$

The case of multiple such solutions is resolved using the preference deviation metric. Since the minima of **ach** obtained in this manner is only justifiable w.r.t. ND_f, we shall say that D' is an efficient solution only w.r.t. the non-dominated solutions generated by NSGA-II. The statement is just a cautious side note to indicate that the effectiveness of the approach is as good as the convergence and diversity preservation abilities of the multi-objective optimizer.

6. EMPIRICAL RESULTS

We applied our methodology to the "adult.data" benchmark data set [8]. The data was extracted from a census bureau database and has been extensively used in studies related to disclosure control. We prepared the data set as described in [1, 7]. All rows with missing values are removed to finally have a total of 30162 rows. The attributes used in this study along with their domain size are listed in Table 1. This gives us a chromosome of length 105 representing a solution.

Attribute	Domain Size
Age (*age*)	20 (granularity=5)
Work Class (*wkc*)	7
Education (*edc*)	16
Marital Status (*mst*)	7
Occupation (*occ*)	14
Race (*rac*)	5
Gender (*gen*)	2
Native Country (*ncy*)	41
Salary Class (*slc*)	2

Table 1: Attributes and domain size from the *adult census* data set.

For NSGA-II, the population size N_{pop} is set at 100 with a maximum of 50,000 function evaluations. Binary crossover is performed on the entire chromosome with rate 0.8. Mutation is performed on the individual encodings of every attribute with a rate of 0.001. The modified selection operator is used for binary tournament selection. Weights on the attributes are assigned equally (1/9), unless otherwise stated. Typical run time of an experiment is observed to be roughly 15 minutes on an Intel Core 2 Duo 2x2.83GHz machine with 2GB RAM and running 64-bit Fedora Core 8. No difference in obtained solutions is observed when running the experiments with different random number seeds.

6.1 Solution efficiency

Fig. 1 illustrates the NWP and NECD values of the non-dominated solutions returned by NSGA-II for different values of k_{pref}. A preference point of $(0.2, 1.0)$ is used in these experiments. Choosing a NECD preference of 1.0 effectively allows NSGA-II to look for low NWP solutions irrespective of the privacy bias they induce. As higher values of k_{pref} are used, the number of feasible solutions obtained decreases. This is likely to happen since the search space is known to be very dense for low values of k_{pref}, while solutions become rare as higher privacy requirements are enforced. Consequently, while reported solutions for $k_{pref} = 2, 5$ and 10 have an effective k close to k_{pref}, higher values are obtained for $k_{pref} = 25, 50$ and 100. However, higher information loss has to be sustained for stronger privacy requirements. An unique feasible minima of **ach** is obtained in all the cases. In confirmation to our theoretical observation, the minima point is a non-dominated point in the NWP vs. NECD objective space w.r.t other feasible solutions returned by NSGA-II. Further, the existence of solutions at effective k values higher than k_{pref} (for example $k = 3$ for $k_{pref} = 2$) strengthens our claim that the optimal solution need not always have effective k value equal to k_{pref}. Once again, the concept of Pareto-dominance helps here in discovering these solutions.

6.2 Impact of population size

Fig. 2 shows the reported solutions for three different settings of population size, $N_{pop} = 100, 250$ and 500. Notice that increasing the population size, while keeping the number of function evaluations fixed, seem to have only marginal impact on the overall solution quality. Solutions are slightly less effective in terms of the preference deviation metric for larger population size, albeit there is no logical pattern in the behavior. Larger populations typically have the potential to explore more parts of the search space. However, the absence of an uniform distribution of solutions in the search space makes this exploration difficult. Even with the capacity to carry better genetic diversity, solutions resulting from most crossover operations tend to map to similar objective values. This is further corroborated by the observation that the number of unique solutions obtained is similar irrespective of the population size used. Large populations have more duplicates which affect the convergence rate of the population. This is primarily due to the higher selective pressure of duplicate solutions which limits the exploratory capabilities of the population. Small populations and higher number of iterations is a key element in solving this problem.

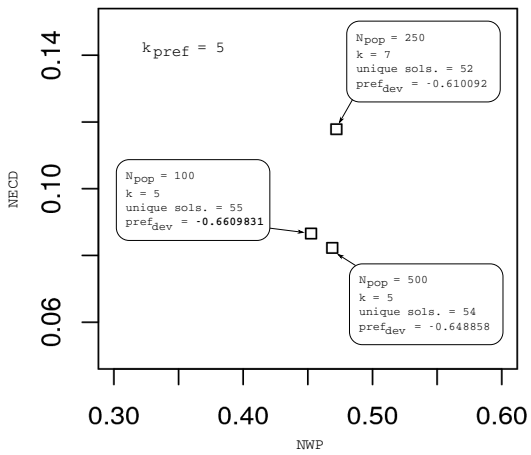

Figure 2: Impact of population size (N_{pop}) on solution quality for $k_{pref} = 5$ and preference point $(0.2, 1.0)$. The number of unique solutions in the final population is similar irrespective of population size.

6.3 Effect of weight vector

Fig. 3 illustrates the solutions obtained for different assignment of weights to the attributes. A preference point of $(0.2, 0.1)$ is used here. As is evident from the solutions, the assignment of equal weights (wv_1) in this problem results in a much higher NWP and NECD. Weight assignments impact the amount of generalization that may be performed for an attribute, which in turn influence the information content of the anonymized data set. Even when all attributes are equally important, higher weights can be assigned to attributes with larger domain sizes to retain as much information as possible. For example, while most solutions in the figure completely suppress (number of partitions=1) the "Native Country" attribute, assigning a higher weight to the attribute (as in wv_2 and wv_5) return solutions with more number of partitions. In general, NSGA-II is seemingly ef-

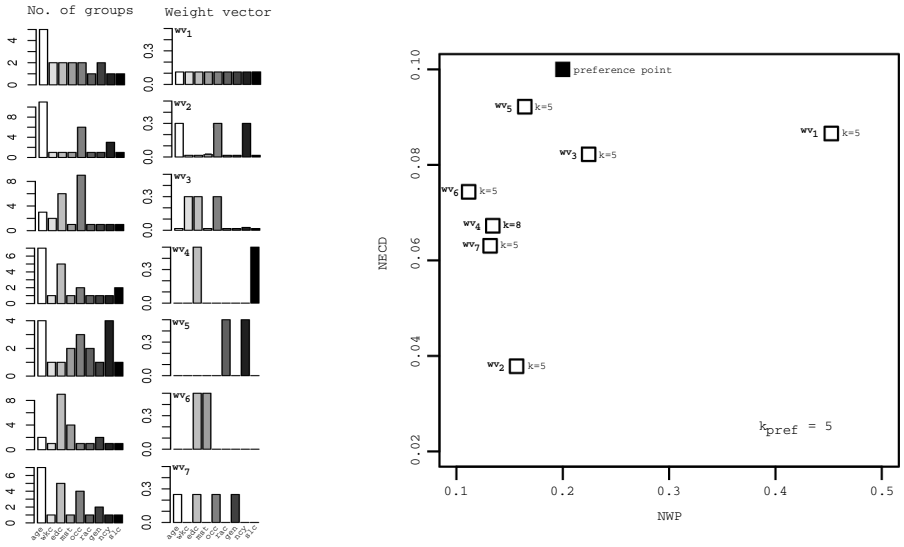

Figure 3: Effect of the weight vector on attainable NWP and NECD. Assigning equal weights, when not required, can result in less effective solutions. Weights also affect the effective k value attainable; for example, the solution for the problem with weight vector wv_4 lies at $k = 8$. Also shown are the number of groupings resulting in the attribute domains for each reported solution. Solutions maintain comparatively higher number of groups (meaning less generalization) for attributes with higher weights.

fective in generating solutions with higher number of partitions in accordance with the weight assignments. The "Age" attribute seems to have some correlation with the other attributes as the generalization performed on it is low even if most of the weight is distributed on other attributes. This experiment with weight vectors provides us with another example (wv_4) demonstrating that the optimal solution need not always be present at $k = k_{pref}$.

6.4 Impact of bias preference

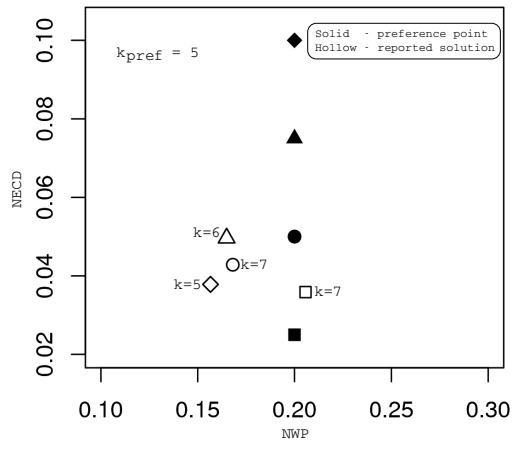

Figure 4: Impact of bias preference on solutions obtained with weight vector wv_2. Heavy bias preference allows exploration of solutions with $k = k_{pref}$. Satisfying lower bias preference requires more generalization resulting in higher effective k. Solutions satisfying the preference values exhibit a trade-off characteristic between the level of privacy and NWP.

Fig. 4 illustrates the impact of setting the NECD preference value. A typical preference of 1.0 effectively means that any level of bias is acceptable. As a result, a solution only needs to perform as much generalization as is necessary to meet the feasibility constraint, assuming that the minimum value of NWP is attained at $k = k_{pref}$. Such a case happens with the weight vector wv_2. However, when the bias preference is dropped below 0.1, solutions are generated with higher NWP (although within the preference value of 0.2) and higher effective k. This happens because the method is now forced to explore solutions with more generalization in order to better meet the low bias preference. More generalization typically yield higher effective k. Notice that as the bias preference is lowered, the effective k increases. It is imperative to ask at this point why a bias preference of 1.0 should not be set for this problem since the best solution (with $k = 5$) is obtained with this setting. The answer lies in the trade-off characteristic of the solutions between the level of privacy and NWP. Note that the $k = 6$ solution has higher NWP at the expense of slightly higher privacy level than the $k = 5$ solution. Since both solutions meet the NWP preference, the $k = 6$ solution is more preferable. In fact, given the four solutions in the figure and the NWP preference of 0.2, the $k = 7$ solution (one marked with a circle) is the solution of choice. This solution overachieves the preference criteria and provides better privacy than the $k = 5$ and $k = 6$ solutions. In general, specifying a very high bias preference may prohibit the method from exploring the trade-off characteristics between privacy level and NWP.

7. CONCLUSIONS

Privacy preserving data dissemination has to minimize the information loss in the anonymized data set while protecting the identity of underlying individuals to the maximum extent possible. Another objective while doing so is to control

the amount of privacy bias induced by the anonymity model being used. In this paper, we have argued that existing approaches address these aspects only partially in the context of the k–anonymity model. Specifically, standard approaches cannot guarantee that higher privacy levels are not possible when attributes have varying levels of significance and data publisher's preferences are taken into account.

In our approach, we propose using a preference based achievement function to scalarize the induced information loss and privacy bias into a single function, and then perform a constrained optimization on this function. We have proved that minima solutions of the proposed function are Pareto-optimal. The constraint in the optimization problem is then treated as a second objective to minimize, providing a method to improve upon the specified privacy levels, if possible. This is facilitated by solving the problem using an evolutionary multi-objective algorithm. Results on a benchmark data set demonstrate the effectiveness of the method in finding solutions that best achieve the preferences of the data publisher. The method is also able to find higher effective k values depending on the weights assigned to different attributes. Parametric studies suggest that using smaller populations have advantages in this problem structure. Furthermore, the bias preference has a direct impact on the exploration of the privacy versus loss trade-off front.

Further work will explore other potential benchmark problems. We are also looking forward to the emergence of other algorithms that address issues such as privacy bias so that a comparative study can be performed on the performance of the proposed approach.

8. ACKNOWLEDGMENTS

This work was partially supported by the U.S. Air Force Office of Scientific Research under contract FA9550-07-1-0042. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the U.S. Air Force or other federal government agencies.

9. REFERENCES

[1] Bayardo, R. J., and Agrawal, R. Data Privacy Through Optimal k-Anonymization. In *Proceedings of the 21st International Conference on Data Engineering* (2005), pp. 217–228.

[2] Coello, C. A. C. Constraint-Handling Using An Evolutionary Multiobjective Optimization Technique. *Civil Engineering and Environmental Systems 17* (2000), 319–346.

[3] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA–II. *IEEE Transactions on Evolutionary Computation 6*, 2 (2002), 182–197.

[4] Fung, B. C. M., Wang, K., and Yu, P. S. Top-Down Specialization for Information and Privacy Preservation. In *Proceedings of the 21st International Conference in Data Engineering* (2005), pp. 205–216.

[5] Golle, P. Revisiting the Uniqueness of Simple Demographics in the US Population. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society* (2006), pp. 77–80.

[6] Hundepool, A., and Willenborg, L. Mu and Tau Argus: Software for Statistical Disclosure Control. In *Proceedings of the Third International Seminar on Statistical Confidentiality* (1996).

[7] Iyengar, V. S. Transforming Data to Satisfy Privacy Constraints. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), pp. 279–288.

[8] Kohavi, R., and Becker, B. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/.

[9] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. Incognito: Efficient Full-Domain k-Anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (2005), pp. 49–60.

[10] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. Mondrian Multidimensional K-Anonymity. In *Proceedings of the 22nd International Conference in Data Engineering* (2006), p. 25.

[11] Loukides, G., and Shao, J. Capturing Data Usefulness and Privacy Protection in k-Anonymisation. In *Proceedings of the 2007 ACM Symposium on Applied Computing* (2007), pp. 370–374.

[12] Meyerson, A., and Williams, R. On the Complexity of Optimal k-Anonymity. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems* (2004), pp. 223–228.

[13] Miettinen, K., and Kirilov, L. Interactive Reference Direction Approach Using Implicit Parametrization for Nonlinear Multiobjective Optimization. *Journal of Multi-Criteria Decision Analysis 13*, 2-3 (2005), 115–123.

[14] Miettinen, K., and Mäkelä, M. M. On Scalarizing Functions in Multiobjective Optimization. *OR Spectrum 24*, 2 (2002), 193–213.

[15] Samarati, P. Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering 13*, 6 (2001), 1010–1027.

[16] Samarati, P., and Sweeney, L. Protecting Privacy when Disclosing Information: k–Anonymity and its Enforcement through Generalization and Suppression. Tech. rep., Computer Science Laboratory, SRI International, 1998.

[17] Sweeney, L. Achieving k–Anonymity Privacy Protection Using Generalization and Suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10*, 5 (2002), 571–588.

[18] Sweeney, L. k–Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10*, 5 (2002), 557–570.

[19] Wang, K., Yu, P., and Chakraborty, S. Bottom-Up Generalization: A Data Mining Solution to Privacy Protection. In *Proceedings of the 4th IEEE International Conference on Data Mining* (2004), pp. 249–256.