Exploring Privacy versus Data Quality Tradeoffs in Anonymization Techniques using Multi-objective Optimization

Rinku Dewri

Department of Computer Science, University of Denver, Denver, CO 80208, USA.

Email: rdewri@cs.du.edu

Indrajit Ray*, Indrakshi Ray and Darrell Whitley

Department of Computer Science, Colorado State University, Fort Collins, CO 80523, USA.

Email: {indrajit, iray, whitley}@cs.colostate.edu

Abstract

Data anonymization techniques have received extensive attention in the privacy research community over the past several years. Various models of privacy preservation have been proposed: k–anonymity, ℓ–diversity and t–closeness, to name a few. An oft-cited drawback of these models is that there is considerable loss in data quality arising from the use of generalization and suppression techniques. Optimization attempts in this context have so far focused on maximizing the data utility for a pre-specified level of privacy. To determine if better privacy levels are obtainable with the same level of data utility, majority of the existing formulations require exhaustive analysis. Further, the data publisher's perspective is often missed in the process. The publisher wishes to maintain a given level of data utility (since the data utility is the revenue earner) and then maximize the level of privacy within acceptable limits. In this paper, we explore this privacy versus data quality trade-off as a multi-objective optimization problem. Our goal is to provide substantial information to a data publisher about the trade-offs available between the privacy level and the information content of an anonymized data set.

Keywords: Privacy, Data utility, Multi-objective optimization, Cost-benefit analysis.

*Corresponding Author: 1873 Campus Delivery, Department of Computer Science, Colorado State University, Fort Collins, CO 80523-1873, USA. Tel. No. +1(970)491-7097 Fax No. +1(970)491-2466 Email: indrajit@cs.colostate.edu

1 Introduction

Various scientific studies, business processes and legal procedures depend on quality data. Large companies have evolved whose sole business is gathering data from various sources, building large data repositories, and then selling the data or their statistical summary for profit. Examples of such large data publishers are credit reporting agencies, financial companies, demographic data providers and so on. These data repositories often contain sensitive personal information such as medical conditions, financial status etc., which if disclosed and/or misused, can have alarming ramifications. Thus, not only the storage of this data is done with strong security controls, but also the dissemination is frequently governed by various privacy requirements and subjected to disclosure control. For privacy protection, the data need to be sanitized of personally identifying attributes before it can be shared. Anonymizing the data, however, is quite challenging. Re-identifying the values in sanitized attributes is not impossible when other publicly available information or an adversary's background knowledge can be linked with the shared data. This is also known as a *linking attack*. 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 [40].

Database researchers have worked hard over the past several years to address such privacy concerns. Earlier techniques such as scrambling and adding noise to the data values [38] address the inference problem in statistical databases without reducing the value of the data. More recently, the *k–anonymity* model has been the subject of wide-scale research in microdata disclosure control [41, 42]. The anonymizing process involves transforming the original data set into a form unrecognizable in terms of the exact data values by using *generalization* and *suppression* schemes. A generalization performs a one-way mapping of the values of personally identifiable attributes, called *quasi-identifiers*, to a form non differentiable from the original values or to a form that induces uncertainty in recognizing them. An example of this is replacing a specific age by an age range. More often than not, it may be impossible to enforce a chosen level of privacy due to the presence of outliers in the data set. Outliers are not pre-defined in a given data set. Rather they depend on the generalization scheme that one is applying on the data. Given a particular generalization, outliers may emerge, making it difficult to achieve a desired level of privacy. In such a situation, a suppression scheme gets rid of the outliers. Suppression works by removing entire tuples making them no longer existent in the data set. A transformed data set of this nature is said to be k–anonymous if each record in it is same as at least $k-1$ other records with respect to the quasi-identifiers. The higher the value of k, the stronger the privacy that the model offers.

An unavoidable consequence of performing such anonymization is a loss in the quality of the information content of the data set. Statistical inferences suffer as more and more diverse data are recoded to the same value, or records are deleted by a suppression scheme. A summary statistic relying on accurate individual information automatically deteriorates when stronger privacy is implemented. Researchers have therefore looked at different methods to obtain an optimal anonymization that results in a minimal loss of information [3, 42, 47, 51, 52]. Since deciding on an anonymization for k–anonymity is NP-hard [5] most studies so far have focused on algorithms to minimize the information loss for a fixed value of k.

As research in this front progressed, other types of attacks have also been identified — *homogeneity attack, background knowledge attack, skewness attack* and *similarity attack* [7, 39]. Models beyond k–anonymity have been proposed to counter these new forms of attacks on anonymized data sets and the hidden sensitive attributes. Two of the more well known models in this class are the *ℓ–diversity* model [7] and the *t–closeness* model [39]. While these models enable one to better guarantee the preservation of privacy in the disseminated data, they still come at a cost of reduced quality of the information.

The (possibly) unavoidable loss in data quality due to anonymizing techniques presents a dilemma to the data publisher. Since the information they provide forms the basis of their revenue, its whole purpose would be lost if the privacy controls prohibit any kind of fruitful inferences being made from the distributed data. In other words, although the organization needs to use some anonymization technique when disseminating the data, it also needs to maintain a pre-determined level of utility in the published data. Proper anonymization thus involves weighing the risk of publicly disseminated information against the statistical utility of the content. In such a situation, it is imperative that the data publisher understands the implications of setting a parameter in a privacy model (for example, k in k–anonymity or ℓ in ℓ–diversity) to a particular value. There is clearly a trade-off involved. Setting the parameter to a "very low" value impacts the privacy of individuals in the database. Picking a "very high" value disrupts the inference of any significant statistical information from the anonymized data set. Furthermore, a data publisher may at times be confronted with a choice of several values of a parameter. This will arise in situations where individuals are allowed an opportunity to specify their desired privacy levels. For example, some users may be content with $k = 2$ (in the k–anonymity model) while others may want $k = 4$. In such cases the publisher needs to determine if some higher parameter value than initially selected is (or is not) possible with the same level of information loss. If a higher value is possible it will do a bonafide service to the individuals whose personal data are in the repository.

We believe that in order to understand the impact of setting the relevant parameters, a data publisher needs to answer questions similar to the following.

1. What level of privacy can one assure given that one may not suppress any record in the data set and can only tolerate an information loss of 25% (say)?

2. What is a good value for k (assuming the k–anonymity model) when one may suppress 10% (say) of the records and be able to tolerate an information loss of (maybe) 25%?

3. Under the "linking attacks" threat model and assuming that the users of the published data sets are likely to have background knowledge about some of the individuals represented in the data set,

is it possible to combine the k–anonymity and the ℓ–diversity models to obtain a generalization that protects against the privacy problems one is worried about?

4. Is there a generalization that gives a high k and a high ℓ value if one is ready to suppress (maybe) 10% of the records and tolerate (say) 20% of information loss?

Unfortunately, answering these questions using existing techniques will require us to try out different k (or ℓ) values to determine what is suitable. Additionally, since the k–anonymity and ℓ–diversity models have been developed to address different types of attack on privacy, one may want to combine the two models. This will require more possibilities to be tried out. Further, such a methodology does not guarantee that better privacy results cannot be obtained without incurring any or an acceptable increase in the information loss. Although recent studies have looked into the development of fast algorithms to minimize the information loss for a particular anonymization technique with a given value for the corresponding parameter, we are not aware of any study that explores the data publisher's dilemma – *given an acceptable level of information loss determine the best k and/or ℓ value that satisfy the privacy requirements of the data set.*

We make five major contributions in this work. First we discuss the formulation of a series of multi-objective optimization problems, the solutions to which provide an in-depth understanding of the trade-off present between the level of privacy and the quality of the anonymized data set. We note that one important feature often overlooked in the specification of a privacy model is the distribution of the privacy parameter across the anonymized data set. The privacy parameter reported on an anonymized data set, for example k in k-anonymity, is a quantifier of the least property satisfied by all tuples in the data set. Very often, this quantity is an inexact resemblance of the privacy level, for it may be the case that a majority of the tuples in the data set actually satisfy a higher privacy property – a higher k value for example. Failure to capture the distribution of the parameter values thereby makes differentiating between two equivalent (in the sense of a privacy model parameter value and utility) anonymizations a difficult task. Techniques are therefore required to capture (or specify) this distribution and make it a part of the optimization process. As our second contribution we build on the concept of *weighted-k anonymity* (introduced earlier in [46]) and include it in one of the multi-objective problem formulations. Third, we provide an analytical discussion on the formulated problems to show how information on this trade-off behavior can be utilized to adequately answer the data publisher's questions. Fourth, we exemplify our approach by using a popular evolutionary algorithm to solve the multi-objective optimization problems relevant to this study. Our last contribution is the design of a multi-objective formulation that can be used to search for generalizations that result in acceptable adherence to more than one privacy property within acceptable utility levels. Towards this end, we show how decision making is affected when trying to use the k–anonymity and ℓ-diversity models simultaneously.

The remainder of the paper is organized as follows: Section 2 reviews some of the existing research in disclosure control. The required background on multi-objective optimization is presented in Section

3. We introduce the terminology used in the paper in Section 4. Section 5 provides a description of the four multi-objective problems we formulate and the underlying motivation behind them. The specifics of the solution methodology with respect to solving the problems using an evolutionary algorithm is given in Section 6, and a discussion of the results so obtained is presented in Section 7. Finally, Section 8 summarizes and concludes the paper.

2 Related Work

Several algorithms have been proposed to find effective k–anonymization. Ciriani et al. provide an extensive survey on the different methods that can be applied to enforce the principle [50]. 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 [3]. μ-argus suffers from the shortcoming that larger combinations of quasi-identifiers are not checked for k–anonymity and hence the property is not always guaranteed. The *Datafly* approach uses a heuristic method to generalize the attribute containing the most distinct sequence of values for a provided subset of quasi-identifiers [34]. Sequences occurring less than k times are suppressed.

Samarati introduced the concept of *full-domain generalization* where a binary search algorithm finds all k–minimal generalizations and the issue of optimality can then be later resolved based on certain preference information provided by the data recipient [42]. The *Incognito* algorithm is also motivated under such grounds [30]. The basic algorithm starts with the generalization lattice of a single attribute and performs a modified bottom-up breadth-first search to determine the possible generalized domains of the attribute that satisfy k–anonymity. Thereafter, the generalization lattice is updated to include more and more number of attributes.

Iyengar proposes a flexible generalization scheme and uses a genetic algorithm to perform k–anonymization on the larger search space that resulted from it [51]. 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. introduce 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 [37]. As another stochastic approach, Winkler propose using simulated annealing to do the optimization [52].

On the more theoretical side, Meyerson and Williams have recently proposed an approximation algorithm that achieves an anonymization with $\mathrm{O}(k \log k)$ of optimal [5]. However, the method is not suitable when larger values of k is desired.

Most of the previous approaches start from the original data set and systematically or greedily generalize it into one that is k–anonymous. Bayardo and Agrawal proposed a complete search method that iteratively constructs less generalized solutions starting from a completely generalized data set [47]. The algorithm starts with a fully generalized data set and systematically specializes it into one that is minimally k–anonymous. It uses a tree search strategy exploiting both cost-based pruning and dynamic search rearrangement [45]. The idea of a *solution cut* is presented by Fung et al. in their approach to top down specialization [8]. 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.

A more general view of k-anonymization is *clustering* with a constraint on the minimum number of objects in every cluster [18, 23, 22, 25]. Clustering techniques use metrics that quantify the distance between tuples and distance between equivalence classes. The basic idea for the algorithm is finding an arbitrary equivalence class of size smaller than k and merging it with the closest equivalence classes to form a larger equivalence class with the smallest distortion. The process is repeated recursively until each equivalence class contains at least k tuples. Minimum distortion is enforced by choosing *closest common generalizations* of attributes.

LeFevre et al. extend the notion of generalizations on attributes to generalization on tuples in the data set [32]. The authors argue that such multidimensional partitioning of the generalization domain show better performance in capturing the underlying multivariate distribution of the attributes, often advantageous in answering queries with predicates on more than just one attribute.

The drawbacks of using k–anonymity are first described by Machanavajjhala et al. [7]. They identify that k–anonymized data sets are susceptible to privacy violations when there is little diversity in the sensitive attributes of a k–anonymous equivalence class. In order to alleviate such privacy breaches, they propose the model of ℓ–diversity which obtains anonymizations with an emphasis on the diversity of sensitive attribute values on a k–anonymous equivalence class. Further work presented by Li et al. show that the ℓ–diversity model is also susceptible to certain types of attacks [39]. To this effect, they emphasize having the t–closeness property that maintains the same distribution of sensitive attribute values in an equivalence class as is present in the entire data set, with a tolerance level of t. They realize that t–closeness does not deal with identity disclosure scenarios and propose that it should be used in conjunction with k–anonymity.

k–anonymity assumes an adversary with full knowledge about the public information (quasi-identifiers) of all individuals in the microdata. This is an unrealistic assumption and is often not required. Variations have thus been proposed with relaxations on the knowledge of an adversary. If an adversary knows the public information of a single individual, then we may generalize the table so that the original public data of every individual maps to the generalized public data of at least k tuples. This is a relaxation of the k–anonymity requirement since the k tuples are not required to form an equivalence class. Such an anonymization results in $(1, k)$*–anonymity*. Another notion is that of $(k, 1)$*–anonymity* where any tuple in the anonymized data set maps to at least k tuples in the original data. When both forms of anonymity are satisfied, it is called (k, k)*–anonymity* [2]. Every k–anonymous table is (k, k)–anonymous, but the reverse is not necessarily true. (k, k)–anonymizations are secure if an adversary has knowledge on a limited number of individuals in the data set, but can be insecure if the adversary has full knowledge on all individuals. Quantified notions of adversarial knowledge is also used in *skyline privacy* [9]. A vector (ℓ, k, m) is used to say that the adversary knows ℓ sensitive values that a target individual does not have, the sensitive values of k individuals other than the target, and m individuals such that if any one has a specific sensitive value then the target also has it. Given such a representation of adversarial knowledge, a released table must guarantee multidimensional privacy such that the adversary's confidence that an individual has a certain sensitive value should not exceed a given threshold.

Combination of different privacy characteristics have also been attempted in certain models. The *(α, k)–anonymity* model extends k–anonymity such that the confidence of associating quasi-identifiers to sensitive values is limited to within α [44]. Under this model, any k–anonymization must also satisfy the *α-deassociation* requirement, i.e. the relative frequency of a given value of a sensitive attribute in an equivalence class must be less than or equal to a user-defined threshold α, where $0 < \alpha < 1$. However, the model has limitations similar to ℓ–diversity. When the frequency of sensitive values in the whole data set is not well-proportioned, the presence of some highly sensitive values with lower frequency will force α to be very close to one. Therefore a generic model, called the *complete (α, k)–anonymity,* uses different α values for different sensitive values [20]. Another model proposed to protect against attribute disclosure is *p–sensitive k–anonymity* [49]. A data set satisfies p–sensitive k–anonymity if it satisfies k–anonymity and the number of distinct values for each sensitive attribute is at least p in each equivalence class. *Extended p–sensitive k–anonymity* further enforces the requirement that no two values of a sensitive attribute in an equivalence class are descendants of a common protected node in a hierarchy tree specified over the sensitive attribute domain [1]. The hierarchy tree captures semantic relationships between possible values. A similar approach to combine k–anonymity and ℓ–diversity is adopted in the *(k, ℓ)–anonymity* model [56].

Most anonymity models focus on an universal approach to privacy where the same amount of preservation is sought for all individuals. As a consequence, such models may be offering insufficient protection to a subset of individuals while applying excessive privacy control to another subset. The notion of *personalized anonymity* eliminates such problems by performing generalizations that satisfy personal requirements [54]. Personal preferences on sensitive attributes are captured as *guarding nodes.* For example, a personal preference may allow "Flu" to be disclosed as the illness of a patient but no disclosure of a "Lung Cancer" patient or the inference that the illness is "Cancer" may be allowed. In such cases, the released microdata must limit the probability of inference of information beyond what is allowed within a threshold.

A different approach to data generalization is adopted in *Anatomy* [53]. It is a data dissemination method designed on top of ℓ–diversity with the difference that quasi-identifiers are not generalized and released in their original form. The quasi-identifiers and the sensitive attribute are released in two different tables, called the *QIT* and *ST* respectively. The tuples in a microdata are first partitioned (without generalizing the quasi-identifiers) into groups such that the ℓ–diversity property holds in every group. QIT comprises of only the quasi-identifier values of the tuples along with the group number to which a tuple belongs. ST lists, for every group, the distinct values of the sensitive attribute in the group along with a count of the number of tuples in the group that has a specific value. Since no generalization is performed on the quasi-identifiers, anatomy preserves any distributions/correlations in the attributes. Enforcing the ℓ–diversity property while grouping the tuples protects ST against homogeneity and background knowledge attacks. Besides such fragmentation, perturbation (addition of noise) is another technique that is used to anonymize data. Although data perturbation is slowly being discarded in microdata anonymization, these methods are still used in statistical disclosure control where answers to summary queries on a statistical data base should not reveal individual values. However, under the light of recent results by Dwork and Yekhanin [12], the efficacy of the technique even for statistical disclosure control has become questionable. Nonetheless, a cost-benefit analysis still remains relevant irrespective of the data modification technique used.

Differential privacy is another notion of privacy proposed by Dwork in the context of statistical databases [11]. The idea is to ensure that the addition or removal of a singe record in the database should not substantially change the outcome of any analysis. In other words, no individual faces an additional risk to privacy because of including his or her record in the database. Dwork has shown that certain noise distributions can help bound the privacy risk to within tolerance levels. McSherry and Talwar extend this principle to the case when noise distributions are not applicable in the problem domain [17].

Quantification of data utility has been approached from different perspectives by researchers. Early notion of information loss is based on the number of generalization steps one has to perform to achieve a given privacy requirement [42]. Such a method assumes that attribute domains can be progressively generalized and a partial order can be imposed on the domain of all generalizations for an attribute. For instance, ZIP codes can be generalized by dropping a digit from right to left at each generalization step. Postal addresses can be generalized to the street, then to the city, to the county, to the state, and so on. Given that such orderings can be imposed, a distance can be computed for each attribute between the microdata and a generalized version of it. The result is a distance vector with an entry for each attribute. Dominance relationships are used to determine if one distance vector is better than other – a better distance vector will have lower distance values for each attribute.

Information loss is also measured in terms of the amount of distortion in a generalized table. In a cell of a generalized table, the ratio of the domain of the value found in the cell to the height of the attribute's hierarchy reports the amount of generalization and thereby measures the cell's distortion [34]. *Precision* of a generalized table is then computed as one minus the sum of all cell distortions (normalized by the total number of cells). Generalizations based on attributes with longer generalization hierarchies typically maintain precision better than generalizations based on attributes with shorter hierarchies.

Further, hierarchies with different heights can provide different precision measures for the same table. A similar distortion measure is also used in [25].

Similar estimation of information loss is used in the *general loss metric* [51]. The general loss metric computes a normalized information loss for each data value in the generalized data set. The requirement here is that information in every column is potentially important and hence a flexible scheme to compute the loss for both numeric and categorical data is required. Consider an attribute containing categorical information that is generalized based on a hierarchy tree (e.g. Fig. 2). The generalized value of this attribute for a certain tuple corresponds to node P in the tree. If the total number of leaf nodes in the tree is M and the number of leaf nodes in the subtree rooted at node P is M_P then loss for this data value is computed as $(M_P - 1)/(M - 1)$. Similarly, if U_i and L_i are the upper and lower bounds of the interval to which a numerical attribute value gets mapped to, the loss is $(U_i - L_i)/(U - L)$, where U and L are the upper and lower bounds of the attribute. The general loss is the sum of loss over all data values. In some attempts, a summation of losses computed by these methods (interval based and tree based) is instead used [22].

A widely used loss metric, called the *discernibility metric*, assigns a penalty to each tuple based on the number of tuples in the anonymized data set that are indistinguishable to each other [47]. Thus, a tuple belonging to an equivalence class of size j is assigned a penalty of j. A suppressed tuple is assigned a penalty equal to the number of tuples in the data set. The idea behind using the size of the equivalence class as a measure of information loss is to penalize generalizations that result in equivalence classes bigger than what is required to enforce a given privacy requirement. A variant of this is to use the *normalized average cluster size* [32].

Data utility is often measured in conjunction with privacy in an attempt to combine both objectives into a single metric. A metric of such nature favors generalizations that result in maximum gain in the information entropy for each unit of anonymity loss resulting from the generalization. Methods employing such a metric progressively increase the amount of generalization, called a *bottom up generalization* approach [33], or decrease it, called a *top down specialization* approach [8], with the objective of maximizing the metric without violating the anonymity requirement. Utility assessment in these methods are motivated by the ability to perform correct classification tasks as typical in data mining applications. A classification metric is also proposed by Iyengar where a tuple is penalized if it gets suppressed or of its class label does not match the majority class label [51].

Another metric, called *usefulness*, measures utility as the average diversity in the tuples belonging to an equivalence class [19]. This measurement is similar to the general loss metric, with differences being in the treatment of interval based attribute domains. For such domains, the loss is assigned as the normalized distance between the maximum and minimum values of the attribute in the equivalence class. A complementary metric, called *protection*, uses the inverse of the tuple diversities as a measure of the privacy factor. The two metrics inherently exhibit a reciprocal relationship, useful when a data publisher wants to modulate the anonymization process towards one objective or the other.

Preliminary metrics to evaluate the effectiveness of anonymized data in answering aggregate queries have also been proposed. Quality here is derived from a normalized difference between the result of evaluating a query on the anonymous data and the result on the original data [32]. Given a totally ordered set of sensitive attribute values, these metrics measure the range of values that is encompassed in a query result [43]. The smaller the range, the higher is the query answering accuracy.

Loss metrics should not only capture the information loss caused by the generalization but also account for the importance of the different attributes. For example, given a disease analysis data set, an "age" attribute may be considered more critical than a "ZIP code" attribute. In such a case, generalizations that are able to maintain the "age" attribute more accurately should be favored. The *weighted normalized certainty penalty* metric uses a weighted sum of the loss measurements in different attributes of the data set [27]. The loss measurement is similar as in the general loss metric and the usefulness metric. Introduction of such preference characteristics indicates that the measurement of utility can be a very subjective matter after all.

The first known attempt of exploring the privacy and utility trade-offs is undertaken by Dewri et al. [46]. The work focuses on a multi-objective optimization formulation based on a model called *weighted-k anonymity*. A similar trade-off analysis is presented by Huang and Du in the problem of optimizing randomized response schemes for privacy protection [55]. While Brickell and Shmatikov argue that even modest gains in syntactic privacy will require the complete loss of data mining utility in a dataset [21], Li and Li observe that the privacy-utility tradeoff is similar to the risk-return tradeoff in financial investment, and apply concepts from Modern Portfolio Theory to choose the right tradeoff [48].

The potential problem in using the above approaches is that they are targeted towards obtaining an optimal generalization for a fixed value of k, ℓ, or t, sometimes in conjunction. Besides running the algorithms multiple times with different values of the parameter(s), no attempt is known to have been made to understand how the generalizations and the related cost metrics change with changes in the parameter values. Our work seeks to fill this gap.

3 Multi-objective Optimization

In real world scenarios, often a problem is formulated to cater to several criteria or design objectives, and a decision choice to optimize these objectives is sought for. An optimum design problem must then be solved with multiple objectives in consideration. This type of decision making falls under the broad category of multi-criteria, multi-objective, or vector optimization problem.

Multi-objective optimization differs from single-objective ones in the cardinality of the optimal set of solutions. Single-objective optimization techniques are aimed towards finding the global optima. In case of multi-objective optimization, there is no such concept of a single optimum solution. This is due to the fact that a solution that optimizes one of the objectives may not have the desired effect on the others. As a result, it is not always possible to determine an optimum that corresponds in the same way to all the objectives under consideration. Decision making under such situations thus require some domain expertise to choose from multiple trade-off solutions depending on the feasibility of implementation.

Formally we can state a multi-objective optimization problem (MOOP) in microdata disclosure control (MDC) as follows:

Definition 1 MDC MOOP: *Let $f_1, \ldots, f_M$ denote M objective functions to maximize while performing a modification of a given table* PT. *Find a generalized table* RT^* *of* PT *which optimizes the M-dimensional vector function*

$$f(\mathsf{RT}) = [f_1(\mathsf{RT}), f_2(\mathsf{RT}), \ldots, f_M(\mathsf{RT})]$$

where RT *is a generalized version of* PT.

The objective functions in this case are either related to the privacy or utility level maintained in an anonymized table. Note that the privacy level can be inferred with respect to different privacy models. Hence the number of objectives can be more than two. In order to find an optimal solution to the MDC MOOP, we must be able to compare anonymizations with respect to all the objectives in hand. However, due to the conflicting nature of the objective functions, a simple objective value comparison between two anonymizations cannot be performed. Most multi-objective algorithms thus use the concept of dominance to compare feasible solutions.

Definition 2 Dominance and Pareto-optimal set: *Given a table* PT *and M objectives to maximize, a generalized table* RT_1 *of* PT *is said to dominate another generalized table* RT_2 *of* PT *if*

1. $\forall i \in \{1, 2, \ldots, M\}$ $\quad f_i(\mathsf{RT}_1) \geq f_i(\mathsf{RT}_2)$ *and*
2. $\exists j \in \{1, 2, \ldots, M\}$ $\quad f_j(\mathsf{RT}_1) > f_j(\mathsf{RT}_2)$

RT_2 *is then said to be dominated by* RT_1*, denoted by* $\mathsf{RT}_2 \preceq \mathsf{RT}_1$. *If the two conditions do not hold,* RT_1 *and* RT_2 *are said to be non-dominated w.r.t. each other, denoted by the* $\npreceq$ *symbol. Further, all generalized tables of* PT *which are not dominated by any possible generalized version of* PT *constitutes the Pareto-optimal set.*

In other words, a Pareto-optimal solution is as good as other solutions in the Pareto-optimal set, and better than other feasible solutions outside the set. The surface generated by these solutions in the objective space is called the *Pareto-front* or *Pareto-surface*. Fig. 1 shows the Pareto-front for a hypothetical two-objective problem, with the dominance relationships between three feasible solutions.

In the context of the k–anonymity problem, the Pareto-front for the two objectives – maximize k and minimize loss – provides the decision maker an understanding of the changes in the information loss when k is varied. Consider two anonymized versions RT_1 and RT_2 of a data set, with corresponding k and $loss$ as $(k_1, loss_1)$ and $(k_2, loss_2)$ respectively. Let us assume that $k_1 < k_2$ and $loss_1 = loss_2$. A decision maker using RT_1, and unaware of RT_2, misses on the fact that a higher k value is possible without incurring any increase in the loss. A multi-objective algorithm using the dominance concept can expose this relationship between RT_1 and RT_2, namely $\mathsf{RT}_1 \preceq \mathsf{RT}_2$. As another example, consider the case with $loss_2 - loss_1 = \epsilon > 0$. RT_1 and RT_2 are then non-dominated solutions, meaning that one objective cannot be improved without degrading the other. However, if ϵ is a relatively small quantity acceptable to the decision maker, RT_2 might be preferable over RT_1. Such trade-off characteristics are not visible to the decision maker until a multi-objective analysis is carried out. Thus, the objective of the analysis is to find the Pareto-optimal set from the set of all possible anonymized versions of a given data set.

The classical way to solve a multi-objective optimization problem is to follow the preference-based approach [16]. Many methods following this approach employ a scalarization of the multiple objectives at hand [15, 31].A relative weight vector for the objectives can help reduce the problem to a single-objective instance, or impose orderings over the preference given to different objectives. However, such methods fail to provide a global picture of the choices available to the decision maker. In fact, the decision of preference has to be made before starting the optimization process. Relatively newer methods have been proposed to make the decision process more interactive.

Evolutionary algorithms for multi-objective optimization (EMO) have been extensively studied and applied to a wide spectrum of real-world problems. One of the major advantages of using evolutionary algorithms is their ability to scan through the global search space simultaneously, instead of restricting to localized regions of gradient shifts. An EMO works with a population of trial solutions trying to converge on to the Pareto-optimal set by filtering out the infeasible or dominated ones. Having multiple solutions from a single run of an EMO is not only an efficient approach but also helps a decision maker obtain an intuitive understanding of the different trade-off options available at hand. The effectiveness of an EMO is thus characterized by its ability to converge to the true Pareto-front and maintain a good distribution of solutions on the front [35, 36].

A number of algorithms have been proposed in this context [10, 29] – NPGA [24], DPGA [6], PAES [26], SPEA2 [14] and NSGA-II [28] to name a few widely referred ones. We employ the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) for the multi-objective optimization in this study. NSGA-II has gained wide popularity in the multi-objective optimization community, partly because of its efficiency in terms of the convergence and diversity of solutions obtained, and partly due to its extensive application to solve real-world problems[1]. However, we would like to highlight that the availability of an algorithm is not sufficient to apply it directly in this problem domain. As in many real world applications, our contribution comes in the form of appropriate formulations of the problem so that the algorithm can be

[1] *ISI Essential Science Indicators* fast breaking paper in engineering for February 2004: http://www.esi-topics.com/fbp/fbp-february2004.html

applied.

4 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 $\Sigma_j = \{\sigma_1, \sigma_2, \ldots, \sigma_{n_j}\}$. 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. For categorical data, obtaining an ordering requires the user to explicitly specify a hierarchy on the values. A hierarchy can be imposed based on how the values for the attribute can be grouped together. Fig. 2 shows an example hierarchy tree for the attribute "marital status". The leaf nodes in this example 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 pre-order traversal of the hierarchy tree [51].

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 and the elements in the subsets maintain the same ordering as in Σ_j. For the "age" attribute having values in the range of $[10, 90]$, a possible generalization can be $\langle \{[10, 30]\}, \{(30, 50]\}, \{(50, 70]\}, \{(70, 90]\} \rangle$. A possible generalization for the "marital status" attribute can be $\langle\{$*Not Married*$\}$, $\{$*spouse-absent*$\}$, $\{$*civ-spouse*$\}$, $\{$*AF-spouse*$\}\rangle$. It is important to note that generalizations for categorical data is dependent on how the hierarchy is specified for it. Further, generalizations are restricted to only those which respect the hierarchy. The generalization is said to be *constrained* in such a case. For example, the generalization $\langle\{$*Never Married, Divorced*$\}$, $\{$*Widowed, Separated*$\}$, $\{$*Married*$\rangle$ is not valid for marital status since the hierarchy tree specifies that the values $\{$*Divorced, Widowed, Separated*$\}$ can only be generalized as *Once-Married*, if at all.

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 at row i and column j in D by $G_j(v)$ where $G_j(v)$ gives the index of the subset to which v belongs to in the generalization G_j. Note that if a particular generalization G_j is equal to the domain of values Σ_j, all values of the corresponding attribute will be transformed to the same subset index 1, in which case all information in that attribute is lost and the *cell is suppressed.*

4.1 k–Anonymity

Tuples in D whose subset indices are equal in every column of D' can be grouped together into *QI-groups*. In other words, a QI-group collects all tuples in D that has the same generalized form for the quasi-identifiers. The k–anonymity problem is then defined as follows.

Definition 3 k–Anonymity problem: *Given a data set D, find a set of generalizations for the attributes in D such that the QI-groups induced by anonymizing D using the generalizations are all of size at least k.*

The problem can also be explained as obtaining the generalizations under which every tuple in D' is same as at least $k-1$ other tuples. Thus, a higher value of k evaluates to a lower chance of privacy breach.

4.2 ℓ–Diversity

The set of attributes can be divided into *sensitive* and *non-sensitive* classes. A sensitive attribute is one whose value must not be revealed (or get revealed) for any tuple in the data set. All other attributes are considered non-sensitive. Then, the ℓ–diversity principle says that every QI-group should contain at least ℓ "well-represented" values for a sensitive attribute [7]. The principle can be instantiated in many different forms depending on the meaning of "well-represented".

Let a_s be a sensitive attribute in a data set with the domain of values $\Sigma_s = \{\sigma_1, \sigma_2, \ldots, \sigma_{n_s}\}$. Further, let $Q_1, \ldots, Q_p$ be the QI-groups induced by a generalization. If $c(\sigma)_j$, where $\sigma \in \Sigma_s$, denotes the count of the number of tuples with the sensitive attribute value σ in Q_j, then one possible instantiation of the ℓ–diversity problem can be stated as follows.

Definition 4 ℓ–Diversity problem: *Given a data set D, find a set of generalizations for the attributes in D such that for each QI-group induced by anonymizing D using the generalizations, the relation*

$$\frac{c(\sigma)_j}{|Q_j|} \leq \frac{1}{\ell} \tag{1}$$

holds for all $\sigma \in \Sigma_s$ and $j = 1, \ldots, p$.

In other words, the ℓ–diversity property guarantees that a sensitive attribute value cannot be associated with a particular tuple with a probability more than $1/\ell$. The higher the value of ℓ, the better is the privacy. Although this instantiation underlines the essence of the ℓ–diversity principle, two other formulations have been suggested in the original work. One of them is to assure that the entropy of each QI-group defined as

$$Entropy(Q_j) = -\sum_{\sigma \in \Sigma_s} \frac{c(\sigma)_j}{|Q_j|} \log\left(\frac{c(\sigma)_j}{|Q_j|}\right) \tag{2}$$

is at least $\log \ell$. This requires that the entropy of the whole table is at least $\log \ell$ which may be difficult to ensure when few values appear more frequently than others. Hence, another instantiation makes sure that the most frequent sensitive values do not appear too frequently in a QI-group and the less frequent values do not occur rarely. This is called *recursive (c, ℓ)–diversity.* If m is the number of sensitive attribute values in a QI-group and r_i is the count of the i^{th} most frequent value in the QI-group then the QI-group is said to be recursive (c, ℓ)–diverse if $r_1 < c(r_\ell, r_{\ell+1}, \ldots, r_m)$. Every QI-group must be recursive (c, ℓ)–diverse for a table to have recursive (c, ℓ)–diversity. The nature of the instantiation is not the focus of this work but the parameters involved in it. We shall use the instantiation given in Def. 4 to demonstrate our methodology.

4.3 (k, ℓ)–Safe

At this stage, we would like to introduce the concept of a (k, ℓ)–*safe* anonymization. Along the lines of models such as general (α, k)–anonymity [44] and p–sensitive k–anonymity [49], a (k, ℓ)–safe data set incorporates the benefits of k–anonymity and ℓ–diversity in an anonymization. From a data publisher's perspective, it is meaningful since it allows one to ascertain the presence of the privacy guards obtainable from both k–anonymity and ℓ–diversity. A high k value in this case prohibits the likelihood of linking attacks, while a high ℓ value prohibits the likelihood of homogeneity and background knowledge attacks. When multiple types of attacks are possible on a data set, a data publisher would most certainly want to safeguard the published data against as many of them as possible. More formally,

Definition 5 (k, ℓ)–Safe: *An anonymized data set is (k, ℓ)–safe if it is k–anonymous and ℓ–diverse.*

The distinction between (α, k)–anonymity and (k, ℓ)–safety lies in the representation of attribute disclosure risks. (α, k)–Anonymity requires that a possible value of a sensitive attribute should appear with a relative frequency not greater than α in a QI-group, also called the α-*deassociation* property. The motivation behind this representation is to prohibit the frequent occurrence of certain highly sensitive values, such as HIV in a disease data set, in a QI-group and hence control the inference probability on these values. (k, ℓ)–Safe, on the other hand, uses an instantiation of ℓ–diversity that specifies the minimum number of distinct values of a sensitive attribute that must appear in a QI-group. This is very much similar to p–sensitive k–anonymity. Nonetheless, ℓ–diversity can very well be instantiated in conformation to α-deassociation or any of the other forms mentioned in the original work. Hence, we use the generic name (k, ℓ)–safe.

4.4 Optimal generalization

The trivial generalization $G_j = \langle \Sigma_j \rangle$, where $j = 1, \ldots, n_{col}$, can provide the highest k and the highest ℓ value. However, such a generalization results in an anonymized data set with little or no statistically significant information. It is often desired that the anonymized data set be useful for some level of statistical analysis. In such a case, the decision on the value of k (or ℓ) is subjected to a loss measurement of the information content in the anonymized data set, usually done using some metric.

An optimization problem defined for a given value of k (or ℓ) tries to find the generalizations that result in a minimal loss given by the metric.

Depending on the distribution of data values in a data set, obtaining a generalization with an acceptable loss for a given value of k (or ℓ) may or may not be possible. This happens when the data set has outliers that cannot be anonymized without overly generalizing the remaining data points. It therefore becomes a requirement that such outliers be suppressed completely in order to avoid an over-generalization. A suppressed tuple is usually considered nonexistent in the data set. The loss metric can account for this suppression in its loss measurement.

When suppression is allowed, an anonymized data set can be made k–anonymous by suppressing all tuples that belong to QI-groups of size less than k. Similar suppression methods can be used to enforce the ℓ–diversity property. The case without suppression can be modeled into the earlier scenario (with suppression) by assigning an infinite loss when suppression is performed [47]. However, it should be noted that the presence of outliers will always force the requirement for suppression, in which case the loss measurement will always become infinite. Furthermore, even though suppression is not allowed, such an approach enforces the k–anonymity (or ℓ–diversity) property by suppressing outliers. If all the data points in the data set has to stay in the anonymized data set as well, the desired privacy properties cannot be ascertained even after adopting such modeling.

We now proceed to formulate a set of four problems, the solutions to which provide an in-depth understanding of the publisher's dilemma. This is accompanied by an example that illustrates the scope of the problem and the corresponding analysis we envisage to help alleviate the problem.

5 Problem Formulation

As stated in the previous section, an optimization algorithm requires a numeric representation of the information loss associated with a particular generalization. A quantified loss value enables the optimization algorithm to compare two generalizations for their relative effectiveness. Loss (cost) metrics assign some notion of penalty to each tuple whose data values get generalized or suppressed, thereby reflecting the total information lost in the anonymization process. In this paper, we use the general loss metric proposed by Iyengar [51]. The general loss metric computes a normalized information loss for each of the data values in an anonymized data set. The assumption here is that information in every column is potentially important and hence a flexible scheme to compute the loss for both numeric and categorical data is required. Note that our use of the general loss metric is only a matter of choice. The optimization process we use is a black box method and is not affected by how information loss is measured.

5.1 Generalization loss

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 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:

$$loss(v_{i,j}) = \frac{|\Sigma_{j_{g_{i,j}}}| - 1}{|\Sigma_j| - 1} \tag{3}$$

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.

$$GL = \sum_{i=1}^{n_{row}} \sum_{j=1}^{n_{col}} loss(v_{i,j}) \tag{4}$$

5.2 Suppression loss

Although the loss due to suppression can be incorporated into the generalization loss, we decided to separate it out for the purpose of our study. When a row is suppressed, all cells in the row are suppressed irrespective of the generalization. Each cell thereby incurs a loss of one (consequence of Eq. (3)). Let n_{sup} be the number of rows to be suppressed in the data set. The suppression loss for the data set is then given as,

$$SL = n_{col} \times n_{sup} \tag{5}$$

5.3 The multi-objective problems

The multi-objective problems we formulate in this paper are intended to analyze and understand the trade-off nature of the generalization and suppression losses when k (or ℓ) is varied. A single objective optimization problem to minimize the generalization loss with a fixed k (or ℓ) will require multiple runs of the algorithm to understand this trade-off. By adopting a multi-objective approach, we can generate a fairly good approximation of the Pareto-front in a single run of the algorithm, which in turn provides us with the requisite information to make a better decision on the choice of k (or ℓ) for anonymization. In this context, we formulate a series of multi-objective problems for our analysis. Although, Problems 1, 2, and 3 are described for the k–anonymity problem, similar analysis can be carried out for ℓ–diversity as well. Problem 4 caters to (k, ℓ)–safety.

Also note that the problems under study are not intended to provide the data publisher a "best" value for the parameter(s) involved in the anonymization technique. Rather, we put forward a methodology to understand the implications of choosing a particular value for the parameter(s) in terms of the resulting privacy and the data utility. Hence, we shall often find that one or more solutions returned by the optimization process are trivially not acceptable either in terms of privacy or utility, or in some cases, both. It is not our objective to consider such solutions as degenerate and prohibit them from appearing in the solution set. After all, they are also a manifestation of the privacy-utility trade-off, which would likely be never selected as a choice by the data publisher, but still possible. For e.g., an extreme solution will correspond to a situation where every tuple in the data set belongs to its own QI-group, thereby resulting in no privacy and maximum utility. Another extremity is the case where all tuples are grouped together in a single QI-group resulting in maximum privacy but no utility. One cannot deny the fact that in the case of privacy versus utility, both of these are possible solutions. The multi-objective optimization formulations do not incorporate the required domain knowledge to identify these extremities (or other such solutions) as impractical. Only the data publisher has the requisite knowledge to make such identification and disregard such solutions. This is often a post-optimization process. Hence, focus in the solution set should be concentrated on the practical solutions reported by the method. Quite often there will be more than one, and the methodology provides the data publisher a distinctive picture of the differences arising between privacy and utility when it makes a decision to choose one solution over another.

5.3.1 Problem 1: Zero suppression

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 the first problem 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 QI-groups and the only way to merge them into a bigger one is by having more generalization. This results in more information loss which is often not acceptable to a 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 size of the QI-groups even when suppression is not allowed.

Let us define an equivalence class E_i as the set of all tuples in QI-groups of size i. Hence an arbitrary equivalence class E_i shall contain all tuples that are i–anonymous. For brevity, we shall use the term "small equivalence class" or "equivalence class with lower i" to mean equivalence classes E_is with relative smaller i than that used in the term "large equivalence class" or "equivalence class with high i". An ideal generalization would then maintain an acceptable level of loss and also keep the number of rows in equivalence classes with lower i relatively fewer than in equivalence classes with higher i. 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 QI-group.

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_i defines a privacy level for its member tuples – every tuple in the equivalence class is same as exactly $i-1$ other tuples in the same class and has a probability of breach equal to $1/i$. For an intuitive comparison, the k in k–anonymity is the smallest value of i such that E_i is non-empty. Note that the term "equivalence class" is typically used to indicate the concept of a QI-group. The notion adopted here is slightly different. Two rows in the original data set belong to the same equivalence class in the typical definition if the generalization transforms them into the same tuple. However, in this formulation, two rows belong to the same equivalence class E_i if a generalization makes them 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:

$$k_{weighted} = \frac{\sum_{i=1}^{n_{row}} (i \cdot |E_i|)}{\sum_{i=1}^{n_{row}} |E_i|} \qquad (6)$$

Recall the concept of local recoding [4] in this context. A local recoding scheme produces a k–anonymization by using an individual generalization function (instead of a global one) for each tuple in the data set. This is a more powerful scheme compared to having a single generalization function since outliers can be easily suppressed without the drawbacks of an over generalization, hence data utility can be maintained. The weighted-k–anonymity based generalization is orthogonal to this concept in certain ways. Local recoding explores the domain of generalization functions and uses multiple points in this domain to recode different subsets of the data set differently. This puts outliers in their own subset(s), thereby making it easy to enforce a given minimum QI-group size. Weighted-k–anonymity, on the other hand, works with a single generalization function and, instead of trying to enforce a fixed minimum QI-group size, it flexibly creates QI-groups of different sizes with no minimum size constraint. The outliers then must lie in smaller QI-groups in order to maximize data utility. The similarity between the two methods is that the outliers get treated differently than the rest of the data set.

The weighted-k is an estimation of the QI-group size distribution, and hence, although the chances are very rare, a high value need not always indicate that there exists no tuple in the anonymized data set appearing in its original form, i.e. in QI-groups of size one. Since the method results in an average case analysis, rather than worst case, such generalizations can appear. We present three justifications to the use of average privacy in multi-objective analysis, rather than worst case privacy.

First, the use of minimum privacy induces a very dense search space for the purpose of multi-objective optimization. Let us assume that an encoding for a candidate generalization is represented by n bits, resulting in a search space of size 2^n. Given that a data set of size N can be generalized to have possible

k values in the range of 1 (no generalization) to N (all suppressed), on the average, there are $2^n/N$ points on the search space that will map to the same k value, i.e. these points will not be distinguishable from each other on measures of privacy. Even with the modest assumption of $n = 50$ and $N = 10^6$, this average number is in the magnitude of 10^9. In addition, this average estimate is only wishful thinking. The mapping will be much denser for smaller k values. On one side, most points in the search space will induce very similar (and low) privacy levels, while on the other, finding one with significantly higher values of k will be extremely difficult. We can say that there is very little diversity in the search points when using minimum privacy as an objective to maximize. This diversity is a desired property to analyze trade-offs. Using average privacy introduces diversity since the distribution of QI-group sizes have significant avenues to be different for different possible generalizations.

Second, in the context of multi-objective optimization, we cannot exclude solutions with QI-group sizes of one since they just represent an aspect of the privacy-utility trade-off. Ideally, if another generalization with the same (or better) level of utility but without the isolated tuples exists, i.e. the tuples are embedded in QI-groups of size more than one, then it would result in a higher value of weighted-k. Moreover, such a generalization will dominate the previous one and will replace it from the solution set.

Third, conformation to worst-case privacy requirements can be achieved once the trade-off front is approximated. Enforcements of worst-case requirements such as minimum QI-group size is essential during optimization in a single objective framework. However, when performing multi-objective analysis, these enforcements are part of the post-optimization strategy. The requirement essentially helps filter out solutions from the Pareto-front approximated in the analysis. In fact, a minimum utility level can also be one such requirement.

Note that in most cases not all QI-groups with all possible sizes will be generated. Hence, certain equivalence classes will be empty. The weighted-k value provides a sufficiently good estimate of the distribution of the QI-group sizes. A high weighted-k value implies that QI-groups with bigger sizes is relatively more abundant than those with very few tuples. 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.*

5.3.2 Problem 2: Maximum allowed suppression

In this problem, we enable suppression and allow the user to specify an acceptable fraction η of the maximum suppression loss possible $(n_{row} \cdot n_{col})$. Such an approach imposes a hard limit on the number of suppressions allowed [47]. However, unlike earlier approaches, by allowing the user to specify a suppression loss limit independent of k, the optimization procedure can be made to explore the trade-off properties of k and generalization loss within the constraint of the imposed suppression loss limitation.

When suppression is allowed within a user specified limit, all tuples belonging to the equivalence classes $E_1, \ldots, E_d$ can be suppressed, where d satisfies the relation

$$\sum_{i=1}^{d}(|E_i| \cdot n_{col}) \leq \eta \cdot n_{row} \cdot n_{col} < \sum_{i=1}^{d+1}(|E_i| \cdot n_{col}). \qquad (7)$$

Thereafter, the resulting data set becomes $(d+1)$–anonymous and also satisfies the suppression loss constraint. We can now define our optimization problem as *finding the generalization that maximizes d and minimizes the generalization loss.* The problem can also be viewed as the maximization of k and minimization of GL satisfying the constraint $SL \leq \eta \cdot n_{row} \cdot n_{col}$. Note that the problem formulation allows the optimization procedure to find generalizations that create equivalence classes with lower i's of smaller size and thereby increase d.

5.3.3 Problem 3: Any suppression

The third problem is formulated as an extension of the second one where the user does not provide a maximum limit on the suppression loss. The challenge here is the computation of k, GL and SL for a generalization without having a baseline to start with. Since the three quantities are dependent on each other for their computation, it is important that we have some base k value to proceed. We adopt the weighted-k value at this point. Although not very precise, the weighted-k value provides a good estimate of the distribution of the QI-groups. If a very high weighted-k value is obtained for a generalization, then the number of tuples in small equivalence classes is sufficiently low, in which case we can suppress them. If the weighted-k value is low, then most of the tuples belong to equivalence classes with low i. In this case, a higher amount of suppression is required to achieve an acceptable k for the anonymized data set. Also, high weighted-k generally implies a high generalization loss. Such trade-off characteristics are the point of analysis in this problem.

To start with, a particular generalization's weighted-k value is first computed. Thereafter, all tuples belonging to an equivalence class E_i with $i < k_{weighted}$ are suppressed, enabling the computation of SL. This makes the k for the anonymized data set equal to at least $k_{weighted}$. The generalization loss GL is then computed from the remaining data set. The multi-objective problem is defined as *finding the generalization that maximizes $k_{weighted}$, and, minimizes GL and SL.*

5.3.4 Problem 4: (k, ℓ)–safety

This problem is motivated by the requirement that a data publisher may impose on obtaining an anonymized data set that is (k, ℓ)–safe. To formulate the problem, we define the equivalence class $E_{i,j}$. A tuple belongs to this equivalence class if it belongs to an i–anonymous and j–diverse QI-group. Next, an order of importance is imposed on the k and ℓ properties. Such an order specifies which property is more desired by the data publisher and enables us to define a total ordering on the equivalence classes $E_{i,j}$. The ordering is obtained by first arranging the equivalence classes w.r.t. an increasing value in the least desired property, and then for a given value in this property, the equivalence classes are rearranged w.r.t. an increasing value in the most desired property. For example, if the ℓ property is more desired (denoted by $k \ll \ell$), then an example total ordering could be $E_{1,1} < E_{1,2} < \ldots < E_{2,1} < E_{2,2} < \ldots$. Otherwise, the ordering would be $E_{1,1} < E_{2,1} < \ldots < E_{1,2} < E_{2,2} < \ldots$. The objective behind such an ordering is to find the first equivalence class E_{d_1,d_2} in that order such that, for a given acceptable fraction of suppression loss η,

$$k \ll \ell : n_{col} \cdot (\sum_{i=1}^{d_1-1} \sum_{j=1}^{\mathcal{L}} |E_{i,j}| + \sum_{j=1}^{d_2} |E_{d_1,j}|) > \eta \cdot n_{row} \cdot n_{col} \tag{8}$$

$$\ell \ll k : n_{col} \cdot (\sum_{i=1}^{\mathcal{K}} \sum_{j=1}^{d_2-1} |E_{i,j}| + \sum_{i=1}^{d_1} |E_{i,d_2}|) > \eta \cdot n_{row} \cdot n_{col} \tag{9}$$

where, $\mathcal{L}$ and $\mathcal{K}$ are the maximum obtainable values for ℓ and k respectively for the given data set. In other words, all tuples belonging to equivalence classes prior to E_{d_1,d_2} in the order can be suppressed without violating the suppression loss constraint. The data set then becomes (d_1, d_2)–safe and satisfies the suppression loss constraint. The multi-objective optimization problem is then defined as *finding the generalization that maximizes d_1, maximizes d_2, and minimizes GL.*

5.4 Illustrative example

Before discussing the solution methodology we would like to present a small example that illustrates the scope of these problems and type of analysis that we envisage to evolve from our solution. Consider the data tuples shown in Table 1. Obtaining a 3-anonymous generalization would require tuples 1 and 2 to be in an equivalence class with one or more of the other tuples. Given the hierarchy tree of the "marital status" attribute (see Fig. 2), such a generalization would result in the suppression of the attribute, since the only ancestor node common to the values is the root node of the hierarchy tree. Tuples 1 and 2 act as outliers in this case which prohibit obtaining 3-anonymity without overly generalizing the "marital status" attribute.

Table 2 shows two different 2-anonymous generalizations possible on the data set. The only difference between the two anonymizations is the extra information present in $T2$ regarding the age range of the married individuals. Let us assume that such information is not pertinent in the utility measure, and hence both anonymizations have the same utility value. As a result, an algorithm searching for an anonymization based on 2-anonymity can return either of $T1$ or $T2$. However, note that anonymization $T1$ offers better privacy preservation when tuples 3 through 7 in the original data set are concerned. Although both anonymizations are 2-anonymous, which is actually the least privacy level in all QI-groups, there does exist groups with higher k values. In $T2$, tuples $3, 4$ and 5 are associated with probability $\frac{1}{3}$ of re-identification, while tuples 6 and 7 are associated with a probability $\frac{1}{2}$. This probability is $\frac{1}{5}$

for all tuples 3 through 7 in $T1$. Such distinctions are not visible when generalizations are sought only to satisfy a fixed minimum size for all QI-groups, ignoring the actual distribution of the sizes induced. We use the weighted-k as a measure of this distribution. The weighted-k in $T1$ evaluates to $\frac{29}{7}$, while that in $T2$ evaluates to $\frac{17}{7}$, clearly marking that $T1$ has better privacy preserving potential than $T2$. If utility in both anonymizations is same then we shall have $T1 \preceq T2$. Otherwise, there is trade-off present between the privacy and utility factors in the two anonymizations.

Further distinction between $T1$ and $T2$ can be made based on the ℓ-diversity property. If "marital status" is considered a sensitive attribute, then QI-groups comprising tuples $\{3, 4, 5\}$ and $\{6, 7\}$ in $T2$ has an ℓ value of one. This make $T2$ 1-diverse, which is equivalent to the ℓ-diversity property being non-existent. $T1$, on the other hand, maintains an ℓ value of 2. Thus, $T2$ is $(2, 1)$-safe whereas $T1$ is $(2, 2)$-safe. If both anonymizations have the same utility, any optimization aimed at generating a 2-anonymous generalization alone will show no preference to $T1$. By adding the ℓ-diversity property as another objective in the (k, ℓ)-safe optimization process, we can help expose the dominance power of $T1$ w.r.t. both identity and attribute disclosure risks.

6 Solution Methodology

Classical approaches developed to handle multiple objectives concentrated on transforming the multi-objective problem into a special form of a single objective problem formulated using certain user-based preferences. However, because of the trade-off nature of multi-objective solutions, the quality of a solution obtained from a transformed single objective problem is contingent on the user-defined parameters. Evolutionary algorithms for multi-objective optimization are *multi-point methods* usually working with a population of solutions and concentrate on obtaining multiple optimal solutions in a single run. We thus employ the NSGA-II [28] algorithm to solve the multi-objective problems defined in the previous section.

6.1 Solution encoding

Before NSGA-II can be applied, a viable representation of the generalization has to be designed for the algorithm to work with. Here we adopt the encoding suggested by Iyengar [51]. 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 used to anonymize the data set. 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 can now be 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. For attributes with a small domain and a defined ordering of the values, the granularization step can be skipped. For categorical data, a similar encoding can be obtained once an ordering on the domain values is imposed as discussed in Section 4. Fig. 3 shows an example generalization encoding for a "workclass" attribute. The individual encoding for each attribute is concatenated to create the overall encoding for a generalization involving all attributes.

6.2 NSGA-II

Similar to a simple genetic algorithm [13], NSGA-II starts with a population P_0 of N random generalizations. A generation index $t = 0, 1, \ldots, Gen_{MAX}$ keeps track of the number of iterations of the algorithm. Each trial generalization is used to create the anonymized data set and the corresponding values of the quantities to be optimized are calculated. Each generation of NSGA-II then proceeds as follows. An offspring population Q_t is first created from the parent population P_t by applying the usual genetic operations of selection, crossover and mutation [13]. For constrained attributes, a special crossover operator is used as discussed in the next subsection. The offspring population also gets evaluated. The parent and offspring populations are then combined to form a population $R_t = P_t \cup Q_t$ of size $2N$. A non-dominated sorting is applied to R_t to rank each solution based on the number of solutions that dominate it. Rank 1 solutions are all non-dominated solutions in the population. A rank r solution is only dominated by solutions of lower ranks.

The population P_{t+1} is generated by selecting N solutions from R_t. The preference of a solution is decided based on its rank; lower the rank, higher the preference. By combining the parent and offspring population, and selecting from them using a non-dominance ranking, NSGA-II implements an elite-preservation strategy where the best solutions obtained are always passed on to the next generation. However, since not all solutions from R_t can be accommodated in P_{t+1}, a choice is likely to be made when the number of solutions of the currently considered rank is more than the remaining positions in P_{t+1}. Instead of making an arbitrary choice, NSGA-II uses an explicit diversity-preservation mechanism. The mechanism, based on a *crowding distance metric* [28], gives more preference to a solution with a lesser density of solutions surrounding it, thereby enforcing diversity in the population. The NSGA-II crowding distance metric for a solution is the sum of the average side-lengths of the cuboid generated by its neighboring solutions. Fig. 4 depicts a single generation of the algorithm. For a problem with M objectives, the overall complexity of one generation of NSGA-II is $O(MN^2)$.

6.3 Crossover for constrained attributes

The usual single point crossover operator in a genetic algorithm randomly chooses a crossover point and creates two offspring by combining parts of the bit string before and after the crossover point from two different parents. As shown in Fig. 5 (left), such an operation can result in an invalid generalization for constrained attributes. Iyengar proposed modifying such invalid generalizations to the nearest valid generalization [51]. However, finding the nearest valid generalization can be time consuming, besides destroying the properties on which the crossover operator is based on. In this regard, Lunacek et al. proposed a special crossover operator that always create valid offspring for constrained attributes [37]. Instead of randomly choosing a crossover point, their operator forces the crossover point to be chosen at a location where the bit value is one for both parents. By doing so, both parts (before and after the crossover point) of both parents can be guaranteed to be valid generalizations individually, which can then be combined without destroying the hierarchy requirement. Fig. 5 (right) shows an instance of this operator.

6.4 Population initialization

In order to be able to use Lunacek et al.'s crossover operator, the validity of the parent solutions must be guaranteed. This implies that the initial population that NSGA-II starts with must contain all valid generalizations for the constrained attributes. For a given hierarchy tree, we use the following algorithm to generate valid generalizations for the constrained attributes in the initial population.

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 a 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.

6.5 Experimental setup

We applied our methodology to the "adult.data" benchmark data set available from the UCI machine learning repository[2]. The data was extracted from a census bureau database and has been extensively used in studies related to k–anonymization. We prepared the data set as described in [47, 51]. All rows with missing values are removed from the data set to finally have a total of 30162 rows. The attributes "age", "education", "race", "gender" and "salary class" are kept unconstrained, while the attributes "workclass", "marital status", "occupation" and "native country" are constrained by defining a hierarchy tree on them. The remaining attributes in the data set are ignored. For Problem 4, the occupation attribute is considered sensitive.

For NSGA-II, we set the population size as 200 for Problem 1 and 2, and 500 for Problem 3 and 4. The maximum number of iterations is set 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

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

as follow: crossover rate = 0.9, mutation rate = 0.1 with binary tournament selection. We ran the algorithm with different initial populations but did not notice any significant difference in the solutions obtained. The results reported here are from one such run.

7 Results and Discussion

Before presenting our results and the analysis, we would like to emphasize that the rationale behind doing the multi-objective analysis is not to come up with a way of determining the best possible value of a model parameter. Our intention is focused at providing a global perspective of what values of the parameter are possible at different levels of data utility. The final choice of a solution depends on other feasibility criteria as well, for example, if the parameter value found at a particular utility level is acceptable to the human subjects involved or not. An inherent human factor (the data publisher or the human subjects) is thus involved in the selection of a final solution. Further, the use of NSGA-II may raise questions on whether the obtained solutions are optimal. It is possible that another algorithm, or a different metric, provides better solutions. However, our problem formulation neither has any dependency on the methodology chosen to solve them nor is particular to the loss metric used. Further, we want to emphasize that the solutions generated by the NSGA-II implementation are *valid* at each iteration of the algorithm owing to the approach we undertake in formulating the problems. For example, a solution always gives a generalization resulting in $(d+1)$–anonymity in Problem 2, and, d_1–anonymous and d_2–diverse in Problem 4. Of course the objective is to maximize these quantities along with the minimization of the information loss.

The parameters associated with NSGA-II did not have any significant effect on the quality of the solutions obtained. We believe that the special crossover operator provides a much faster rate of convergence as compared to the genetic algorithm implementation by Iyengar [51]. The following results are obtained from the standard settings as mentioned in the previous section.

The term *loss* in the following discussion signify the total information loss as a result of generalization and suppression, i.e. $loss = GL + SL$. The differentiation between GL and SL is made wherever appropriate.

7.1 Problem 1: Zero suppression

Fig. 6 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 QI-group sizes on the anonymized data set. As expected, the generalization loss increases as the distribution gets more inclined towards bigger sizes. In the absence of suppression, a single group size is often hard to enforce for all tuples in the data set. Thus, a solution here results in tuples being distributed in QI-groups of varying sizes. A higher k-weighted value signifies that most tuples belong to QI-groups of relatively much bigger sizes, in which case, the generalization loss is higher. A solution with low k-weighted value results in tuples being distributed mostly in small QI-groups.

The inset figures in the plot depict the cumulative distribution of the number of tuples belonging to an equivalence class E_i (y-axis) with respect to different i 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 for a larger section of the data set. Low generalization losses are only possible when most tuples belong to equivalence classes of lower i 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 i values exist for the solution. The other solution shows an abrupt increase signifying that the tuples either belong to equivalence classes with very small i or ones with very large i. The sought balance in the distribution may therefore exist with an acceptable level of generalization loss.

7.2 Problem 2: Maximum allowed suppression

Fig. 7 shows the trade-off between k and *loss* in Problem 2 when a maximum of 10% suppression loss is allowed. Recall that the k value in this problem is $d + 1$. The top-leftmost plot shows all the solutions obtained for the problem. Each subsequent plot (follow arrows) is a magnification of the steepest part in the previous plot. Each plot shows the presence of locally flat regions where a substantial increase in the k value does not have a comparatively high increase in the *loss*. These regions can be of interest to a data publisher since it allows one to provide higher levels of data privacy without compromising much on the information content. Also, since the solutions corresponding to these flat regions evaluate to distantly separated k values, an analysis based on a single objective formulation with a fixed k shall require a much higher number of runs of the algorithm to identify such trade-off characteristics.

Interestingly, the trend of the solutions is similar in each plot. The existence of such repeated characteristics on the non-dominated front suggests that a data publisher's choice of a specific k, no matter how big or small, can have avenues for improvement, specially when the choice falls in the locally flat regions. A choice of k made on the rising parts of the front is seemingly not a good choice since the user would then be paying a high cost in degraded data quality without getting much improvement on the privacy factor. The rational decision choice in such a case would be to lower the k value to a flat region of the front. We observed similar trends in the solutions when the suppression loss was reduced to a low 1%.

7.3 Problem 3: Any suppression

The trade-off characteristics in Problem 3 are depicted in Fig. 8. Preliminary observations from the plot indicate that an increase in generalization loss results in a decrease in the suppression loss. A similar trend is observed when the k value increases. Since the k values in this case are computed directly from the weighted-k, an explanation for these observations is possible. A high generalization loss signifies that most tuples in the data set belong to large equivalence classes, thereby inducing a high weighted-k. This implies a low accumulation in suppression loss resulting from the deletion of tuples in equivalence classes with $i < k_{weighted}$. Also, as $k_{weighted}$ increases, the equivalence class size distribution incline more towards the ones with high i values resulting in lesser number of tuples available for suppression.

The benefit of solving this problem comes in the form of an approximate solution set available for first-level analysis. For example, Fig. 9 (left) shows the solutions from the set when the suppression loss is set at a maximum allowable limit of 20%. Although GL and SL are conflicting objectives here, the analysis is intended to see if an acceptable level of balance can be obtained between the two with a reasonably good value of k. The encircled region in the plot show that three solutions around the point ($k = 5000, GL = 35\%, SL = 17\%$) are available in this case, and hence a more specific analysis can be performed. A similar solution is found when the analysis is performed by setting the generalization loss limit to 30% (Fig. 9 (right)).

7.4 Problem 4: (k, ℓ)–safety

Fig. 10 depicts a subset of the solutions obtained for Problem 4. The solutions correspond to a suppression loss limit set as $\eta = 0.1$. Further, the plots only show solutions for which the $loss$ is less than 30%. The existence of multiple solutions for a fixed value of ℓ (or k) signifies that there is a trade-off involved in the amount of information loss and the value of k (or ℓ). An observation to make in here is the number of solutions obtained for varying values of k (or ℓ). When the preference is inclined towards the k–anonymity property, the solutions obtained give more choices for the parameter k than ℓ (Fig. 10 (left)). Similarly, when preference ordering is changed to $k \ll \ell$ (Fig. 10 (right)), more choices are available for the ℓ parameter. More importantly, any solution in either of the the two plots satisfy the 30% constraint on $loss$ and hence is a viable solution to a data publisher's request with the same constraints. To choose a single solution, we can follow the same preference ordering as was used while defining the optimization problem. For example, if the ℓ–diversity property is more desirable, we can choose the solution with the highest value of k from the set of solutions with the highest value of ℓ (a $(15, 7)$–safe solution in the plot).

We can extend our analysis to see if improvements are obtainable without much increase in the information loss. Often, the data publisher's constraint on the information loss is specified without an understanding of the trade-offs possible. A multi-objective analysis reveals the nature of these trade-offs among different objectives and can provide suggestions on how one might be improved. For example, Fig. 11 (left) shows solutions when the data publisher has given a 20% constraint on the information $loss$. For a good balance between the two desired privacy properties, say the data publisher chooses the $(20, 4)$–safe solution. Solutions with $\ell = 3, 5$, or 6 are avoided at this point because of the low value of k associated with them. Fig. 11 (right) shows the solutions to the same problem but with a slightly higher *loss* limit of 21%. The encircled solutions depict the new choices that are now revealed to the data publisher. For a small amount of increase in the *loss* limit, the data publisher now has a choice – $(20, 5)$–safe – which offers the same value of k as in the old choice, but with a higher value of ℓ. Further, the earlier reluctance to choose a solution with $\ell = 3$ is reduced after the revelation of the $(22, 3)$–safe solution. In fact, there is even a candidate solution for $\ell = 6$ and a much better value in k. With this information, the data publisher now has some idea about the trade-offs possible between privacy and information loss. In fact, the trade-off analysis in this case may motivate the data publisher to relax the loss constraint, which is not a big relaxation in itself, and reap the benefits of better privacy.

7.5 Resolving the dilemma

It is possible to analyze the trade-off surface generated for a particular data set and provide the data publisher with an analytical form for it. Given that an approximation of the Pareto-front is known, an analytical form can be derived through a polynomial curve fitting approach. However, it must be noted that analyzing the privacy-utility trade-off in this manner is rather problem specific. It cannot be ascertained that the Pareto-front always has a definitive structure for all data sets. Any theoretical analysis motivated from Pareto behavior in a data set is limited to that particular data set, and is not directly extensible to another. We cannot say for sure if the Pareto front will be similar for different data sets with similar distributions. Understanding the trade-off is rather empirical in this work, represented directly in terms of the parameter values and the resulting utility (represented by the value of the utility metric of choice). Nonetheless, it is not always true that empirical results are just tuples of $\langle$privacy,utility$\rangle$ values without providing much insight into the trade-off involved. Observing the Pareto-front on a graphical plot can reveal underlying traits. A classic case of this is seen in Fig. 7. Our observations indicate here that a major part of the Pareto-front is flat, or steep, in this data set. This signify that the data publisher has more flexibility in choosing a k value for a given level of utility. The steep jumps in utility signify that there exist points for which utility can often be improved significantly with a slight deterioration in privacy.

To summarize the above discussion, we go back to the questions asked by the data publisher in Section 1 and try to provide answers to them w.r.t. the benchmark data set.

1. We can generalize the data set in such a way that more number of tuples have a low probability of being identified by a linking attack. There is a generalization that results in 22% loss and attains a weighted average k value of 2528 (from Fig. 6). The inset figure shows that a substantial fraction of the tuples belong to sufficiently big QI-groups.

2. For the constraints given, a generalization with $k = 14$ is known (from Fig. 7). However, if the information loss constraint can be relaxed to 26%, a solution with $k = 36$ is known. Note that

analysis of the nature performed in Problem 3 can be used to provide further suggestions on the trade-offs available for suppression loss.

3. A (k, ℓ)–safe solution can provide the benefits of both k–anonymity and ℓ–diversity. A generalization with a high value of k and ℓ can be an answer. However, it is required that the more desired property be specified for better analysis.

4. For the given constraints, and assuming that ℓ–diversity is more desired, a solution with $k = 20$ and $\ell = 4$ offers a good balance between the two privacy measures (from Fig. 11). There are other solutions with trade-offs in the k and ℓ values. However, if the information loss constraint is relaxed to 21%, the ℓ value can be increased to 5. Besides, this will also allow two additional solutions: $(22, 3)$–safe and $(18, 6)$–safe.

8 Conclusions

In this paper, we investigate the problem of data privacy preservation from the perspective of a data publisher who must guarantee a specific level of utility on the disseminated data in addition to preserving the privacy of individuals represented in the data set. The conflicting nature of the two objectives motivate us to perform a multi-objective analysis on the problem, the solutions to which present a data publisher with the knowledge of different trade-off properties existent between the privacy and utility factors.

We present empirical results to demonstrate that the choice of the parameter value in the k–anonymity problem can be made in an informed manner rather than arbitrarily. The multi-objective problems are formulated to cater to differing requirements of a decision maker, primarily focused on the maximization of the k value and minimization of the losses.

For generalizations without suppression, a 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 data set without compromising much on its information content. When suppression is allowed up to a hard limit, the user's choice of k should be based on an analysis similar to that performed in Problem 2. Typically, the nature of the non-dominated solution set provides invaluable information on whether an anonymization exists to improve a particular value of the model parameter without much degradation in quality of the data. First-level explorations in this context can begin with gaining an overall understanding of the trade-off characteristics in the search space.

Our results also indicate that different privacy models can be combined and optimized to result in minimal information loss. However, the trade-off picture is better portrayed in cases when the model parameters are kept separated and formulated as multiple objectives. We use (k, ℓ)–safety as the combination of k–anonymity and ℓ–diversity models with the objective of demonstrating how a multi-objective formulation can be devised to search for generalizations that result in acceptable adherence to more than one privacy property and within acceptable utility levels. Trade-off analysis on the empirical results lends credence to the fact that often a data publisher's choice of a particular solution can be augmented with information to suggest possible improvements on one or more objectives.

The formulations presented in the paper also address the data publisher's dilemma. They provide a methodology to analyze the problem of data anonymization in manners that appeal to the actual entity that disseminates the data. We believe that such an analysis not only reinstates the data publisher's confidence in its choice of a particular privacy model parameter, 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 direction can start with examination of the framework with other models of privacy preservation. Real valued parametrization of t makes the t–closeness model an interesting subsequent candidate. Hybrid models catering to different forms of attacks are also required. Work on this can begin with an exploration of what trade-offs are generated when looking for the existence of two, or more, privacy properties simultaneously. We believe that transitioning these different models into the real world requires us to synchronize our perspective of the problem with those that actually deal with it.

9 Acknowledgements

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.

References

[1] A. Campman and T. M. Truta, Extended p-Sensitive k-Anonymity, *Studia Universitatis Babes-Bolyai Informatica*, **51**:2 (2006), 19–30.

[2] A. Gionis, A. Mazza, and T. Tassa, k-Anonymization Revisited, in: *Proceedings of the 24th International Conference on Data Engineering*, 2008, pp. 744–753.

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

[4] A. Takemura, *Local Recoding by Maximum Weight Matching for Disclosure Control of Microdata Sets.* CIRJE F-Series CIRJE-F-40, CIRJE, Faculty of Economics, University of Tokyo, 1999.

[5] A. Meyerson and R. Williams, 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.

[6] A. Osyczka and S. Kundu, A New Method to Solve Generalized Multicriteria Optimization Problems Using the Simple Genetic Algorithm, *Structural Optimization*, **10**:2 (1995), 94–99.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, ℓ–Diversity: Privacy Beyond k–Anonymity, in: *ICDE 2006: Proceedings of the 22nd International Conference on Data Engineering*, 2006, p. 24.

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

[9] B. Chen, K. LeFevre, and R. Ramakrishnan, Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge, in: *Proceedings of the 33rd International Conference on Very Large Data Bases*, 2007, pp. 770–781.

[10] C. A. C. Coello, An Updated Survey of GA-Based Multiobjective Optimization Techniques, *ACM Computing Surveys*, **32**:2 (2000), 109–143.

[11] C. Dwork, Differential Privacy, *Automata, Languages and Programming*, **4052** (2006), 1–12.

[12] C. Dwork and S. Yekhanin, New Efficient Attacks on Statistical Disclosure Control Mechanisms, in: *Proceedings of the 28th Annual Conference on Cryptology*, 2008, pp. 469–480.

[13] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, 1989.

[14] E. Zitzler, M. Laumanns, and L. Thiele, SPEA2: Improving the strength pareto evolutionary algorithm, in: *Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Probelms*, 2002, pp. 95–100.

[15] E. Miglierina and E. Molho, Scalarization and Stability in Vector Optimization, *Journal of Optimization Theory and Applications*, **114**:3 (2002), 657–670.

[16] E. Triantaphllou, *Multi-Criteria Decision Making Methods.* Springer, 2000.

[17] F. McSherry and K. Talwar, Mechanism Design via Differential Privacy, in: *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, 2007, pp. 94–103.

[18] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, Achieving Anonymity Via Clustering, in: *Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2006, pp. 153–162.

[19] G. Loukides and J. Shao, Capturing Data Usefulness and Privacy Protection in K-Anonymisation, in: *Proceedings of the 2007 ACM Symposium on Applied Computing*, 2007, pp. 370–374.

[20] H. Jian-min, Y. Hui-qun, Y. Juan, and C. Ting-ting, A Complete (α,k)-Anonymity Model for Sensitive Values Individuation Preservation, in: *Proceedings of the International Symposium on Electronic Commerce and Security*, 2008, pp. 318–323.

[21] J. Brickell and V. Shmatikov, The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing, in: *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 70–78.

[22] J. W. Byun, A. Karma, E. Bertino, and N. Li, *Efficient k-Anonymization Using Clustering Techniques*, in: Advances in Databases: Concepts, Systems and Applications, Springer Berlin/Heidelberg, 2007, pp. 188–200.

[23] J. Domingo-Ferrer and J. M. Mateo-Sanz, Practical Data-Oriented Microaggregation for Statistical Disclosure Control, *IEEE Transactions on Knowledge and Data Engineering*, **14**:1 (2002), 189–201.

[24] J. Horn, N. Nafploitis, and D. Goldberg, A Niched Pareto Genetic Algorithm for Multiobjective Optimization, in: *Proceedings of the First IEEE Conference on Evolutionary Computation*, 1994, pp. 82–87.

[25] J. Li, R. C. Wong, A. W. Fu, and J. Pei, Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures, in: *Proceedings of 8th International Conference on Data Warehousing and Knowledge Discovery*, 2006, pp. 405–416.

[26] J. D. Knowles and D. W. Corne, Approximating the Non-dominated Front using Pareto Archived Evolution Strategy, *Evolutionary Computation*, **8**:2 (2000), 149–172.

[27] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu, Utility-Based Anonymization Using Local Recodings, in: *Proceedings of the 12th Annual SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 785–790.

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

[29] K. Deb, *Multi-objective Optimization Using Evolutionary Algorithms.* John Wiley & Sons Inc., 2001.

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

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

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

[33] K. Wang, P. Yu, and S. Chakraborty, 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.

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

[35] M. P. Hanse and A. Jaszkiewicz, *Evaluating the Quality of Approximations of the Non-dominated Set.* IMM Technical Report IMM-REP-1998-7, Institute of Mathematical Modeling, Technical University of Denmark, 1998.

[36] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler, Combining Convergence and Diversity in Evolutionary Multi-objective Optimization, *Evolutionary Computation*, **10**:3 (2002), 263–282.

[37] 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*, 2006, pp. 1713–1720.

[38] N. R. Adam and J. C. Wortman, Security-control Methods for Statistical Databases – A Comparative Study, *ACM Computing Surveys*, **21**:4 (1989), 515–556.

[39] N. Li, T. Li, and S. Venkatasubramanian, t–Closeness: Privacy Beyond k–Anonymity and ℓ–Diversity, in: *ICDE 2007: Proceedings of the 23rd International Conference on Data Engineering*, 2007, pp. 106–115.

[40] P. Golle, 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.

[41] P. Samarati and L. Sweeney, Generalizing Data to Provide Anonymity when Disclosing Information, in: *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 1998.

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

[43] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, Aggregate Query Answering on Anonymized Tables, in: *Proceedings of the 23rd International Conference on Data Engineering*, 2007, pp. 116–125.

[44] R. C. Wong, J. Li, A. Fu, and K. Wang, (α,k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing, in: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 754–759.

[45] R. J. Bayardo, R. Agrawal, and D. Gunopulos, Constraint-Based Rule Mining in Large, Dense Databases, in: *Proceedings of the 15th International Conference on Data Engineering*, 1999, pp. 188–197.

[46] R. Dewri, I. Ray, I. Ray, and D. Whitley, On the Optimal Selection of k in the k-Anonymity Problem, in: *Proceedings of the 24th International Conference on Data Engineering*, 2008, pp. 1364–1366.

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

[48] T. Li and N. Li, On the Tradeoff Between Privacy and Utility in Data Publishing, in: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, 517–526.

[49] T. M. Truta and B. Vinay, Privacy Protection: p-Sensitive k-Anonymity Property, in: *Proceedings of the 22nd International Conference on Data Engineering Workshops*, 2006, p. 94.

[50] V. Ciriani, S. D. C. di Vimercati, S. Foresti, and P. Samarati, k-Anonymity, *Secure Data Management in Decentralized Systems*, **33** (2007), 323–353.

[51] V. S. Iyengar, 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.

[52] W. Winkler, *Using Simulated Annealing for k–Anonymity.* Technical report, US Census Bureau Statistical Research Division, 2002.

[53] X. Xiao and Y. Tao, Anatomy: Simple and Effective Privacy Preservation, in: *Proceedings of the 32nd International Conference on Very Large Data Bases*, 2006, pp. 139–150.

[54] X. Xiao and Y. Tao, Personalized Privacy Preservation, in: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, 2006, pp. 229–240.

[55] Z. Huang and W. Du, OptRR: Optimizing Randomized Response Schemes for Privacy-Preserving Data Mining, in: *Proceedings of the 24th International Conference on Data Engineering*, 2008, pp. 705–714.

[56] Z. Li, G. Zhan, and X. Ye, Towards an Anti-inference (k,ℓ)-Anonymity Model with Value Association Rules, in: *Proceedings of the Database and Expert Systems Applications*, 2006, pp. 883–893.

Tuple ID	*Age*	*Marital Status*
1	15	Never Married
2	17	Never Married
3	20	Civ-Spouse
4	26	AF-Spouse
5	28	AF-Spouse
6	30	Civ-Spouse
7	30	AF-Spouse

Table 1: Original data set T.

$T1$:

ID	*Age*	*Marital Status*
1	10-19	Not Married
2	10-19	Not Married
3	20-39	Married
4	20-39	Married
5	20-39	Married
6	20-39	Married
7	20-39	Married

$T2$:

ID	*Age*	*Marital Status*
1	10-19	Not Married
2	10-19	Not Married
3	20-29	Married
4	20-29	Married
5	20-29	Married
6	30-39	Married
7	30-39	Married

Table 2: 2-anonymous generalization of original data set T.

Fig. 1: Pareto-front for a hypothetical two-objective problem.

Fig. 2: Hierarchy tree for the *marital status* attribute. Numbering on the leaf nodes indicate their ordering in $\Sigma_{marital\ status}$.

Fig. 3: Example generalization encoding for the *workclass* constrained attribute.

Fig. 4: One generation of NSGA-II.

Fig. 5: Usual single point crossover (left) and special crossover for constrained attributes (right).

Fig. 6: Solutions to Problem 1 found by NSGA-II. Inset figures show cumulative distribution of $|E_i|$ as i increases.

Fig. 7: Solutions to Problem 2 ($\eta = 10\%$) found by NSGA-II. Top-leftmost plot shows all obtained solutions. Each subsequent plot (follow arrows) is a magnification of a region of the previous plot.

Fig. 8: Solutions to Problem 3 found by NSGA-II. Read x-axis label for a plot from the text-box along the same column and y-axis label from the text-box along the same row.Trade-off characteristics are visible across different pairs of the objective functions.

Fig. 9: Problem 3 solutions for $\%SL < 0.2$ (left) and $\%GL < 0.3$ (right). Encircled solutions can be of interest to a user.

Fig. 10: Problem 4 solutions for $\eta = 0.1$ and $\%loss < 0.3$. Left plot shows solutions when k–anonymity is more desired and right plot shows solutions when ℓ–diversity is more desired.

Fig. 11: Problem 4 solutions for $\eta = 0.1$ and $k \ll \ell$. Left plot shows solutions when maximum $loss$ allowed is 20%. Right plot shows solutions when maximum $loss$ is increased to 21%. The quality of solutions as well as the number of choices available increase considerably for a slight increase in the $loss$ limit.

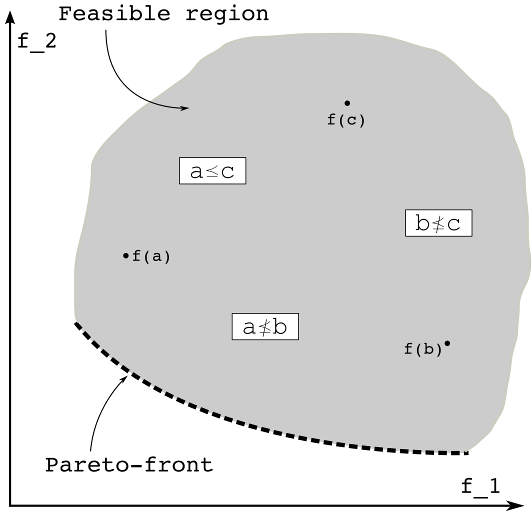

Fig. 1: Pareto-front for a hypothetical two-objective problem.

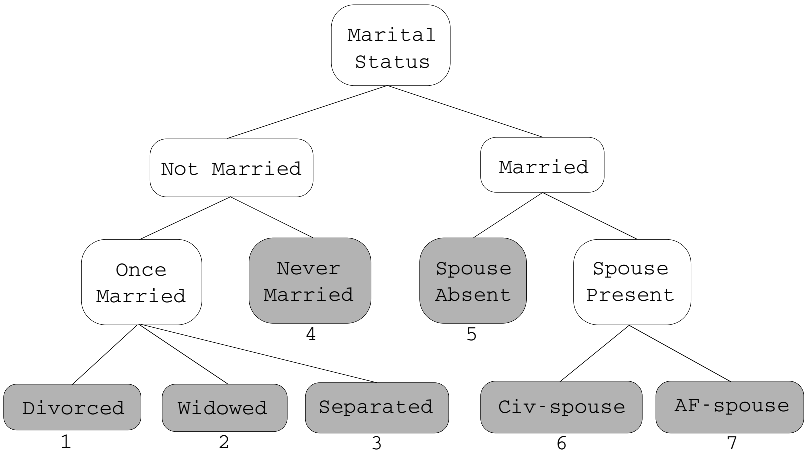

Fig. 2: Hierarchy tree for the *marital status* attribute. Numbering on the leaf nodes indicate their ordering in $\Sigma_{marital\ status}$.

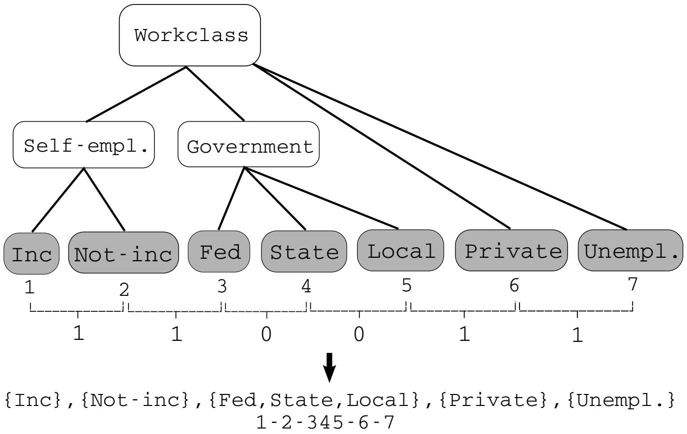

Fig. 3: Example generalization encoding for the *workclass* constrained attribute.

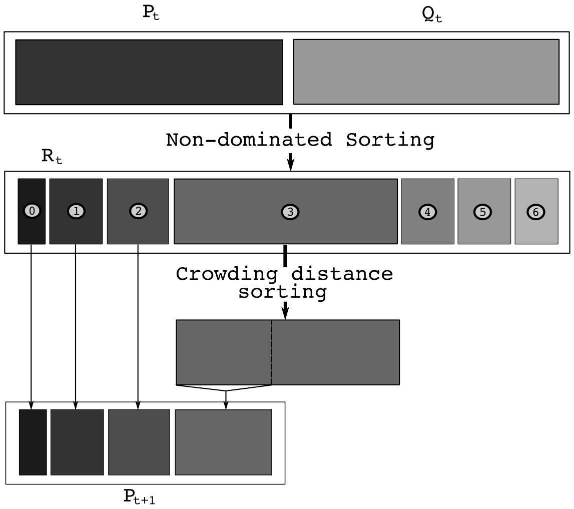

Fig. 4: One generation of NSGA-II.

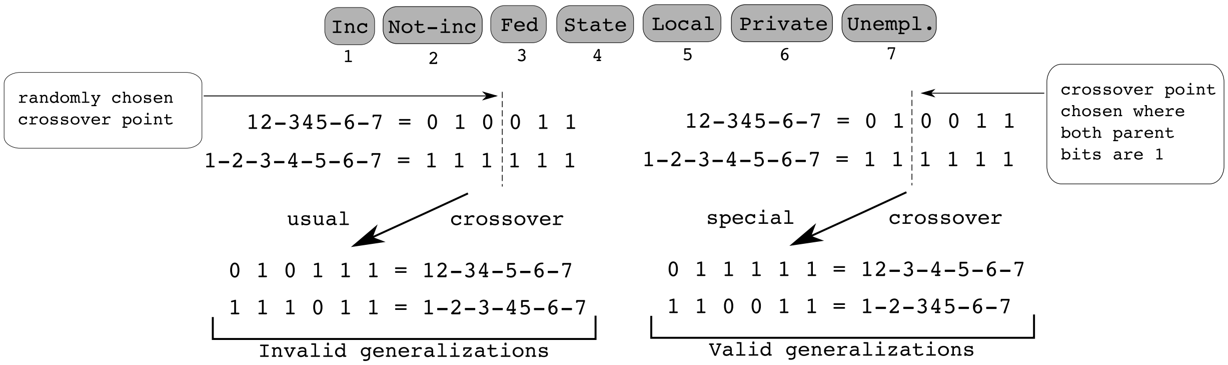

Fig. 5: Usual single point crossover (left) and special crossover for constrained attributes (right).

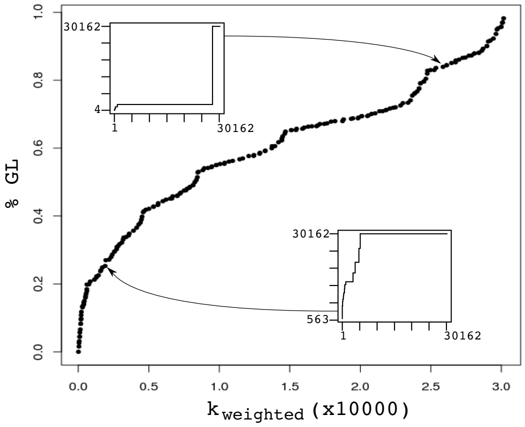

Fig. 6: Solutions to Problem 1 found by NSGA-II. Inset figures show cumulative distribution of $|E_i|$ as i increases.

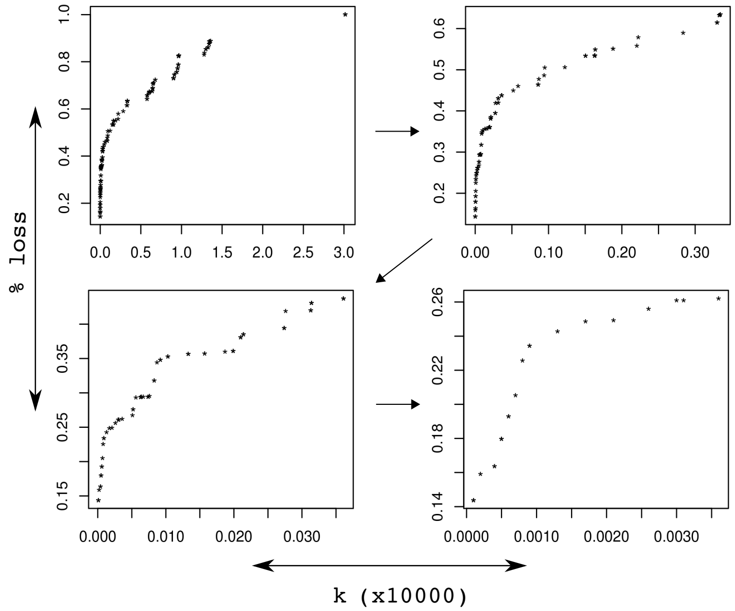

Fig. 7: Solutions to Problem 2 (η = 10%) found by NSGA-II. Top-leftmost plot shows all obtained solutions. Each subsequent plot (follow arrows) is a magnification of a region of the previous plot.

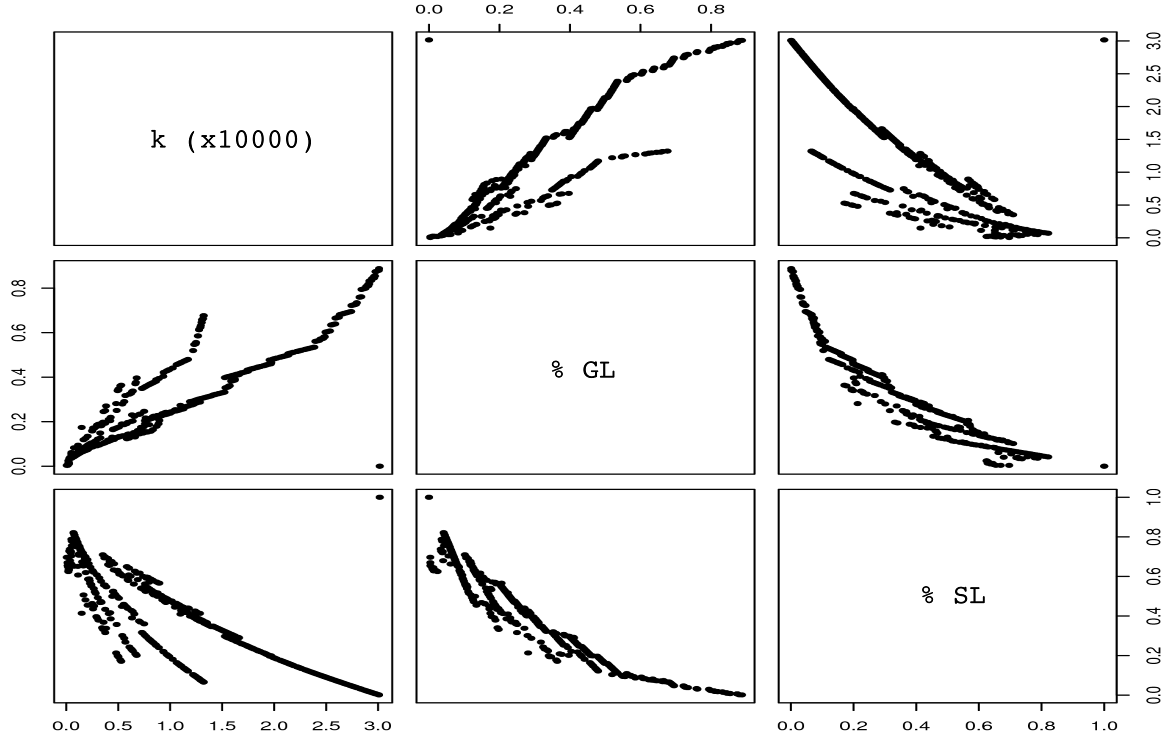

Fig. 8: Solutions to Problem 3 found by NSGA-II. Read x-axis label for a plot from the text-box along the same column and y-axis label from the text-box along the same row.Trade-off characteristics are visible across different pairs of the objective functions.

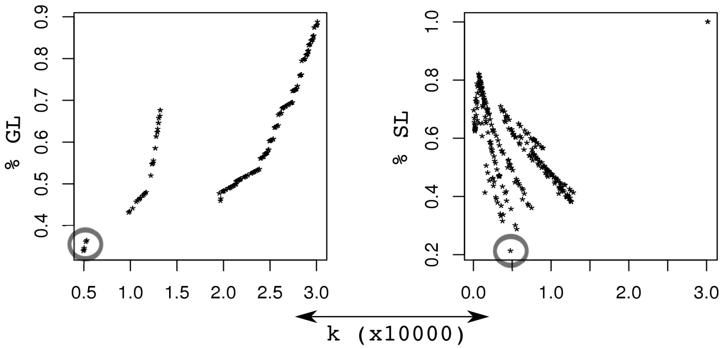

Fig. 9: Problem 3 solutions for $\%SL < 0.2$ (left) and $\%GL < 0.3$ (right). Encircled solutions can be of interest to a user.

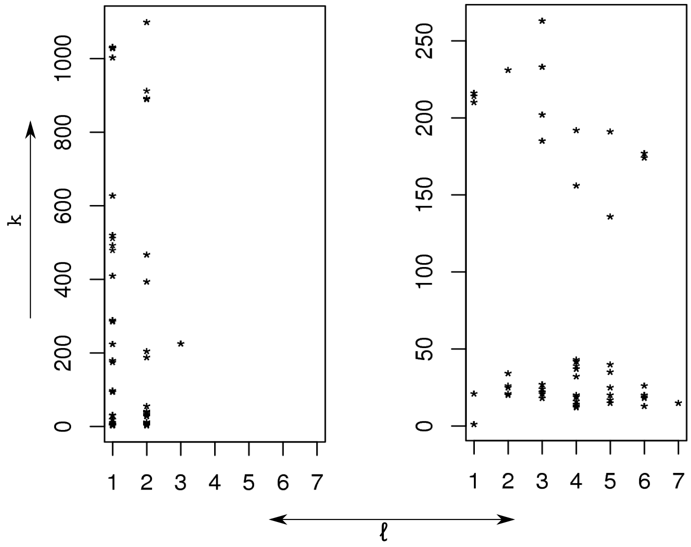

Fig. 10: Problem 4 solutions for $\eta = 0.1$ and $\%loss < 0.3$. Left plot shows solutions when k–anonymity is more desired and right plot shows solutions when ℓ–diversity is more desired.

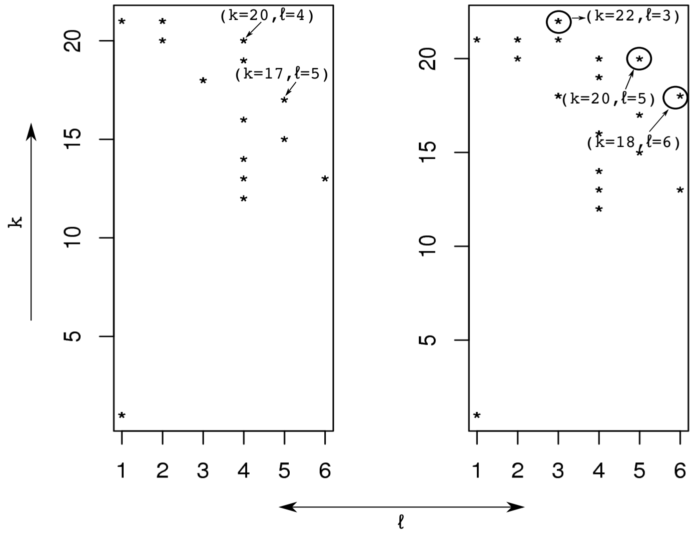

Fig. 11: Problem 4 solutions for $\eta = 0.1$ and $k \ll \ell$. Left plot shows solutions when maximum *loss* allowed is 20%. Right plot shows solutions when maximum *loss* is increased to 21%. The quality of solutions as well as the number of choices available increase considerably for a slight increase in the *loss* limit.