On the Comparison of Microdata Disclosure Control Algorithms

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

ABSTRACT

Privacy models such as k-anonymity and ℓ-diversity typically offer an aggregate or scalar notion of the privacy property that holds collectively on the entire anonymized data set. However, they fail to give an accurate measure of privacy with respect to the individual tuples. For example, two anonymizations achieving the same value of k in the k-anonymity model will be considered equally good with respect to privacy protection. However, it is quite possible that for one of the anonymizations a majority of the individual tuples have lesser probabilities of privacy breaches than their counterparts in the other anonymization. We therefore reject the notion that all anonymizations satisfying a particular privacy property, such as k-anonymity, are equally good. The scalar or aggregate value used in privacy models is often biased towards a fraction of the data set, resulting in higher privacy for some individuals and minimalistic for others. Consequently, to better compare anonymization algorithms, there is a need to formalize and measure this bias. Towards this end, we advocate the use of vector-based methods for representing privacy and other measurable properties of an anonymization. We represent the measure of a given property for an anonymized data set using a *property vector*. Anonymizations are then compared using *quality index functions* that quantify the effectiveness of the property vectors. A formal analysis with respect to their scope and limitations is provided. Finally, we present preference based techniques when comparisons are to be made across multiple properties induced by anonymizations.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*security, integrity, and protection*

General Terms

Performance, Theory

EDBT 2009, March 24–26, 2009, Saint Petersburg, Russia.

Keywords

Data privacy, Anonymization bias, Performance evaluation

1. INTRODUCTION

Microdata disclosure control involves transforming the actual data into a form unrecognizable in terms of the exact values by using *generalization* and *suppression* techniques [16]. Generalization of data is performed by grouping together data attribute values into a more general one. An example of this is replacing a specific age by an age range. Data suppression on the other hand removes entire tuples making them no longer existent in the data set.

Table 1: Hypothetical microdata.

		Zip Code	Age	Marital Status
$\mathcal{T}_1$:	1	13053	28	CF-Spouse
	2	13268	41	Separated
	3	13268	39	Never Married
	4	13053	26	CF-Spouse
	5	13253	50	Divorced
	6	13253	55	Spouse Absent
	7	13250	49	Divorced
	8	13052	31	Spouse Present
	9	13269	42	Separated
	10	13250	47	Separated

An unavoidable consequence of performing such anonymization is a loss in the quality of the data set. Researchers have therefore looked at different methods to obtain an optimal anonymization that results in a minimal loss of information [1, 2, 3, 11, 14, 18, 20, 22]. A typical disclosure control algorithm searches over the space of anonymizations satisfying a particular privacy model, seeking the one with highest utility. An implicit assumption in such optimization attempts is that all anonymizations satisfying a particular privacy property fare equally well in preserving the privacy of individuals. For example, in the k-anonymity model where the measure of privacy is given by the size of the minimum equivalence class in the anonymized data set, two anonymizations of the same data set achieving the same value of k will be considered equally good with respect to privacy protection. Comparative studies based on such an assumption ignores the fact that an anonymization can introduce unwanted bias towards a certain fraction of the individuals represented in the data set. This bias, which we term the

Table 2: Two 3-anonymous generalizations of $\mathcal{T}_1$. Real values of marital status are shown in italics. Left table is denoted as $\mathcal{T}_{3a}$ and right table as $\mathcal{T}_{3b}$.

	Zip Code	Age	Marital Status
1	1305*	(25,35]	Married (*CF-Spouse*)
4	1305*	(25,35]	Married (*CF-Spouse*)
8	1305*	(25,35]	Married (*Spouse Present*)
2	1326*	(35,45]	Not Married (*Separated*)
3	1326*	(35,45]	Not Married (*Never Married*)
9	1326*	(35,45]	Not Married (*Separated*)
5	1325*	(45,55]	Not Married (*Divorced*)
6	1325*	(45,55]	Not Married (*Spouse Absent*)
7	1325*	(45,55]	Not Married (*Divorced*)
10	1325*	(45,55]	Not Married (*Separated*)

	Zip Code	Age	Marital Status
1	130**	(15,35]	Married (*CF-Spouse*)
4	130**	(15,35]	Married (*CF-Spouse*)
8	130**	(15,35]	Married (*Spouse Present*)
2	132**	(35,55]	Not Married (*Separated*)
3	132**	(35,55]	Not Married (*Never Married*)
5	132**	(35,55]	Not Married (*Divorced*)
6	132**	(35,55]	Not Married (*Spouse Absent*)
7	132**	(35,55]	Not Married (*Divorced*)
9	132**	(35,55]	Not Married (*Separated*)
10	132**	(35,55]	Not Married (*Separated*)

anonymization bias, stems from the fact that current privacy models offer only a collective representation of the level of privacy, resulting in higher privacy for some individuals and minimalistic for others. Under such a scenario, the results of comparing the effectiveness of various anonymizations can be misleading.

Let us consider the data set $\mathcal{T}_1$ shown in Table 1. The data set contains 10 tuples with 3 attributes in each – *Zip Code, Age* and *Marital Status*. Table 2 shows two possible 3-anonymous ($\mathcal{T}_{3a}$ and $\mathcal{T}_{3b}$) generalizations of $\mathcal{T}_1$. The level of privacy inferred from $\mathcal{T}_{3a}$ and $\mathcal{T}_{3b}$ is based on 3-anonymity which is essentially the minimum size of an equivalence class. The notion of privacy assumed here is in a minimalistic sense, meaning that every tuple has at most a $\frac{1}{3}$ probability of privacy breach. Thus, both $\mathcal{T}_{3a}$ and $\mathcal{T}_{3b}$ are considered to have the same level of privacy. However, we argue that $\mathcal{T}_{3b}$ should rightfully be evaluated as providing better privacy. This is because tuples $\{2, 3, 5, 6, 7, 9, 10\}$ in $\mathcal{T}_{3b}$ has $\frac{1}{7}$ probability of breach, lower than their counterparts in $\mathcal{T}_{3a}$. In general, with models like k-anonymity and others based on equivalence class sizes, such subtle information is likely to be lost. This is because these privacy measurements are based on certain aggregate property of the anonymization, namely the minimum equivalence class size in this case. What differentiates $\mathcal{T}_{3a}$ and $\mathcal{T}_{3b}$ is the data utility factor which in some sense is orthogonal to the privacy requirement. From a data utility perspective $\mathcal{T}_{3a}$ is perhaps better since the attributes Zip Code and Age are less generalized in $\mathcal{T}_{3a}$ than in $\mathcal{T}_{3b}$. Thus, either of $\mathcal{T}_{3a}$ or $\mathcal{T}_{3b}$ can be preferable depending on a higher utility or a better privacy requirement.

In this work, we focus on identifying ways of comparing anonymizations when such bias is known to exist. We reject a categorical statement such as "4-anonymity is better than 3-anonymity," but seek alternative ways of comparing anonymizations. Towards this end, we introduce a vector-based representation of privacy to address the problem of bias that is induced by the scalar representation. Each property, such as privacy or utility, is associated with a *property vector*, where each element gives a measure of the property for an individual anonymized tuple. Such a representation would signify, for example, the privacy level present for every individual in the data set under a particular privacy model. This will not only allow one to capture the anonymization bias introduced by existing privacy models, but also enable one to perform comparisons between various anonymizations based on the difference in distribution of the privacy levels.

We propose the notion of *quality index functions* that can be used to evaluate the effectiveness of an anonymization and then formally analyze the characteristics of such functions. An m-ary quality index function assigns a real number to a combination of m property vectors. This quantitative estimate is useful in measuring the quality of the property vector. If the quality index value for one instance of a property vector produced by an anonymization is better than another instance produced by a different anonymization, we will say that the first anonymization is preferred over the second. Unary quality index functions are limited in their ability in performing comparisons and can measure only the aggregate properties of the anonymizations. Towards this end, we explore other methods of comparison that allows one to quantify the differences in the values of the property measured across the tuples instead of a minimalistic estimate. We also present various preference based techniques when comparisons are to be made across multiple properties.

The remainder of the paper is organized as follows. The idea of anonymization bias is elaborated upon in Section 2. Section 3 defines the concepts pertinent to the remaining discussion. Section 4 explores the limitations of performing a comparative study using strict comparisons. The requirement for other methods of comparison follows from this in Section 5. A number of comparators are suggested in this section. Section 6 reviews some of the existing works in disclosure control. Finally, Section 7 concludes the paper with a discussion on future extensions.

2. ANONYMIZATION BIAS

Let us revisit the data in Table 1. Table 3 shows a 4-anonymous generalization of the table. We note that further discrepancy beyond those discussed in Section 1 is evident when we start looking at the anonymizations from an user's perspective. Typically, a 4-anonymous generalization is considered to provide higher privacy than a 3-anonymous one. Again, this idea is based on a minimalistic notion of privacy, keeping in mind the entire data set and a certain property satisfied by the tuples in it. However, it is worth noting that attacks on the anonymized data sets could be targeted towards a particular subset of the individuals represented in the data set. In such a situation, a user needs to be concerned about her own level of privacy, rather than that maintained collectively. For example, if user 8 is to choose between the anonymizations $\mathcal{T}_{3b}$ and $\mathcal{T}_4$, the choice would

Table 3: A 4-anonymous generalization of $\mathcal{T}_1$.

		Zip Code	Age	Marital Status
$\mathcal{T}_4$:	1	13***	(20,40]	*
	3	13***	(20,40]	*
	4	13***	(20,40]	*
	8	13***	(20,40]	*
	2	13***	(40,60]	*
	5	13***	(40,60]	*
	6	13***	(40,60]	*
	7	13***	(40,60]	*
	9	13***	(40,60]	*
	10	13***	(40,60]	*

be the latter which conforms to our understanding that 4-anonymity is better. However, if user 3 is in question then the 3-anonymous generalization $\mathcal{T}_{3b}$ is in fact better than $\mathcal{T}_4$. Fig. 1 plots the size of the equivalence class for each tuple in the three different generalizations. The plot tells us that different anonymizations can in fact be better for different individuals. This in some way disrupts our understanding of "better" and "poor" privacy.

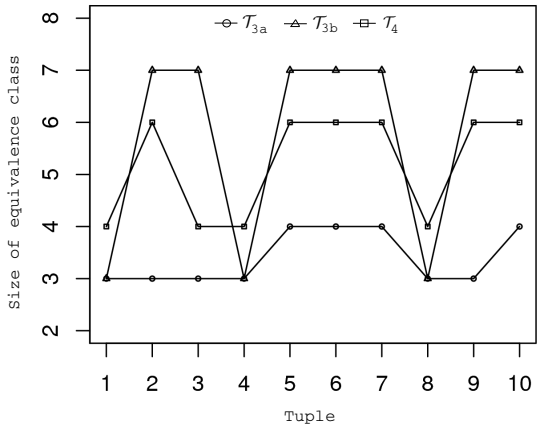

Figure 1: The size of equivalence class to which a tuple in $\mathcal{T}_1$ belongs to for different anonymizations. Two different anonymizations with the same collective privacy level can have different privacy levels for individual tuples.

These fundamental problems are often ignored while performing comparative studies. The notion of privacy assumed in most studies is limited to some overall measure, which can result in anonymizations being *biased* towards some fraction of the data set. Although removing this bias can be a difficult task, no attempt is known to have been made to measure it, less provide privacy measures keeping the bias in consideration. Comparative studies typically assume that if parameters are set similarly in a privacy model, for example k in k-anonymity, then the resulting level of privacy would also be similar. Thereafter, most of the focus during optimization is directed towards obtaining higher utility. With the anonymization bias in picture, the very assumption in the first step of an optimization procedure does not hold any longer. Our work in this paper is not targeted towards defining a new privacy model that overcomes this bias, but to find ways of comparing anonymizations when the bias is known to exist.

Note that the appearance of bias is not limited to k-anonymity alone. When individual measures of privacy are not considered, such bias can appear in any privacy model. The bias can be present even in a personalized privacy setting, such as in the model presented by Xiao and Tao [21]. Personalized privacy in such a model is achieved by constraining the probability of privacy breach for an individual, depending on personal preferences of a breach, to an upper bound. Nonetheless, the individual probabilities need not be same for all tuples, thereby biasing a generalization scheme in more favor towards some tuples than others.

It is imperative to ask if an anonymization can be strictly better than another. It is known that privacy and utility are two conflicting facets of an anonymization, indicating that an anonymization better in one aspect (privacy or utility) is likely to suffer on the other. However, even if utility is not considered as a criteria for obtaining a better generalization, how easy is it to establish that one anonymization is better than another? The answer could be subjective depending on how one defines a "better" anonymization. We shall explore the limitations and alternatives to establish the superiority of an anonymization for various definitions in this context.

3. PRELIMINARIES

Let $\Phi_1, \Phi_2, \ldots, \Phi_a$ represent the domains of a attributes. A data set of size $\mathcal{N}$ defined over these attributes is a collection of $\mathcal{N}$ tuples of the form $(\phi_1, \phi_2, \ldots, \phi_a) \in \Phi_1 \times \Phi_2 \times \ldots \times \Phi_a$. An *anonymization* of a data set is achieved by performing generalization/suppression of the tuples, resulting in an *anonymized data set* unidentifiable from the original one. Since suppression of tuples can be represented as a special case of generalization, we adhere to the term "generalization" to mean both. Further, although tuples suppressed during an anonymization are usually eliminated, we assume that they still exist in the anonymized data set in an overly generalized form. This enables us to say that both the original data set and the anonymized one are of the same size. An anonymized data set is then subjected to a variety of property measurements. A *property* here refers to a scalar quantity signifying a privacy, utility or any other measurable feature of a tuple in the anonymized data set. This gives us a vector of values representing the property value for every tuple of the anonymized data set.

Definition 1. (**Property Vector**) A property vector $\mathcal{D}$ for a data set of size $\mathcal{N}$ is an $\mathcal{N}$-dimensional vector $(d_1, d_2, \ldots, d_{\mathcal{N}})$ with $d_i \in \mathbb{R}; 1 \leq i \leq \mathcal{N}$ specifying a measure of a property for the i^{th} tuple of the data set.

A property signifies the grounds under which a comparison is made between two anonymizations. Consider that we are performing k-anonymization on a data set. A generalization scheme to do so results in multiple equivalence classes, the desired property being that all such equivalence classes are of size at least k. If we pick our privacy property to be the "*size of the equivalence class to which a tuple belongs*", then each tuple will have an associated integer. This results in a property vector $s = (s_1, s_2, \ldots, s_{\mathcal{N}})$ for a data set of size $\mathcal{N}$. For example, the equivalence class property vector induced in $\mathcal{T}_{3a}$ is $(3, 3, 3, 3, 4, 4, 4, 3, 3, 4)$. Another example of a property could be the contribution made by a tuple to

the total information loss. If measurements on the diversity of sensitive attributes in an equivalence class is desired, then the property can be the number of times the sensitive attribute value of a tuple appears in its equivalence class. Considering "Marital Status" as a sensitive attribute, such a property vector for $\mathcal{T}_{3a}$ will be $(2, 2, 1, 2, 2, 1, 2, 1, 2, 1)$.

Ideally, any number of properties can be studied on a given anonymization. An anonymization is only a representation of the generalization scheme, inducing different property vectors for different properties. For example, one may be interested in analyzing an anonymization w.r.t. both k-anonymity and ℓ-diversity. In such a case, the property vectors to consider are the ones generated by the properties - *size of equivalence class of a tuple* and *count of the sensitive attribute value of a tuple in its equivalence class*. For an anonymization, a property vector due to the first property relates to k-anonymity, while that from the latter one relates to ℓ-diversity. We thus use the notion of *r-property anonymization* to indicate that the set of properties decided upon for a comparison process is restricted to a pre-specified set of r properties. The objective of a comparison is to decide if one anonymization is better than another w.r.t. the specified properties.

Definition 2. (r–**Property Anonymization**) Let Δ be the set of all data sets of size $\mathcal{N}$. Let Υ be the set of all elements $\upsilon \in 2^{\mathbb{R}^{\mathcal{N}}}$ such that $|\upsilon| = r$. A r–property anonymization $\mathcal{G}$ is a function $\mathcal{G} : \Delta \rightarrow \Upsilon$ which induces a set of r $\mathcal{N}$-dimensional property vectors $\mathcal{G}(\delta)$ on a data set $\delta \in \Delta$.

Note that an r-property anonymization does not mean that there are restrictions on how an anonymization is done. It only indicates that, for a given anonymization, r different property vectors are chosen to proceed with the comparison. The idea is to project an anonymized data set into a set of $\mathcal{N}$-dimensional vectors with regard to r properties, and then compare the resulting vectors for different anonymization schemes. For example, the aforementioned example of analyzing the size of equivalence class and the number of sensitive attribute values of a tuple in its equivalence class will be referred to as 2–property anonymization.

Comparisons between anonymizations is done by defining a *comparator*, denoted by $\rhd$. A comparator is an ordering operation defined on sets of property vectors. An example is the dominance-based comparator $\succeq$ – under a 1-property anonymization, say $\Upsilon_1 = \{(d^1_1, \ldots, d^1_{\mathcal{N}})\}$ and $\Upsilon_2 = \{(d^2_1, \ldots, d^2_{\mathcal{N}})\}$, then $\Upsilon_1 \succeq \Upsilon_2$ iff $\forall 1 \leq i \leq \mathcal{N}, d^1_i \geq d^2_i$. In other words, comparators are user-defined ways of evaluating the superiority of a property vector. An anonymization is better than another only w.r.t. the comparator used in comparing the induced set of property vectors. Therefore, given a comparator $\rhd$, the relation $\mathcal{G}_1 \rhd \mathcal{G}_2 \iff \Upsilon_1 \rhd \Upsilon_2$ is implicitly assumed to be true. Note that the definition of a comparator need not always be explicitly made in terms of the values in a property vector. For example, a comparator may be defined just to check if more values in one property vector is higher than the corresponding values in another vector. Hence, we use *quality index functions* on property vectors to quantitatively measure the competence of a set of property vectors.

Definition 3. (m**-ary Quality Index**) Let Π be the set of all property vectors w.r.t a particular property. An m-ary quality index P is a function $P : \Pi^m \rightarrow \mathbb{R}$ which assigns a combination of m property vectors $\mathcal{D}_1, \ldots, \mathcal{D}_m$ a real value $P(\mathcal{D}_1, \ldots, \mathcal{D}_m)$.

Since comparisons are usually performed by applying an anonymization on the same data set, we shall restrict Π to be the set of all property vectors of the same size, i.e. the size $\mathcal{N}$ of the data set. Based on the same reasoning, a quality index may also use the original data set while mapping property vectors to real numbers.

A commonly used method of performing comparisons is through *unary* quality index functions (1-ary). Unary quality indices are functions applied independently on anonymizations. They measure one or more feature (privacy/utility) of an anonymization, and the quantitative value is considered representative of the measured feature of the anonymization. For example, k-anonymity is an unary quality index on the equivalence class size property vector, given as $P_{k-anon}(s) = \min_{s_i}(s)$ ($= 3$ for $\mathcal{T}_{3a}$). Another possible quality index could be the average size of the equivalence class maintained in the anonymized data set, i.e. $P_{s-avg}(s) = \sum s_i/\mathcal{N}$ ($= 3\times3+3\times3+4\times4/10$=3.4 for $\mathcal{T}_{3a}$). Other models like ℓ-diversity and t-closeness results in other property vectors depending on the property being measured, for example ℓ-diversity uses a count of the number of times the sensitive attribute value of a tuple is represented in an equivalence class. With this property, we shall have a unary quality index $\ell = P_{\ell-div}((2, 2, 1, 2, 2, 1, 2, 1, 2, 1))$ value of 1 for $\mathcal{T}_{3a}$ (considering "Marital Status" as the sensitive attribute). Once again, the ℓ or t values for an anonymization is a quality measurement on the property vectors, the minimum values in these models. Certain forms of utility measurements can also be captured from property vectors. A loss measurement, such as the *general loss metric* [7], computes a normalized loss quantity for every tuple of the data set. For such metrics, a property vector specifies the loss resulting from each tuple in the data set. Thereafter, the quality index is some form of aggregation of the individual losses.

Note that unary indices only allow the measurement of an aggregate property of an anonymization. This limits any kind of comparison against the bias that may be present across anonymizations. More specifically, comparisons are not possible across the property values maintained by a tuple in two different anonymizations. This problem is eliminated by *binary* indices (2-ary) since both anonymizations are now available to allow comparison of individual components of the induced property vectors. A binary quality index has two anonymizations as input and the real-valued output signifies a relative measure of one anonymization's effectiveness over another. For example, a binary quality index such as $P_{binary}(s, t) = |\{s_i | s_i > t_i\}|$ counts the number of entries in the property vector s that has higher property values than the corresponding entries in t. Note that s and t in this case are property vectors measuring the same property in two different anonymizations. For the size of equivalence class property in $\mathcal{T}_{3a}$, with property vector $s = (3, 3, 3, 3, 4, 4, 4, 3, 3, 4)$, and $\mathcal{T}_{3b}$, with property vector $t = (3, 7, 7, 3, 7, 7, 7, 3, 7, 7)$, we have $P_{binary}(s, t) = 0$ and $P_{binary}(t, s) = 7$. These index values indicate that if an 1-property anonymization is analyzed w.r.t. the size of equivalence class property, then anonymization $\mathcal{T}_{3b}$ inducing the property vector t is preferable over $\mathcal{T}_{3a}$.

Quality estimation based on index functions stresses on

Table 4: Strict comparators based on dominance relationships.

Comparator ($\rhd$)	$\mathcal{D}_1 \rhd \mathcal{D}_2$	$\Upsilon_1 \rhd \Upsilon_2$	$\mathcal{G}_1 \rhd \mathcal{G}_2$
Weak dominance ($\succeq$)	$\forall i; d_i^1 \geq d_i^2$	$\forall \mathcal{D}_i \in \Upsilon_1, \mathcal{D}'_i \in \Upsilon_2; \mathcal{D}_i \succeq \mathcal{D}'_i$	$\mathcal{G}_1$ is not worse than $\mathcal{G}_2$
Strong dominance ($\succ$)	$\forall i; d_i^1 \geq d_i^2$ $\wedge \exists j; d_j^1 > d_j^2$	$\forall \mathcal{D}_i \in \Upsilon_1, \mathcal{D}'_i \in \Upsilon_2; \mathcal{D}_i \succeq \mathcal{D}'_i$ $\wedge \exists \mathcal{D}_j \in \Upsilon_1, \mathcal{D}'_j \in \Upsilon_2; \mathcal{D}_j \succ \mathcal{D}'_j$	$\mathcal{G}_1$ is better than $\mathcal{G}_2$
Non-dominance ($\|$)	$\exists i, j; d_i^1 < d_i^2 \wedge d_j^1 > d_j^2$	$\exists \mathcal{D}_i \in \Upsilon_1, \mathcal{D}'_i \in \Upsilon_2; \mathcal{D}_i \succ \mathcal{D}'_i$ $\wedge \exists \mathcal{D}_j \in \Upsilon_1, \mathcal{D}'_j \in \Upsilon_2; \mathcal{D}'_j \succ \mathcal{D}_j$	$\mathcal{G}_1$ and $\mathcal{G}_2$ are incomparable
User defined ($\blacktriangleright$-better)	$\mathcal{D}_1$ is $\blacktriangleright$-better than $\mathcal{D}_2$	Υ_1 is $\blacktriangleright$-better than Υ_2	$\mathcal{G}_1$ is $\blacktriangleright$-better than $\mathcal{G}_2$

the fact that a particular anonymization can be interpreted with respect to many different privacy properties and utility measurements. The superiority of one anonymization to another is thus dependent on what privacy properties are taken into consideration while performing the comparison. If quality indices are used to establish this superiority, then our concern is how many of them are needed to do so.

4. STRICT COMPARISONS

Most algorithms in disclosure control are designed to obtain anonymizations that can maximize the utility of the anonymized data while satisfying a pre-specified privacy property. An anonymization is considered to be better than another if it provides higher utility within the constraints of the privacy requirement. Privacy, as given by a model, and utility are two properties induced by an anonymization. A disclosure control algorithm scans through the space of anonymizations satisfying a privacy property to find the one with maximum utility. Performance of one algorithm is said to be better than another if it is able to find an anonymization with higher utility.

This form of comparison suffers from the fact that the privacy level measured from an anonymization is a scalar quantity. It is known that maximum privacy and maximum utility are orthogonal objectives that cannot be achieved at the same time. Hence, it is imperative that when an anonymization with a better utility is found by an algorithm, the privacy factor must suffer. However, this facet may not be exposed if scalar measures are used to represent privacy. The anonymization bias plays an important role here in explaining a degraded performance in privacy from high utility anonymizations. Further, optimization attempts are also rare where emphasis is laid on obtaining anonymizations that satisfy more than one privacy property.

Discrepancies of the above nature prompts us to consider vector based measurements of the properties induced by an anonymization. Our perspective of an anonymization is that of a source that induces various properties, both in terms of privacy and utility, and more importantly, the properties can be measured on each tuple in the data set. Thereafter, methods to compare these property vectors (one or more) are devised to evaluate the competence of an anonymization.

The first question we ask is the feasibility of performing *strict comparisons*. A strict comparison between property vectors follows from the concept of dominance, widely used in the multi-objective optimization community. The notions of *weak* and *strong* dominance enables us to make strong statements about the superiority of an anonymization. Weak dominance says that every measured value of a property on every tuple of the data set after an anonymization must be at least as good as the value measured from the corresponding tuples with another anonymization. This establishes a "*not worse than*" relationship between vectors. Strong dominance offers a stricter notion and establishes the "*better than*" relationship (Table 4). The *non-dominance* relationship signifies incomparable vectors. We are interested in strict comparisons based on dominance because they provide a framework to undoubtedly justify why an anonymization is better than another. It can potentially eliminate the effects of anonymization bias during a comparative study. However, the following discussion shows that adopting a dominance based comparative method could be rather impractical.

THEOREM 1. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two property vectors measuring the same property and induced by the 1–property anonymizations $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively on a data set of size $\mathcal{N}$. If $\mathcal{P} = (P_1, P_2, \ldots, P_n)$ be a vector of n unique unary quality indices such that*

$$\forall 1 \leq i \leq n : P_i(\mathcal{D}_1) \geq P_i(\mathcal{D}_2) \iff \mathcal{D}_1 \succeq \mathcal{D}_2,$$

then the number of indices is at least equal to the size of the data set, i.e. $n \geq \mathcal{N}$.

PROOF. The proof follows from the fact that the number of open hypercubes required to cover $\mathbb{R}^{\mathcal{N}}$ is finite. We shall show that if $n < \mathcal{N}$, then infinite such open hypercubes can be defined. The proof is by induction. Consider the two non-comparable property vectors $\mathcal{D}_1 = (a, b)$ and $\mathcal{D}_2 = (b, a)$ with $a, b \in \mathbb{R}$ and $n = 1$. Then either $P_1(\mathcal{D}_1) \geq P_1(\mathcal{D}_2)$ or $P_1(\mathcal{D}_1) < P_1(\mathcal{D}_2)$. This implies that either $\mathcal{D}_1 \succeq \mathcal{D}_2$ or $\mathcal{D}_2 \succeq \mathcal{D}_1$, which leads to a contradiction since $\mathcal{D}_1 \parallel \mathcal{D}_2$. Hence the theorem holds for $\mathcal{N} = 2$.

Let us suppose that the theorem holds for data sets of size $\mathcal{N}-1$. Consider the two property vectors $\mathcal{D}_1 = (a, a, \ldots, a, c)$ and $\mathcal{D}_2 = (b, b, \ldots, b, c)$ with $a, b \in \mathbb{R}$ and $a < b$ on a data set of size $\mathcal{N}$. Further, let us assume that there exists a combination of $n < \mathcal{N}$ unary quality indices satisfying the relation in the theorem.

We first show that $\forall 1 \leq i \leq n : P_i(\mathcal{D}_1) < P_i(\mathcal{D}_2)$.

If $\exists j$ such that $P_j(\mathcal{D}_1) > P_j(\mathcal{D}_2)$ then $\mathcal{D}_2 \not\succeq \mathcal{D}_1$ which results in a contradiction.

Next, consider a property vector $\mathcal{D} \in \mathcal{D}_c = \{(d_1, d_2, \ldots, d_{\mathcal{N}-1}, c) | \forall 1 \leq i \leq \mathcal{N}-1 : a < d_i < b\}; c \in \mathbb{R}$. We then have $\mathcal{D}_2 \succeq \mathcal{D} \succeq \mathcal{D}_1$ leading to $\forall 1 \leq i \leq n : P_i(\mathcal{D}_2) \geq P_i(\mathcal{D}) \geq P_i(\mathcal{D}_1)$. If $\exists j$ such that $P_j(\mathcal{D}_2) = P_j(\mathcal{D}_1)$, then $P_j(\mathcal{D}_2) = P_j(\mathcal{D}) = P_j(\mathcal{D}_1)$. With this we can now prove that the number of indices required for data sets of size $\mathcal{N}-1$ can be less than $\mathcal{N}-1$, contrary to the hypothesis assumed. To do so, we consider two property vectors $\mathcal{D}_p = (p_1, p_2, \ldots, p_{\mathcal{N}-1})$ and $\mathcal{D}_q = (q_1, q_2, \ldots, q_{\mathcal{N}-1})$ on a data set of size $\mathcal{N}-1$, such that $\forall 1 \leq i \leq \mathcal{N}-1 : a < p_i \leq q_i < b$. Hence, $\mathcal{D}_q \succeq \mathcal{D}_p$. Next, we expand the two vectors by concatenating the same arbitrary element $c \in \mathbb{R}$, the concatenated vectors being denoted by $\mathcal{D}_{p|c}$ and $\mathcal{D}_{q|c}$ respectively.

Since $\forall 1 \leq i \leq n : P_i(\mathcal{D}_{q|c}) \geq P_i(\mathcal{D}_{p|c}) \iff \mathcal{D}_{q|c} \succeq \mathcal{D}_{p|c}$ and $(\mathcal{D}_q \succeq \mathcal{D}_p) \iff (\mathcal{D}_{q|c} \succeq \mathcal{D}_{p|c})$, we have $\forall 1 \leq i \leq n : P_i(\mathcal{D}_{q|c}) \geq P_i(\mathcal{D}_{p|c}) \iff \mathcal{D}_q \succeq \mathcal{D}_p$. Also, using the result that $P_j(\mathcal{D}_2) = P_j(\mathcal{D}) = P_j(\mathcal{D}_1)$ for any $\mathcal{D} \in \mathcal{D}_c$, we have $P_j(\mathcal{D}_{q|c}) = P_j(\mathcal{D}_{p|c})$. Hence the result of P_j can be ignored and one can write $\forall 1 \leq i \neq j \leq n : P_i(\mathcal{D}_{q|c}) \geq P_i(\mathcal{D}_{p|c}) \iff \mathcal{D}_q \succeq \mathcal{D}_p$, which means that less than $\mathcal{N}-1$ (recall $n < \mathcal{N}$) quality indices can be used to compare $\mathcal{D}_p$ and $\mathcal{D}_q$. Since this is contrary to the assumed hypothesis, the existence of such j is not possible.

Thus, $\forall 1 \leq i \leq n : P_i(\mathcal{D}_1) < P_i(\mathcal{D}_2)$.

Let us now consider the open hyperrectangle $\mathcal{I}_c = \{(r_1, r_2, \ldots, r_n) \in \mathbb{R}^n | \forall 1 \leq i \leq n : P_i(\mathcal{D}_1) < r_i < P_i(\mathcal{D}_2)\}$. Also, for $f > c$, $\mathcal{I}_c \cap \mathcal{I}_f = \phi$; if $\exists (r_1, r_2, \ldots, r_n) \in \mathcal{I}_c \cap \mathcal{I}_f$ then $\forall 1 \leq i \leq n : [P_i((a, a, \ldots, a, c)) < r_i < P_i((b, b, \ldots, b, c))] \wedge [P_i((a, a, \ldots, a, f)) < r_i < P_i((b, b, \ldots, b, f))]$, giving $\forall 1 \leq i \leq n : P((a, a, \ldots, a, f)) < P_i((b, b, \ldots, b, c))$, and implying that $(b, b, \ldots, b, c) \succeq (a, a, \ldots, a, f)$ which is absurd. Hence, since $\mathbb{R}$ is uncountable and c is chosen arbitrarily in $\mathbb{R}$, there are uncountably many disjoint open hyperrectangles in the n-dimensional quality index space $\mathbb{R}^n$. This is in contradiction to the fact that $\mathbb{R}^n$ contains countably many disjoint open hyperrectangles. Therefore, $n \not< \mathcal{N}$ which implies $n \geq \mathcal{N}$. $\square$

A similar proof can be given for the case when strong dominance has to be inferred between two property vectors, i.e. for the equivalence relation $[\forall 1 \leq i \leq n : P_i(\mathcal{D}_1) \geq P_i(\mathcal{D}_2) \wedge \exists j \in \{1, \ldots, n\} | P_j(\mathcal{D}_1) > P_j(\mathcal{D}_2)] \iff \mathcal{D}_1 \succ \mathcal{D}_2$ to hold, n must be at least $\mathcal{N}$.

An important question here is whether all possible $\mathcal{N}$-dimensional property vectors are valid given a specific privacy measurement and a specific data set. The answer is a strict no. For example, if we consider the size of the equivalence class as the privacy property, one commonly used in models like k-anonymity and ℓ-diversity, we would find that the measurements are dependent. In other words, if a tuple belongs to an equivalence class of size s, then there exists $s-1$ other tuples that belong to the same equivalence class. Hence the measurement (size of equivalence class) will be the same for all s tuples. This restricts us from attaining all possible property vectors. This then raises the question if Theorem 1 is valid when the set of possible property vectors is actually a subset of the set of all possible property vectors. We show that the theorem in fact holds for such a subset as well.

COROLLARY 1. *Let Π be the set of all property vectors that can be defined for a data set of size $\mathcal{N}$. Let $\mathcal{D} \subseteq \Pi$ be a set of property vectors measuring a particular property such that $\forall \mathcal{D}_1, \mathcal{D}_2 \in \mathcal{D}$,*

$$\forall 1 \leq i \leq n : P_i(\mathcal{D}_1) \geq P_i(\mathcal{D}_2) \iff \mathcal{D}_1 \succeq \mathcal{D}_2$$

Then, $n \geq \mathcal{N}$.

PROOF. Let us assume that $n < \mathcal{N}$. The proof is given by constructing the maximal superset $\mathcal{D}_M$ of $\mathcal{D}$ on which the relationship holds. Let $a = (a_1, a_2, \ldots, a_\mathcal{N}), b = (b_1, b_2, \ldots, b_\mathcal{N}) \in \mathcal{D}$. Hence we can say, $\forall 1 \leq i \leq n : P_i(a) \geq P_i(b) \iff a \succeq b$. Also then $a, b \in D_M$. Now consider the following vectors.

- $x \in \mathcal{X} = \{(a_1 c_1, a_2 c_2, \ldots, a_\mathcal{N} c_\mathcal{N}) | \forall 1 \leq i \leq \mathcal{N}; c_i \geq 1\}$
- $y \in \mathcal{Y} = \{(b_1 + (a_1 - b_1)e_1, \ldots, b_\mathcal{N} + (a_\mathcal{N} - b_\mathcal{N})e_\mathcal{N}) | \forall 1 \leq i \leq \mathcal{N}; 0 \leq e_i \leq 1\}$
- $z \in \mathcal{Z} = \{(b_1/d_1, b_2/d_2, \ldots, b_\mathcal{N}/d_\mathcal{N}) | \forall 1 \leq i \leq \mathcal{N}; d_i \geq 1\}$

We have the following relation on these vectors: $a \succeq b \iff x \succeq a \succeq y \succeq b \succeq z$. Hence, by applying the quality indices P_i's on a and b one can compare two property vectors belonging to two different sets from $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$. Note that one can still not assert that two vectors belonging to the same set ($\mathcal{X}$, $\mathcal{Y}$ or $\mathcal{Z}$) can be compared in the same manner. Thus, we arbitrarily choose three vectors, one each from $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, and include it in D_M, i.e. $D_M = \{x, a, y, b, z\}$. Hence, given two vectors in D_M, we can find at least three other vectors that satisfy the relationship, thereby increasing the cardinality of D_M. Since the choice of the new vectors is arbitrary, we can continue the process as many times as possible, every time increasing the cardinality of D_M. Then, in the limit, D_M will become equal to Π. This contradicts the result from Theorem 1 saying that $n \geq \mathcal{N}$ quality indices are required to compare two property vectors in Π. Therefore, the assumed hypothesis $n < \mathcal{N}$ is incorrect, implying $n \geq \mathcal{N}$. $\square$

The situation becomes worse if comparisons are to be made against multiple properties. On first thought, it is possible to map a set of property vectors to an unique property vector of size $\mathcal{N}$ and Theorem 1 could suggest that the lower bound on n is therefore $\mathcal{N}$. Let $\Upsilon_p = \{\mathcal{D}_{p1}, \mathcal{D}_{p2}, \ldots, \mathcal{D}_{pr}\}$ be a set of r property vectors, where $\mathcal{D}_{pi} = (d^p_{i1}, d^p_{i2}, \ldots, d^p_{i\mathcal{N}})$; $1 \leq i \leq r$. Consider the vector $d_j \in D_j = \{(d^p_{1j}, d^p_{2j}, \ldots, d^p_{rj}) \in \mathbb{R}^r\}$ for some $j \in \{1, ..., \mathcal{N}\}$. Note that $|D_j| = \mathbb{R}^r$ and since the cardinality of $\mathbb{R}^r$ and $\mathbb{R}$ is the same, there exists a bijective function $f_j : D_j \to \mathbb{R}$. Now define the function $\mathcal{F} : \Upsilon \to \mathbb{R}^\mathcal{N}$ which maps each set of r property vectors to a vector of size $\mathcal{N}$ as $\mathcal{F}(\Upsilon_p \in \Upsilon) = (f_1(d_1), f_2(d_2), \ldots, f_\mathcal{N}(d_\mathcal{N}))$. Since every f_j is a bijective function, $\mathcal{F}$ is bijective too. However, a bijective mapping is not sufficient to imply the equivalence relation $\mathcal{F}(\Upsilon_1) \succeq \mathcal{F}(\Upsilon_2) \iff \Upsilon_1 \succeq \Upsilon_2$ and hence the lower bound ($n \geq \mathcal{N}$) given by Theorem 1 would be incorrect. It can be shown that for the relation to hold, n should at least be equal to $r\mathcal{N}$. The following corollary gives this lower bound on the number of quality indices required to compare two sets of property vectors. Note that the corollary uses notions of quality index functions extended to sets of property vectors.

COROLLARY 2. *Let $\Upsilon_1 = \{\mathcal{D}_{11}, \mathcal{D}_{12}, \ldots, \mathcal{D}_{1r}\}$ and $\Upsilon_2 = \{\mathcal{D}_{21}, \mathcal{D}_{22}, \ldots, \mathcal{D}_{2r}\}$ be two sets of property vectors induced by the r–property anonymizations $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively on a data set of size $\mathcal{N}$. If $\mathcal{P} = (P_1, P_2, \ldots, P_n)$ be a vector of n unary quality indices such that*

$$\forall 1 \leq i \leq n : P_i(\Upsilon_1) \geq P_i(\Upsilon_2) \iff \Upsilon_1 \succeq \Upsilon_2$$

then $n \geq r\mathcal{N}$.

PROOF. Let us assume that there exists quality indices $P_i; 1 \leq i < r\mathcal{N}$ such that the equivalence relation holds. Now, $\Upsilon_1 \succeq \Upsilon_2 \iff \forall 1 \leq j \leq r; \mathcal{D}_{1j} \succeq \mathcal{D}_{2j}$. If $\mathcal{D}_{pj} = (d^p_{j1}, d^p_{j2}, \ldots, d^p_{j\mathcal{N}})$ then $\mathcal{D}_{1j} \succeq \mathcal{D}_{2j} \iff \forall 1 \leq z \leq \mathcal{N}; d^1_{jz} \geq d^2_{jz}$. Therefore, $\Upsilon_1 \succeq \Upsilon_2 \iff \forall 1 \leq j \leq r; \forall 1 \leq z \leq \mathcal{N}; d^1_{jz} \geq d^2_{jz}$. By transitivity, we then have $\forall 1 \leq i < r\mathcal{N} : P_i(\Upsilon_1) \geq P_i(\Upsilon_2) \iff \forall 1 \leq j \leq r; \forall 1 \leq z \leq \mathcal{N}; d^1_{jz} \geq d^2_{jz}$.

Let us now consider a data set of size $r\mathcal{N}$. For two property vectors $\mathcal{D}_1$ and $\mathcal{D}_2$ defined on this data set, $\mathcal{D}_1 \succeq \mathcal{D}_2$ if and only if every component of $\mathcal{D}_1$ is greater than or equal to

the corresponding component of $\mathcal{D}_2$. We first divide a property vector on this data set into r equal sections, thereby resulting in r $\mathcal{N}$-dimensional vectors. The quality indices P_i can then be applied to these resulting vectors. The equivalence relation $\forall 1 \leq i < r\mathcal{N} : P_i(\Upsilon_1) \geq P_i(\Upsilon_2) \iff \forall 1 \leq j \leq r; \forall 1 \leq z \leq \mathcal{N}; d^1_{jz} \geq d^2_{jz}$ can then be used to state that less than $r\mathcal{N}$ quality indices can be used to establish a dominance relation between $\mathcal{D}_1$ and $\mathcal{D}_2$. This contradicts the result from Theorem 1 which states that we would require at least $r\mathcal{N}$ quality indices to compare any two property vectors on a data set of size $r\mathcal{N}$. Hence, no such combination of quality indices with $n < r\mathcal{N}$ can exist. □

The results till this point indicate that it is rather impractical to determine the superiority of an anonymization based on notions of weak or strong dominance, and with unary quality indices. This prompts us to consider other methods of representing the quality of anonymizations relative to one another by defining ▶-better (read as *metric better*) comparators. For example, we can define a $\blacktriangleright_{cov}$-better comparator such that, given two equivalence size property vectors $\mathcal{D}_1$ and $\mathcal{D}_2$, $\mathcal{D}_1 \blacktriangleright_{cov} \mathcal{D}_2$ if more tuples in $\mathcal{D}_1$ have a higher equivalence class size than the corresponding number in $\mathcal{D}_2$. Another example is the $\blacktriangleright_{min}$-better comparator typically used in models such as k-anonymity – $\mathcal{D}_1 \blacktriangleright_{min} \mathcal{D}_2$ if $\min_{d^1_i}(\mathcal{D}_1) > \min_{d^2_i}(\mathcal{D}_2)$. In fact, current methods of comparison are all based on some ▶-better comparator.

▶-better comparators may naturally induce the quality indices to be used to infer the relationship, for example P_{k-anon} is the quality index to be used to infer $\blacktriangleright_{min}$-better. However, as mentioned earlier, comparators such as $\blacktriangleright_{min}$ are limited by the fact that they are based on some aggregate property of the vectors, ignoring the anonymization bias altogether. On the other hand, $\blacktriangleright_{cov}$ is a rational candidate since the relationship is based on the output from comparing multiple tuples in the two property vectors. Note that $\blacktriangleright_{cov}$ in fact induces a binary quality index as presented in the next section.

5. ▶-BETTER COMPARATORS

Comparisons with dominance based comparators suffer from the drawback that the number of unary quality indices required is impractically large. This also follows from our intuition that comparison of two $\mathcal{N}$-dimensional vectors cannot be accomplished with less than $\mathcal{N}$ scalar quantities without losing information. Besides this difficulty, dominance based comparison is a rather strict way of evaluating the quality of an anonymization. It is not unlikely that a property vector is not able to dominate another vector because of low property values for a minor fraction of the tuples. This effectuates to saying that the anonymization bias could be present negligibly resulting in a non-dominance relationship.

Although removing the effects of anonymization bias is hard (or perhaps impossible), methods can be devised to keep it in consideration during a comparative study. We therefore seek other comparators, called *▶-better comparators,* that can capture the quality of anonymizations together with the bias they introduce. ▶-better comparators provide a weaker notion of superiority than dominance-based ones, nonetheless their objective is also targeted towards identifying anonymizations with better utilization of the bias. Hence, we emphasize that such comparators pay adequate attention to the property values across all tuples of the anonymized data set.

In the following discussion, we introduce a number of ▶-better comparators and the corresponding unary/binary quality index they induce. All expressions below are in terms of two property vectors $\mathcal{D}_1 = (d^1_1, d^1_2, \ldots, d^1_{\mathcal{N}})$ and $\mathcal{D}_2 = (d^2_1, d^2_2, \ldots, d^2_{\mathcal{N}})$ measuring a given property when induced by two different anonymizations on a data set of size $\mathcal{N}$. Without loss of generality, we assume that a higher value of a property measurement for a tuple is better.

5.1 $\blacktriangleright_{rank}$-better

Consider the $\mathcal{N}$-dimensional space of all property vectors for a given property. Let $\mathcal{D}_{max}$ be a *point of interest* and all points are assigned a *rank* based on their distance from $\mathcal{D}_{max}$. We then say $\mathcal{D}_1 \blacktriangleright_{rank} \mathcal{D}_2$ if the rank of $\mathcal{D}_1$ is lower than the rank of $\mathcal{D}_2$. A visualization of this on a 2-dimensional space is depicted in Fig. 2. The point $\mathcal{D}_{max}$ here is the property vector that is the most desired one, quite often the property vector that offers the maximum measure of the property for every tuple in the data set. The arcs surrounding $\mathcal{D}_{max}$ are the locus of points at the same distance from $\mathcal{D}_{max}$. Note that any two points on the same arc are incomparable vectors and are assigned the same rank. Points on two different arcs are compared based on how close they are to achieving the most desired property vector. The *rank-based* unary quality index

$$P_{rank}(\mathcal{D}_1) = \| \mathcal{D}_1 - \mathcal{D}_{max} \|$$

can then be used to infer the relationship $P_{rank}(\mathcal{D}_1) < P_{rank}(\mathcal{D}_2) \iff \mathcal{D}_1 \blacktriangleright_{rank} \mathcal{D}_2$. It is also possible to associate a tolerance level ϵ to the rank such that two property vectors differing in rank by ϵ or less are considered equally good.

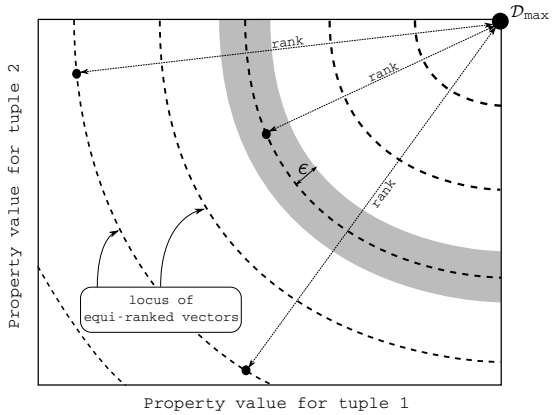

Figure 2: The rank-based $\blacktriangleright_{rank}$-better comparator assigns ranks to property vectors based on the distance from a point of interest $\mathcal{D}_{max}$.

The direct implication of equi-ranked property vectors is that two different anonymizations are equivalent in terms of their ability to pursue the most desirable levels of the property being measured. In other words, the amount of bias each has to overcome (or introduce) to reach the desirable level is equivalent. In a comparative setting, the rank of a

property vector can be viewed as an estimate of the bias present in an anonymization w.r.t. a particular property.

5.2 $\blacktriangleright_{cov}$-better

The *coverage* comparator $\blacktriangleright_{cov}$ compares two property vectors based on the fraction of tuples in one that has a better measurement of the property than in the other. This comparator follows from the intuition that an anonymization better than another should be able to retain higher values of the measured property for more individuals represented in the data set. With this comparator and the equivalence class size property, the aforementioned anonymization $\mathcal{T}_4$ is $\blacktriangleright_{cov}$-better than $\mathcal{T}_{3a}$, and $\mathcal{T}_{3b}$ is $\blacktriangleright_{cov}$-better than $\mathcal{T}_4$. The quality index induced from this comparator is binary in nature, given as

$$P_{cov}(\mathcal{D}_1, \mathcal{D}_2) = \frac{|\{d_i^1 | d_i^1 \geq d_i^2\}|}{\mathcal{N}}$$

and satisfying $P_{cov}(\mathcal{D}_1, \mathcal{D}_2) > P_{cov}(\mathcal{D}_2, \mathcal{D}_1) \iff \mathcal{D}_1 \blacktriangleright_{cov} \mathcal{D}_2$. Note that if $P_{cov}(\mathcal{D}_1, \mathcal{D}_2) = 1$ and $P_{cov}(\mathcal{D}_2, \mathcal{D}_1) = 0$, then $\mathcal{D}_1 \succ \mathcal{D}_2$, and vice versa. Fig. 3 shows the computation of the coverage-based quality index for two hypothetical property vectors.

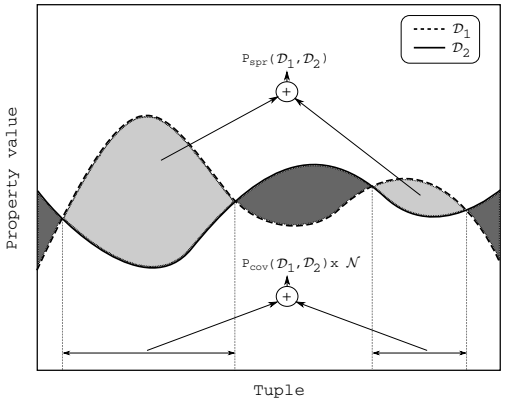

Figure 3: Computation of the quality index functions based on the $\blacktriangleright_{cov}$ and $\blacktriangleright_{spr}$ comparators. P_{cov} is based on the number of tuples with better property value while P_{spr} is based on the actual difference in magnitude of the measured property value.

The coverage comparator is useful when two anonymizations demonstrate similar levels of collective privacy, for example both are k-anonymous for some given k. The comparator then identifies what fraction of the tuples is favored by the skewness in the distribution of privacy levels. A higher value of $P_{cov}(\mathcal{D}_1, \mathcal{D}_2)$, compared to $P_{cov}(\mathcal{D}_2, \mathcal{D}_1)$, implies that more tuples benefit from the skewed distribution of property values in one anonymization than in the other. Such a comparison helps justify that the bias introduced by an anonymization can be useful in providing better property values for a larger fraction of the data set.

5.3 $\blacktriangleright_{spr}$-better

The coverage comparator does not take into consideration the difference in magnitude of a measured property for a given tuple when comparing it across two different property vectors. It is possible that for two vectors $\mathcal{D}_1$ and $\mathcal{D}_2$ with $P_{cov}(\mathcal{D}_1, \mathcal{D}_2) \approx P_{cov}(\mathcal{D}_2, \mathcal{D}_1)$ (the quality index values are close), $\mathcal{D}_1$ maintains much better values in the property for the tuples on which it is superior than $\mathcal{D}_2$ compared to those on which $\mathcal{D}_2$ is superior. For example, consider the hypothetical property vectors $\mathcal{D}_1 = (2, 2, 3, 4, 5)$ and $\mathcal{D}_2 = (3, 2, 4, 2, 3)$. In this case, $P_{cov}(\mathcal{D}_1, \mathcal{D}_2) = P_{cov}(\mathcal{D}_2, \mathcal{D}_1) = \frac{3}{5}$. However, one can argue that $\mathcal{D}_1$ is a superior vector since the difference in magnitude of the measured property is higher (a value of 2) in the two tuples where it is better than $\mathcal{D}_2$. This difference is only 1 in the two tuples where $\mathcal{D}_2$ is better. The *spread* comparator $\blacktriangleright_{spr}$ is based on the total amount of variation (or spread) present between tuples w.r.t. a property. The quality index based on this comparator is given as

$$P_{spr}(\mathcal{D}_1, \mathcal{D}_2) = \sum_{i=1}^{\mathcal{N}} max(d_i^1 - d_i^2, 0)$$

and measures the total difference in magnitude of the measured property for the tuples on which $\mathcal{D}_1$ performs better than $\mathcal{D}_2$. We then say $P_{spr}(\mathcal{D}_1, \mathcal{D}_2) > P_{spr}(\mathcal{D}_2, \mathcal{D}_1) \iff \mathcal{D}_1 \blacktriangleright_{spr} \mathcal{D}_2$. We also have $P_{spr}(\mathcal{D}_1, \mathcal{D}_2) = 0 \iff \mathcal{D}_2 \succeq \mathcal{D}_1$. Fig. 3 shows the difference in computation of the coverage-based and spread-based quality indices.

The spread comparator uses a quantification of the leverage availed by individual tuples from the skewed distribution of property values. This is crucial when the fraction of tuples benefited, as given by the P_{cov} quality index, are equivalent. The P_{spr} quality index quantifies the utilization of the bias in terms of differences in observed property values. Even for the case when two anonymizations have different collective privacy levels, this quantification differentiates them further, often counter to established preferential norms. For example, consider the equivalence class size property vector $(3, 3, 3, 5, 5, 5, 5, 5, 3, 3, 3, 4, 4, 4, 4)$ from a 3-anonymous generalization and $(2, 2, 6, 6, 6, 6, 6, 6, 3, 3, 3, 4, 4, 4, 4)$ from a 2-anonymous generalization. Both anonymizations show signs of bias towards a certain fraction of the tuples. The 3-anonymous generalization will be a typical choice when predefined notions of "better privacy" is used. However, the 2-anonymous generalization achieves better privacy for 6 more tuples (tuples 3 to 8) at the expense of reducing the privacy levels of tuple 1 and 2. This is a reasonable justification to choosing the 2-anonymous generalization instead. The P_{spr} quality index values compare at 2 and 8, thereby revealing this reasoning. In fact, the P_{cov} index also points at the same.

5.4 $\blacktriangleright_{hv}$-better

Another way of measuring the superiority of an anonymization is to ask how good it is against other possible anonymizations apart from the one it is compared to. Such a method refrains itself from performing relative comparisons and instead adopts a "tournament" style mechanism. In a tournament mechanism, a candidate a is preferred over candidate b not because a performs better than b, but because a performs better than a larger number of other candidates than the number of candidates over which b performs better.

For a given property vector $\mathcal{D}_i$, we consider the set of property vectors $\Psi_i = \{\mathcal{D}_j | \mathcal{D}_i \succeq \mathcal{D}_j\}$. Fig. 4 depicts this set, for a 2-dimensional space, as the volume enclosed by all property vectors which are not better than the vector in question under the $\succeq$ comparator. When performing comparisons for two property vectors, the *hypervolume* comparator $\blacktriangleright_{hv}$ assigns superiority based on the hypervolume enclosed by

points which are not superior to both property vectors under the $\succeq$ comparator. In Fig. 4, $\mathcal{D}_1$ is $\succeq$-better than all points in region A, none of which appears in Ψ_2. Similarly, region B has all points that do not appear in Ψ_1 and region C has all points that appear in $\Psi_1 \cap \Psi_2$. Thus, the quality index

$$P_{hv}(\mathcal{D}_1, \mathcal{D}_2) = \prod_{i=1}^{\mathcal{N}} d_i^1 - \prod_{i=1}^{\mathcal{N}} min(d_i^1, d_i^2)$$

is a measurement of the hypervolume on which $\mathcal{D}_1$ is solely $\succeq$-better, giving us the relationship $P_{hv}(\mathcal{D}_1, \mathcal{D}_2) > P_{hv}(\mathcal{D}_2, \mathcal{D}_1) \iff \mathcal{D}_1 \blacktriangleright_{hv} \mathcal{D}_2$. Further, if $P_{hv}(\mathcal{D}_1, \mathcal{D}_2) = 0$ then $\mathcal{D}_2 \succeq \mathcal{D}_1$, and vice versa. Note that the subtraction of the commonly dominated hypervolume is only used to signify what the quality index wishes to compute, but is otherwise not required during a comparison.

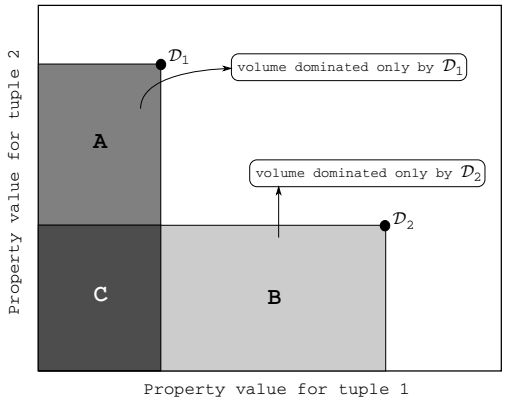

Figure 4: The hypervolume comparator $\blacktriangleright_{hv}$ gives preference to property vectors that solely outperform more property vectors – $\mathcal{D}_2$ in this case since volume of region $B >$ volume of region A.

The hypervolume comparator checks the effectiveness of an anonymization w.r.t. possibly unseen anonymizations. Such anonymizations are captured in the dominated hypervolume. This expands the comparison beyond the anonymizations considered in the comparative process. A higher hypervolume for an anonymization indicates that a larger number of other anonymizations (possibly generated by other algorithms) will induce worse property values than it.

Let us consider the property vector $s = (3, 3, 3, 5, 5, 5, 5, 5)$ induced by an anonymization S for a particular property. Let $t = (4, 4, 4, 4, 4, 4, 4, 4)$ be the property vector induced by anonymization T for the same property. Any anonymization U inducing the property vector $u = (u_1, u_2, \ldots, u_8)$ is worse than S if $u_i < 3; i = 1, 2, 3$ and $u_i < 5; i = 4, \ldots, 8$. The hypervolume is a measure of such anonymizations. Similarly, U is worse than T if $u_i < 4; i = 1, \ldots, 8$. In this case, $P_{hv}(s, t) > P_{hv}(t, s)$ indicating that the number of possible anonymizations that is worse than S is more than that is worse than T. If the volume enclosed by all property vectors is finite, then one can also say that the number of possible anonymizations better than S is less than that is better than T. This method of looking into the effectiveness of an anonymization is complementary to the method behind P_{cov} and P_{spr}. While the latter two facilitates a comparison on a one-to-one basis, comparisons via P_{hv} involves a broader extent of the space of property vectors.

These four quality indices can be used to compare property vectors measuring a single property only. It is true that an anonymization evaluated as being better w.r.t. a particular property can be worse in the context of another. Hence, for the case when more than one property is being measured on an anonymization, other mechanisms are required to weigh the importance of the different properties when making comparisons. We now consider sets of property vectors, instead of a single one, and suggest three indices to compare two such sets. Let us assume that comparisons are to be made across the sets $\Upsilon_1 = \{\mathcal{D}_{11}, \mathcal{D}_{12}, \ldots, \mathcal{D}_{1r}\}$ and $\Upsilon_2 = \{\mathcal{D}_{21}, \mathcal{D}_{22}, \ldots, \mathcal{D}_{2r}\}$ induced by two r–property anonymizations $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively. In the following expressions, we use the notation $P(X, Y)$ to indicate a binary quality index defined to compare two property vectors X and Y. Note that different quality indices can be used to compare different properties. Also, without any loss of generality, we assume that a higher value of the quality index is desirable; otherwise we can negate the index value.

5.5 $\blacktriangleright_{WTD}$-better

The first method to compare sets of property vectors is based on the classical weighted sum approach. Weight assignments are useful when the properties analyzed are orthogonal in nature, such as privacy and utility. The *weight-based* comparator $\blacktriangleright_{WTD}$ requires a vector $W = (w_1, \ldots w_r)$ such that the weight w_i signifies the importance of the i^{th} property being measured. Typically the weights are chosen such that, for $1 \leq i \leq r$, $0 < w_i < 1$ and $\sum_{i=1}^{r} w_i = 1$. The quality index is given as

$$P_{WTD}(\Upsilon_1, \Upsilon_2) = \sum_{i=1}^{r} [w_i \cdot P(\mathcal{D}_{1i}, \mathcal{D}_{2i})]$$

and comparisons are made using the relationship $P_{WTD}(\Upsilon_1, \Upsilon_2) > P_{WTD}(\Upsilon_2, \Upsilon_1) \iff \Upsilon_1 \blacktriangleright_{WTD} \Upsilon_2$. It is advisable to normalize the P values before computing the weighted sum.

Consider the 3-anonymous generalizations in $\mathcal{T}_{3a}$ and $\mathcal{T}_{3b}$. The size of equivalence class property vectors for the two generalizations are $p_a = (3, 3, 3, 3, 4, 4, 4, 3, 3, 4)$ and $p_b = (3, 7, 7, 3, 7, 7, 7, 3, 7, 7)$ respectively. Using Iyengar's data utility metric [7], the utility property vectors for them are $u_a = (2.03, 1.7, 1.7, 2.03, 1.6, 1.6, 1.6, 2.03, 1.7, 1.6)$ and $u_b = (2.03, 0.97, 0.97, 2.03, 0.97, 0.97, 0.97, 2.03, 0.97, 0.97)$ respectively. Using the coverage comparator we have, $P_{cov}(p_a, p_b) = 0.3 < 1 = P_{cov}(p_b, p_a)$ and $P_{cov}(u_a, u_b) = 1 > 0.3 = P_{cov}(u_b, u_a)$. Thus, if equal weights are assigned to both privacy and utility, then generalizations $\mathcal{T}_{3a}$ and $\mathcal{T}_{3b}$ are equally good.

5.6 $\blacktriangleright_{LEX}$-better

The weight-based comparator suffers from the drawback that it may sometimes be difficult to assign weight values specifying the preference of a property. An alternative to this is to instead specify a lexicographic ordering of the different properties. For example, when privacy levels depicted by different privacy models are to be used in conjunction to decide on an anonymization, the property vectors induced from different privacy properties can be ordered in descending order of relevance. Such a method orders the elements of a set of property vectors such that the most desirable property is the first element, the second most desirable property as the second element, and so on. With this ordering in

place, we define the following quality index.

$$P_{LEX}(\Upsilon_1, \Upsilon_2) = \min_i \{1 \leq i \leq r | P(\mathcal{D}_{1i}, \mathcal{D}_{2i}) - P(\mathcal{D}_{2i}, \mathcal{D}_{1i}) > \epsilon_i\}$$

Here, $\epsilon = (\epsilon_1, \ldots, \epsilon_r)$ is a *significance* vector where ϵ_i gives the maximum tolerable difference in the P values for the i^{th} property. Thus, two property vectors $\mathcal{D}_{1i}$ and $\mathcal{D}_{2i}$ are considered to be equally good if $|P(\mathcal{D}_{1i}, \mathcal{D}_{2i}) - P(\mathcal{D}_{2i}, \mathcal{D}_{1i})| \leq \epsilon_i$. The ϵ-*lexicographic* comparator $\blacktriangleright_{LEX}$ assigns superiority to a set of property vectors based on the property on which it is more superior, given the ordering on the properties and the significance vector. $P_{LEX}(\Upsilon_1, \Upsilon_2)$ thereby computes the first property on the ordering where Υ_1 is superior. If $P_{LEX}(\Upsilon_1, \Upsilon_2) < P_{LEX}(\Upsilon_2, \Upsilon_1)$, then Υ_1 has a more desirable property where it is superior to Υ_2, i.e. $P_{LEX}(\Upsilon_1, \Upsilon_2) < P_{LEX}(\Upsilon_2, \Upsilon_1) \iff \Upsilon_1 \blacktriangleright_{LEX} \Upsilon_2$.

5.7 $\blacktriangleright_{GOAL}$-better

The *goal-based* comparator $\blacktriangleright_{GOAL}$ is useful when the competence of an anonymization can be measured in terms of its closeness to a desirable level (or goal). In such a situation, a goal vector $G = (g_1, \ldots, g_r)$ specifies the values desired in the quality indices used – the P functions. The quality index

$$P_{GOAL}(\Upsilon_1, \Upsilon_2) = \sum_{i=1}^{r} [P(\mathcal{D}_{1i}, \mathcal{D}_{2i}) - g_i]^2$$

then computes the sum-of-squares error of the quality indices from the goal values. The comparison is performed with the relationship $P_{GOAL}(\Upsilon_1, \Upsilon_2) < P_{GOAL}(\Upsilon_2, \Upsilon_1) \iff \Upsilon_1 \blacktriangleright_{GOAL} \Upsilon_2$. We can also use unary performance indices as replacements for the binary functions P. The use of unary indices simplifies the specification of the goal vector in terms of the goal property vectors instead. If $\mathcal{D}_{g_1}, \ldots, \mathcal{D}_{g_r}$ are the goal property vectors, then the goal vector can be formulated as $G = (P_1(\mathcal{D}_{g_1}), \ldots, P_r(\mathcal{D}_{g_r}))$, where P_is are unary quality indices.

6. RELATED WORK

Several algorithms have been proposed to find effective k–anonymization [17]. The μ-argus algorithm is based on the greedy generalization of infrequently occurring combinations of *quasi-identifiers* and suppresses outliers to meet the k–anonymity requirement [6]. 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 [16].

Sweeney's Datafly approach uses a heuristic method to generalize the attribute containing the most distinct sequence of values for a provided subset of quasi-identifiers [16]. Sequences occurring less than k times are suppressed. In the same work, Sweeney proposes a theoretical algorithm that can exhaustively search all potential generalizations to find the one that minimally distorts the data during anonymization. Samarati's algorithm [15] can identify all k–minimal generalizations, out of which an optimal generalization can be chosen based on certain preference information provided by the data recipient.

Iyengar proposes a flexible generalization scheme and uses a genetic algorithm to perform k–anonymization on the larger search space that resulted from it [7]. 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 [12]. In order to obtain a guaranteed optimal solution, Bayardo and Agrawal propose a complete search method that iteratively constructs less generalized solutions starting from a completely generalized data set [1].

LeFevre et al. extend the notion of generalizations on attributes to generalization on tuples in the data set [9]. The authors argue that such multidimensional partitioning of the generalization domain shows better performance in capturing the underlying multivariate distribution of the attributes, often advantageous in answering queries with predicates on more than just one attribute.

Xiao and Tao argue that the privacy obtained by a particular generalization may not be sufficient in terms of an individual's personal requirement [21]. To this end, they propose a model where each individual specifies a *guarding node* on attribute values. A generalization scheme is required to prohibit divulgence of any information finer than the value of the guarding node.

The drawbacks of using k–anonymity are first described by Machanavajjhala et al. [13]. 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 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 [10]. 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. Truta and Vinay propose p-sensitive k-anonymity to enforce diversity in the sensitive attribute values in an equivalence class [19]. According to the model, every equivalence class is required to have at least p distinct values of the sensitive attribute. The model suffers from the drawback that attribute values may not be uniformly distributed in a data set and hence obtaining p distinct values in an equivalence class can become impossible.

Dewri et al. [2] presents a multi-objective optimization formulation to explore the privacy and utility trade-offs in microdata anonymization. Their work is based on the notion of *weighted-k–anonymity* which does not constrain k in k–anonymity to a particular value, but rather explores the resulting utility when the weighted equivalence class size varies across anonymizations. Although a potential drawback of their work is the induction of low k (even $k = 1$) values in the solution set, their approach points out that changes in the utility of an anonymization is a resultant of the difference in the distribution of the k values across different tuples of the data set. Another multi-objective analysis is presented by Huang and Du for the problem of optimizing randomized response schemes for privacy protection [5].

Problems similar to the ones encountered here are known to exist in the multi-objective optimization community as well. Quality assessment of solution sets in this community is often difficult because of the existence of non-dominance relationships between one or more members of two differ-

ent sets. Hansen and Jaszkiewicz propose the use of quality measures that induce a linear ordering on the space of all possible solution sets [4]. Knowles and Corne [8] provide a critical overview of existing quality measures and show that most existing measures do not cater to the "ordering" requirement proposed by Hanse and Jaszkiewicz. Later work presented by Zitzler et al. explore the limitations of a comparative study done under the light of *quality indicators* [23]. We have found that a principle theorem proved in their work is equally applicable in the case of anonymization comparisons. The analysis presented in their work serves as a backbone for this study.

7. CONCLUSION AND EXTENSIONS

In this paper, we explore an often ignored factor in the comparison of anonymizations reported by microdata disclosure control algorithms. This factor, which we call the anonymization bias, results in anonymizations being skewed towards a fraction of the data set w.r.t. a privacy property. Hence, the attainment of a collective privacy level is not sufficient to conclude that two anonymizations offer the same level of privacy. We therefore introduce property vectors as an alternate representation of a property (privacy/utility) measured on an anonymization. This representation helps indicate the privacy level of every individual in the data set separately. Further, the issue of comparing anonymizations is addressed by the usage of quality index functions that gives an absolute or a relative quantitative estimate of the quality of property vectors. Our initial conclusion on such a comparative method is that unary quality indices are limited in their ability to perform comparisons, specifically when strict inferences like "not worse than" or "better than" are to be made between anonymizations. As a result, we explore alternative methods of comparison, keeping in mind that comparators defined on such grounds should make adequate effort to quantify the differences in values of the property measured across the tuples of the entire data set instead of a minimalistic estimate. Estimates based on rank, spread and hypervolume are thus suggested. We also present various preference based techniques when comparisons are to be made across multiple properties induced by anonymizations.

An immediate extension of this work is the identification of privacy measures that address the anonymization bias. Moreover, vector-based representations of privacy would require a rethinking of the utility optimization problem since trade-offs between privacy and utility now becomes more apparent. The currently adopted optimization framework in disclosure control is single objective in nature. The scalar quantification of privacy enables the framework to direct its search for higher utility anonymization while satisfying a privacy constraint. If vector representations of privacy are adopted, then the framework has to undergo changes. This is primarily because the vector representation allows one to distinguish between anonymizations even when a typical privacy constraint is satisfied by both. Note that the current framework only makes this distinction based on utility, whereas the vector representation enables the distinction to be made at the privacy front as well. Finding "good" anonymizations thus converts into a multi-objective problem. Although the multi-objective nature of the privacy versus utility problem is well understood in the community, it has remained irrelevant under the pretext of scalar privacy representations. However, under the light of vector representations, privacy should no longer be imposed only as a constraint in the framework but rather handled directly as an objective to maximize. We leave the exploration of this frontier for a later study.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

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

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

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

[4] Hanse, M. P., and Jaszkiewicz, A. 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.

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

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

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

[8] Knowles, J. D., and Corne, D. W. On Metrics for Comparing Non-Dominated Sets. In *Proceedings of the Congress on Evolutionary Computation* (Honolulu, HI, USA, 2002), pp. 711–716.

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

[10] Li, N., Li, T., and Venkatasubramanian, S. t–Closeness: Privacy Beyond k–Anonymity and ℓ–Diversity. In *Proceedings of the 23rd International Conference on Data Engineering* (Istanbul, Turkey, 2007), pp. 106–115.

[11] Loukides, G., and Shao, J. Capturing Data Usefulness and Privacy Protection in

K-Anonymisation. In *Proceedings of the 2007 ACM Symposium on Applied Computing* (Seoul, Korea, 2007), pp. 370–374.

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

[13] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. ℓ–Diversity: Privacy Beyond k–Anonymity. In *Proceedings of the 22nd International Conference on Data Engineering* (Atlanta, GA, USA, 2006), p. 24.

[14] Nergiz, M. E., and Clifton, C. Thoughts on k-anonymization. *Data and Knowledge Engineering 63*, 3 (2007), 622–645.

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

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

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

[18] Takemura, A. 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.

[19] Truta, T. M., and Vinay, B. Privacy Protection: p-Sensitive k-Anonymity Property. In *Proceedings of the 22nd International Conference on Data Engineering Workshops* (Atlanta, GA, USA, 2006), p. 94.

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

[21] Xiao, X., and Tao, Y. Personalized Privacy Preservation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (Chicago, IL, USA, 2006), pp. 229–240.

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

[23] Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., and da Fonseca, V. G. Performance Assessment of Multiobjective Optimizers: An Analysis and Review. *IEEE Transactions on Evolutionary Computation 7*, 2 (2003), 117–132.