A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints

GABRIEL GHINITA, PANAGIOTIS KARRAS, and PANOS KALNIS
National University of Singapore
and
NIKOS MAMOULIS
University of Hong Kong

Recent research studied the problem of publishing microdata without revealing sensitive information, leading to the privacy-preserving paradigms of k-anonymity and ℓ-diversity. k-anonymity protects against the identification of an individual's record. ℓ-diversity, in addition, safeguards against the association of an individual with specific sensitive information. However, existing approaches suffer from at least one of the following drawbacks: (i) ℓ-diversification is solved by techniques developed for the simpler k-anonymization problem, causing unnecessary information loss. (ii) The anonymization process is inefficient in terms of computational and I/O cost. (iii) Previous research focused exclusively on the privacy-constrained problem and ignored the equally important accuracy-constrained (or dual) anonymization problem.

In this article, we propose a framework for efficient anonymization of microdata that addresses these deficiencies. First, we focus on one-dimensional (i.e., single-attribute) quasi-identifiers, and study the properties of optimal solutions under the k-anonymity and ℓ-diversity models for the privacy-constrained (i.e., direct) and the accuracy-constrained (i.e., dual) anonymization problems. Guided by these properties, we develop efficient heuristics to solve the one-dimensional problems in linear time. Finally, we generalize our solutions to multidimensional quasi-identifiers using space-mapping techniques. Extensive experimental evaluation shows that our techniques clearly outperform the existing approaches in terms of execution time and information loss.

Categories and Subject Descriptors: H.2.0 [**Database Management**]: General—*Security, integrity, and protection*

General Terms: Design, Experimentation, Security

Additional Key Words and Phrases: Privacy, anonymity

This work was partially supported by grant HKU 715108E from Hong Kong RGC.

Authors' addresses: G. Ghinita, P. Karras, P. Kalnis, National University of Singapore, Computing 1, Computing Drive, Singapore 117417; email: {ghinitag,karras,kalnis}@comp.nus.edu.sg; N. Mamoulis, University of Hong Kong, Pokfulam Road, Hong Kong; email: nikos@cs.hku.hk.

DOI 10.1145/1538909.1538911 http://doi.acm.org/10.1145/1538909.1538911

ACM Reference Format:
Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. 2009. A framework for efficient data anonymization under privacy and accuracy constraints. ACM Trans. Database Syst., 34, 2, Article 9 (June 2009), 47 pages, DOI = 10.1145/1538909.1538911 http://doi.acm.org/10.1145/1538909.1538911

1. INTRODUCTION

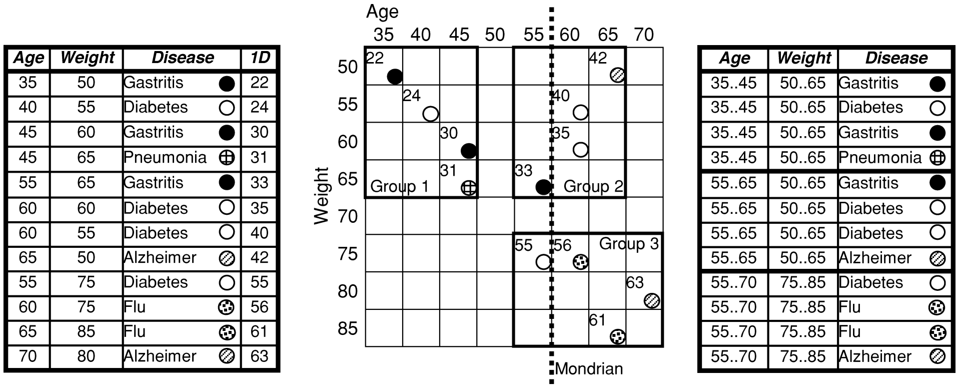
Organizations, such as hospitals, need to release microdata (e.g., medical records) for research and other public benefit purposes. However, sensitive personal information (e.g., medical condition of a specific person) may be revealed in this process. Conventionally, identifying attributes such as name or social security number are not disclosed, in order to protect privacy. Still, recent research [Froomkin 2000; Sweeney 2002] has demonstrated that this is not sufficient, due to the existence of *quasi-identifiers* in the released microdata. Quasi-identifiers are sets of attributes (e.g., ⟨ZIP, Gender, DateOfBirth⟩) which can be joined with information obtained from diverse sources (e.g., public voting registration data) in order to reveal the identity of individual records.

To address this threat, Samarati [2001] and Sweeney [2002] proposed the k-anonymity model: For every record in a released table there should be at least $k - 1$ other records identical to it along a set of quasi-identifying attributes. Records with identical quasi-identifier values constitute an *equivalence class*. k-anonymity is commonly achieved either by generalization (e.g., show only the area code instead of the exact phone number) or suppression (i.e., hide some values of the quasi-identifier), both of which inevitably lead to information loss. Still, the data should remain as accurate as possible in order to be useful in practice. Hence a trade-off between privacy and information loss emerges.

Recently, the concept of ℓ-diversity [Machanavajjhala et al. 2006] was introduced to address the limitations of k-anonymity. The latter may disclose sensitive information when there are many identical Sensitive Attribute (*SA*) values within an equivalence class[1] (e.g., all persons suffer from the same disease). ℓ-diversity prevents uniformity and background knowledge attacks by ensuring that at least ℓ *SA* values are *well represented* in each equivalence class (e.g., the probability to associate a tuple with an *SA* value is bounded by $1/\ell$ [Xiao and Tao 2006a]). Machanavajjhala et al. [2006] suggest that any k-anonymization algorithm can be adapted to achieve ℓ-diversity. However, the following example demonstrates that such an approach may yield excessive information loss.

Consider the privacy-constrained anonymization problem for the microdata in Figure 1(a), where the combination of ⟨*Age*, *Weight*⟩ is the quasi-identifier and Disease is the sensitive attribute. Let the required privacy constraint, within the k-anonymity model, be $k = 4$. The current state-of-the-art k-anonymization

[1] k-anonymity remains a useful concept, suitable for cases where the sensitive attribute is implicit or omitted (e.g., a database containing information about convicted persons, regardless of specific crimes).



Age	Weight	Disease	1D
35	50	Gastritis ●	22
40	55	Diabetes ○	24
45	60	Gastritis ●	30
45	65	Pneumonia ⊕	31
55	65	Gastritis ●	33
60	60	Diabetes ○	35
60	55	Diabetes ○	40
65	50	Alzheimer ⊘	42
55	75	Diabetes ○	55
60	75	Flu ⊛	56
65	85	Flu ⊛	61
70	80	Alzheimer ⊘	63

(a) original microdata

(b) k-anonymous groups

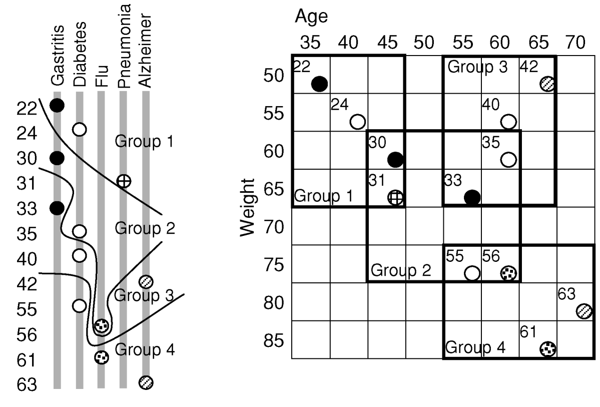
Age	Weight	Disease
35..45	50..65	Gastritis ●
35..45	50..65	Diabetes ○
35..45	50..65	Gastritis ●
35..45	50..65	Pneumonia ⊕
55..65	50..65	Gastritis ●
55..65	50..65	Diabetes ○
55..65	50..65	Diabetes ○
55..65	50..65	Alzheimer ⊘
55..70	75..85	Diabetes ○
55..70	75..85	Flu ⊛
55..70	75..85	Flu ⊛
55..70	75..85	Alzheimer ⊘

(c) published data

Fig. 1. k-anonymization example ($k = 4$).

algorithm (i.e., Mondrian [LeFevre et al. 2006a]) sorts the data points along each dimension (i.e., *Age* and *Weight*), and partitions across the dimension with the widest normalized range of values. In our example, the normalized ranges for both dimensions are the same. Mondrian selects the first one (i.e., Age) and splits it into segments $35-55$ and $60-70$ (see Figure 1(b)). Further partitioning is not possible because any split would result in groups with less than 4 records. We propose a different approach. First, we map the multidimensional quasi-identifier to a 1D value. In this example we use an 8 × 8 Hilbert space filling curve (see Section 6 for details); other mappings are also possible. The resulting sorted 1D values are shown in Figure 1(a) (column 1D). Next, we partition the 1D space. We prove that the optimal 1D partitions are nonoverlapping and contain between k and $2k - 1$ records. We obtain 3 groups which correspond to 1D ranges [22..31], [33..42], and [55..63]. The resulting 2D partitions are enclosed by three rectangles in Figure 1(b). In this example, our method causes less information loss because the extents of the obtained groups are smaller than in the case of Mondrian. For instance, consider the query "Find how many persons are in the age segment $35-45$ and weight interval $50-60$": The correct answer is 3. Assuming that records are uniformly distributed within each group, our method returns the answer $4 \times 9/12 = 3$ (there are 4 records in $Group_1$, 9 data space cells that match the query, and a total of 12 cells in $Group_1$). On the other hand, the answer obtained with Mondrian is $6 \times 9/40 = 1.35$ (from the group situated to the left of the dotted line). Clearly, our k-anonymization algorithm is more accurate.

The advantages of our approach are even more prominent with the ℓ-diversification problem. This problem is more difficult because, in order to cover a variety of SA values, the optimal 1D partitioning may have to include overlapping ranges. For example, if ℓ=3, group 2 in Figure 2(a) contains tuples {30, 35, 56}, whereas the third group contains tuples {33, 40, 42}. Nevertheless, we prove that there exist optimal partitionings consisting of only consecutive ranges with respect to each individual value of the sensitive attribute. Based on this property, we develop a heuristic which essentially groups together records



(a) 1D l-diversity

(b) l-diverse groups

Age	Weight	Disease
35..45	50..65	Gastritis ●
35..45	50..65	Diabetes ○
35..45	50..65	Pneumonia ⊕
45..60	60..75	Gastritis ●
45..60	60..75	Diabetes ○
45..60	60..75	Flu ⊛
55..65	50..65	Gastritis ●
55..65	50..65	Diabetes ○
55..65	50..65	Alzheimer ⊘
55..70	75..85	Diabetes ○
55..70	75..85	Flu ⊛
55..70	75..85	Alzheimer ⊘

(c) published data

Fig. 2. ℓ-diversification example ($\ell = 3$).

that are close to each other in the 1D space, but have different sensitive attribute values. The four resulting groups[2] are shown in Figure 2(b). From the result we can infer, for instance, that no person younger than 55 suffers from Alzheimer's. On the other hand, if we use Mondrian, we cannot partition the space at all because any possible disjoint partitioning would violate the ℓ-diversity property. For example, if the Age axis was split into segments $35 - 55$ and $60 - 70$ (i.e., as in the k-anonymity case), then gastritis would appear in the left-side partition with probability $3/6$, which is larger than the allowed $1/\ell = 1/3$. Since Mondrian includes all tuples in the same partition, young or old persons are ascribed the same probability to suffer from Alzheimer's. Obviously the resulting information loss is unacceptable.

The previous example demonstrates that existing techniques for the privacy-constrained k-anonymization problem, such as Mondrian, are not appropriate for the ℓ-diversification problem. In Section 2 we also explain that Anatomy [Xiao and Tao 2006a], which is an ℓ-diversity-specific method, exhibits high information loss, despite relaxing the privacy requirements (i.e., it publishes the *exact* quasi-identifier). Moreover, while our techniques resemble clustering, our experiments show that existing clustering-based anonymization techniques (e.g., Xu et al. [2006]) are worse in terms of information loss and considerably slower.

So far, research efforts focused on the privacy-constrained anonymization problem, which minimizes information loss for a given value of k or ℓ; we call this the *direct* anonymization problem. However, the resulting information loss may be high, rendering the published data useless for specific applications. In practice, the data recipient may require certain bounds on the amount of information loss. For instance, it is well known that the occurrence of certain diseases is highly correlated to age (e.g., Alzheimer's can only occur in elderly patients). To ensure that anonymized hospital records make practical sense,

[2] Note that although groups may overlap in their quasi-identifier extents, each record belongs to *exactly one* group.

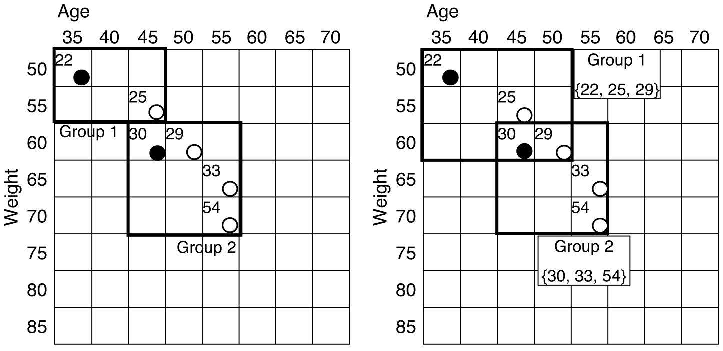
```
Iterative Privacy-Constrained Solution for the Dual problem (IPCSD)
Input: set of records $\mathcal{R}$, accuracy bound $E$, precision threshold $Thr$
1. $\ell_{min} = 1$, $\ell_{max} = |\mathcal{R}|/(number\ of\ occurrences\ of\ most\ freq.\ SA\ value\ in\ \mathcal{R})$
2. do
3.    $\ell = (\ell_{max} + \ell_{min})/2$
4.    $\mathcal{P}$ =run $\ell$-diversification (i.e., privacy-constrained) on $\mathcal{R}$ with parameter $\ell$
5.    if $(InformationLoss(\mathcal{P}) \leq E)$ then $\ell_{min} = \ell$
6.    else $\ell_{max} = \ell$
7. while$(\ell_{max} - \ell_{min} > Thr)$
8. output $\ell_{min}$
```

Fig. 3. Iterative privacy-constrained solution for the accuracy-constrained (i.e., dual) problem.

a medical researcher may require that no anonymized group should span a range on attribute *Age* larger than 10 years. Motivated by such scenarios, we introduce the accuracy-constrained or dual anonymization problem. Let E be the maximum acceptable amount of information loss (the metric is formally defined in Section 2). The accuracy-constrained anonymization problem finds the maximum degree of privacy (i.e., k or ℓ) that can be achieved such that information loss does not exceed E. Subsequently, the data publisher can assess whether the attainable privacy under this constraint is satisfactory, and can decide whether it makes sense to publish the data at all. To the best of our knowledge, the dual problem has not been addressed previously, despite its important practical applications.

A possible solution for the dual problem is to use an existing method for privacy-constrained anonymization, as shown in Figure 3 (we consider the ℓ-diversity case). The algorithm, called Iterative Privacy-Constrained Solution for the Dual problem (*IPCSD*), performs a binary search to find the maximum value of ℓ for which the information loss does not exceed E. The ℓ_{min} value of 1 (line 1) corresponds to no privacy, whereas ℓ_{max} is the maximum achievable privacy, and is a characteristic of the dataset. As we will formally discuss in Section 2.3, ℓ_{max} is equal to the total number of records divided by the number of occurrences of the *SA* value with the highest frequency. The algorithm stops when the search interval for the ℓ value is reduced below a certain threshold *Thr*. IPCSD is a generic solution that can be used in conjunction with any privacy-constrained ℓ-diversification method, such as our proposed 1D method described earlier, or Mondrian. The invocation of a particular method is done in line 4 of the pseudocode.

Because it is not specifically tailored for the dual problem, IPCSD can yield unsatisfactory results. Consider the example of Figure 4 and assume the E bound requires that the span of each group along any of the quasi-identifier attributes does not exceed 15. IPCSD, used in conjunction with Mondrian, will give the result in Figure 4(a), with a maximum achievable privacy metric of $\ell = 4/3$ (ℓ is the inverse of the maximum association probability between a record and an *SA* value, which is 3/4 for group 2). It is easy to see that all splits, except that between *Weight* values 55 and 60, leave on one side of the split only records with the same *SA* value, hence association probability is 100% (no privacy). The solution depicted in the example is the only one where



(a) IPCSD (b) our solution

Fig. 4. Accuracy-constrained ℓ-diversification example.

Mondrian offers some amount of privacy. As we will show in our experimental evaluation, IPCSD used in conjunction with our direct 1D method suffers from similar drawbacks.

In general, IPCSD (in conjunction with any privacy-constrained method) fails to find a good solution because it imposes the privacy bound on all groups and locally minimizes their extents. This approach is not useful for the accuracy-constrained problem, where we would prefer to preserve large extents (as long as the E bound is satisfied) and locally maximize privacy. Furthermore, the binary search process employed by IPCSD is based on the assumption that information loss increases with ℓ; this assumption holds in theory, and would be realized by an optimal privacy-constrained ℓ-diversification algorithm. Nevertheless, existing techniques are not optimal, hence do not guarantee this monotonicity. Therefore, IPCSD may yield low privacy if the search finds a local maximum of ℓ under constraint E, instead of a global one. Finally, the iterative search process may result in a large computational overhead in practice.[3]

Motivated by these limitations, we propose an efficient, specialized algorithm to solve the accuracy-constrained problem. We map the multidimensional dual problem to 1D and study the properties of an optimal 1D solution. Based on these properties we develop a heuristic that, for an information loss bound E, achieves considerably better privacy than IPCSD. On the same example, our method produces the solution of Figure 4(b). The obtained ℓ value is $3/2$, corresponding to a probability of association between a record and an SA value of $2/3 = 66\%$ (because there are at most two records with the same SA value in each group of three records).

1.1 Contributions

We present a framework for solving efficiently the privacy-constrained (i.e., direct) and accuracy-constrained (i.e., dual) anonymization problems, by mapping

[3] An attempt to overcome the lack of monotonicity using global optimization techniques (e.g., simulated annealing) would incur even larger overhead, and still not guarantee finding the global maximum privacy.

the multidimensional quasi-identifiers to 1D space.[4] Specifically:

(i) For k-anonymization, we develop an optimal algorithm for the direct problem with 1D quasi-identifiers, which has running time linear in the size of the dataset.

(ii) For the more complex direct ℓ-diversification problem, we study theoretically the properties of possible optimal 1D solutions. Guided by these properties, we propose an efficient heuristic algorithm with linear-time complexity in the data size.

(iii) We study the accuracy-constrained problem in the context of the ℓ-diversity paradigm. We derive a polynomial-time optimal solution in the 1D space, and propose an efficient 1D heuristic which runs linearly in the data size. Our work is the first to address the dual anonymization problem.

(iv) We generalize our algorithms to multidimensional quasi-identifiers,[5] by mapping them to 1D space. Given a sorted input, the I/O cost is very low, since our algorithms scan the data only once. As case studies, we consider mappings based on the Hilbert space filling curve and iDistance [Zhang et al. 2005].

(v) The experimental results show that our algorithms consistently outperform existing generalization methods by a wide margin, in terms of both information loss and running time.

The rest of this article is organized as follows: Section 2 contains essential definitions and surveys the related work. Section 3 and Section 4 present our solutions for the 1D privacy-constrained problem under the k-anonymity and ℓ-diversity paradigms, respectively. Section 5 addresses the 1D accuracy-constrained problem. Section 6 extends our algorithms to the general case of multidimensional quasi-identifiers. Section 7 presents the experimental evaluation and Section 8 concludes the article.

2. BACKGROUND AND RELATED WORK

This section introduces the data model and terminology used in the article, and presents the related work. For the ease of reference, Table I summarizes the notations used throughout the work.

Definition 1 (*Quasi-Identifier*). Given a database table $T(A_1, A_2, \ldots, A_n)$, a quasi-identifier attribute set $Q_T = \{A_1, A_2, \ldots, A_d\} \subseteq \{A_1, A_2, \ldots, A_n\}$ is a set of attributes that can be joined with external information in order to reveal the personal identity of individual records [Samarati 2001; Sweeney 2002].

A set of tuples that are indistinguishable in the projection of T on Q_T is called an *equivalence class* or, alternatively, an anonymized group. Two

[4] This work is an extended version of Ghinita et al. [2007]. Additional contributions include the novel accuracy-constrained (i.e., dual) anonymization problem (Section 2.5), an optimal solution as well as an heuristic for the dual problem (Section 5), and the corresponding experimental evaluation (Section 7.3). We also added an optimal algorithm for ℓ-diversification (Section 4.2).

[5] We emphasize that our optimal solutions apply for 1D quasi-identifiers only; in the multidimensional space, we propose heuristic solutions.

Table I. Summary of Notations

Symbol	Description
k	Degree of Anonymity
ℓ	Degree of Diversity
E	Accuracy Bound
G	k-anonymous/ℓ-diverse group
Q_T	quasi-identifier
$NCP(G)$	Normalized Certainty Penalty of group G
$\mathcal{P}$	Partitioning of data into groups
$GCP(\mathcal{P})$	Global Certainty Penalty of Partitioning $\mathcal{P}$
$\mathcal{IL}_1$	Average-Extent Information Loss Metric
$\mathcal{IL}_\infty$	Maximum-Extent Information Loss Metric
$\mathcal{PM}$	Privacy Metric
m	Cardinality of sensitive attribute domain
N	Dataset Cardinality
$D_1, \ldots, D_m$	Data domains based on SA value
G^q	sub-set of G in domain D_q, $1 \leq q \leq m$
r_i	record with index i in the data sequence
b_i, e_i	*begin* and *end* records of group G_i
$\mathbf{b_i}, \mathbf{e_i}$	*begin* and *end* boundaries of group G_i
$\mathbf{enditem(a)}$	*end* item of boundary $\mathbf{a}$
$I_{r_i}^E$	E-extent interval ending at r_i

commonly employed techniques to create anonymized groups are generalization and suppression[6] [Sweeney 2002]. Generalization defines equivalence classes for tuples as multidimensional ranges in the Q_T space, and replaces their actual Q_T values with a representative value of the whole range of the class (e.g., replaces the city with the state). Generalization ranges are usually specified by a generalization hierarchy, or taxonomy tree (e.g., city→state→country). Suppression excludes some Q_T attributes or entire records (known as outliers) from the microdata.

The privacy-preserving transformation of the microdata is referred to as recoding. Two models exist: In global recoding, a particular detailed value must be mapped to the same generalized value in all records. Local recoding, on the other hand, allows the same detailed value to be mapped to different generalized values in each equivalence class. Local recoding is more flexible and has the potential to achieve lower information loss [LeFevre et al. 2006a]. The recoding process can also be classified into single-dimensional, where the mapping is performed for each attribute individually, and multidimensional, which maps the Cartesian product of multiple attributes. Multidimensional mappings are more accurate; nevertheless initial research focused on single-dimensional ones due to simplicity. In this article, we develop local recoding, multidimensional transformations.

All privacy-preserving transformations cause information loss, which must be minimized in order to maintain the ability to extract meaningful information from the published data. Next we discuss suitable information loss metrics.

[6] Permutation is another alternative. We review permutation-based methods for ℓ-diversity in Section 2.3.

2.1 Information Loss Metrics

A variety of information loss metrics have been proposed. The Classification Metric (*CM*) [Iyengar 2002] is suitable when the purpose of the anonymized data is to train a classifier. Each record is assigned a class label, and information loss is computed based on the adherence of a tuple to the majority class of its group. However, it is not clear how *CM* can be extended to support general purpose applications. The Discernibility Metric (*DM*) [Bayardo and Agrawal 2005], on the other hand, measures the cardinality of the equivalence class. Although classes with few records are desirable, *DM* does not capture the distribution of records in the Q_T space. More accurate are the Generalized Loss Metric [Iyengar 2002] and the similar Normalized Certainty Penalty (*NCP*) [Xu et al. 2006]. The latter factors in the extent of each class in the Q_T space. For numerical attributes the *NCP* of an equivalence class G is defined as

$$NCP_{A_{Num}}(G) = \frac{max^G_{A_{Num}} - min^G_{A_{Num}}}{max_{A_{Num}} - min_{A_{Num}}},$$

where the numerator and denominator represent the ranges of attribute A_{Num} for the class G and the entire attribute domain, respectively. In the case of categorical attributes, where no total order or distance function exists, NCP is defined with respect to the taxonomy tree of the attribute. We have

$$NCP_{A_{Cat}}(G) = \begin{cases} 0, & card(u) = 1 \\ card(u)/|A_{Cat}|, & otherwise \end{cases},$$

where u is the lowest common ancestor of all A_{Cat} values included in G, $card(u)$ is the number of leaves (i.e., attribute values) in the subtree of u, and $|A_{Cat}|$ is the total number of distinct A_{Cat} values. The *NCP* of class G over all quasi-identifier attributes is

$$NCP(G) = \sum_{i=1}^{d} w_i \cdot NCP_{A_i}(G), \tag{1}$$

where d is the number of attributes in Q_T (i.e., the dimensionality). A_i is either a numerical or categorical attribute and has a weight w_i, where $\sum w_i = 1$.

NCP measures information loss for a single equivalence class. Xu et al. [2006] characterize the information loss of an entire partitioning by summing the *NCP* over all tuples in each group. For the sake of comparison with previous work, we adopt a normalized formulation of the aggregate version of *NCP*, called the Global Certainty Penalty (*GCP*). Let partitioning $\mathcal{P}$ be the set of all equivalence classes in the released anonymized table. The *GCP* for $\mathcal{P}$ is defined as

$$GCP(\mathcal{P}) = \frac{\sum_{G \in \mathcal{P}} |G| \cdot NCP(G)}{d \cdot N}, \tag{2}$$

where N denotes the number of records in the original table (i.e., microdata), $|G|$ is the cardinality of group G, and d is the dimensionality of Q_T. The advantage of this formulation is its ability to measure information loss among tables with varying cardinality and dimensionality. Furthermore, *GCP* is between 0 and

1, where 0 signifies no information loss (i.e., the original microdata) and 1 corresponds to total information loss (i.e., there is only one equivalence class covering all records in the table).

Furthermore, in addition to adopting *GCP*, we introduce a broader class of information loss metrics expressed as Minkowski-norms on group extents. In particular, we focus on the average-extent metric

$$\mathcal{IL}_1(\mathcal{P}) = \text{avg}_{G\in\mathcal{P}} \left(max^G_{Q_T} - min^G_{Q_T}\right) \tag{3}$$

and the maximum-extent metric

$$\mathcal{IL}_\infty(\mathcal{P}) = \max_{G\in\mathcal{P}} \left(max^G_{Q_T} - min^G_{Q_T}\right). \tag{4}$$

To facilitate presentation, we refer to these metrics using the unified notation $\mathcal{IL}$. Both metrics, as well as *GCP*, have the following property.

Definition 2 (*Superadditivity*). Given an equivalence class G and two subsets G_1 and G_2 such that $G = G_1 \cup G_2$ and $G_1 \cap G_2 = \emptyset$, an information loss metric $\mathcal{IL}$ is called superadditive if $\mathcal{IL}(\mathcal{P}) \geq \mathcal{IL}((\mathcal{P}\diagdown\{G\}) \cup \{G_1\} \cup \{G_2\})$.

In other words, a superadditive information loss metric has the property that applying additional group divisions will never degrade the quality of the partitioning [Muthukrishnan and Suel 2005].

We develop our theoretical results mainly on top of the $\mathcal{IL}$ metrics. Some of our optimal solutions hold for the *GCP* metric as well, whereas others can be extended as heuristics that work with *GCP* and yield low information loss in practice.

2.2 Privacy-Constrained k-Anonymization

Definition 3 (*k-Anonymity*). A database table T with a quasi-identifier attribute set Q_T conforms to the k-anonymity property, if and only if each unique tuple in the projection of T on Q_T occurs at least k times [Samarati 2001; Sweeney 2002].

An optimal solution to the k-anonymization problem should minimize information loss. Formally, we have the next problem.

Problem 1 *(Privacy-Constrained k-Anonymization)*. Given a table T, a quasi-identifier set Q_T, and a privacy bound expressed as the degree of anonymity k, determine a partitioning $\mathcal{P}$ of T such that each partition $G \in \mathcal{P}$ has at least k records, and $\mathcal{IL}(\mathcal{P})$ is minimized.

Meyerson and Williams [2004] proved that optimal k-anonymization for multidimensional quasi-identifiers is NP-hard under the suppression model. They proposed an approximate algorithm that minimizes the number of suppressed values; the approximation bound is $O(k \cdot log\, k)$. Aggarwal et al. [2005] improved this bound to $O(k)$, whereas Park and Shim [2007] proposed an algorithm that achieves an $O(\log k)$-approximation bound, but runs in time exponential to k. The work in Byun et al. [2007] proves that k-anonymization is *NP*-hard under the generalization model as well (an information loss metric similar to *GCP* is used), by showing that suppression is a special case of generalization. Several generalization approaches limit the search space by considering only global recoding. Bayardo and Agrawal [2005] proposed an optimal algorithm for single-dimensional global recoding with respect to the CM and DM metrics. Incognito [LeFevre et al. 2005] takes a dynamic programming approach, and finds an optimal solution for any metric by considering all possible generalizations, but only for global, single-dimensional recoding.

To address the inflexibility of single-dimensional recoding, Mondrian [LeFevre et al. 2006a] employs multidimensional global recoding, which achieves finer granularity. Mondrian partitions the space recursively across the dimension with the widest normalized range of values. Mondrian can also support a limited version of local recoding: If many points fall on the boundary of two anonymized groups, they may be divided between the two groups. Because Mondrian uses space partitioning, the data points within a group are not necessarily close to each other in the Q_T space (e.g., points 22 and 55 in Figure 1(b)), causing high information loss. In Iwuchukwu and Naughton [2007], data records are bulk-loaded into a R$^+$-tree index, and each resulting leaf node corresponds to an anonymized group. Similar to Mondrian, this technique employs multidimensional global recoding (since R$^+$-tree leaf nodes do not overlap).

Most existing multidimensional local recoding methods are based on clustering. In Aggarwal et al. [2006] k-anonymization is treated as a special clustering problem, called r-cellular clustering. A constant factor approximation of the optimal solution is proposed, but the bound only holds for the Euclidean distance metric. Furthermore, the computation and I/O cost are high in practice. Xu et al. [2006] propose agglomerative and divisive recursive clustering algorithms, which attempt to minimize the *NCP* metric. The latter (called TopDown in the following) is the best of the two. TopDown performs a two-step clustering: First, all records are in one cluster which is recursively divided as long as there are at least $2k$ records in each cluster. In the second step, the clusters with less than k members are either grouped together, or they borrow records from clusters with more than k records. The complexity of TopDown is $O(N^2)$. In our experiments, we show that TopDown is inefficient in terms of information loss and computational cost. Independently of our work, Wong et al. [2006] proposed a solution to privacy-constrained k-anonymization based on dimensionality reduction, similar to the one we introduce in Section 3. However, their work deals only with k-anonymity. Moreover, the cost is quadratic to the database size, whereas our solution has linear complexity.

2.3 Privacy-Constrained ℓ-Diversification

A database table T with a quasi-identifier attribute set Q_T and a Sensitive Attribute *SA* conforms to the ℓ-diversity property if and only if each equivalence class in T with respect to Q_T has at least ℓ well-represented values of the sensitive attribute. Machanavajjhala et al. [2006] proposed two interpretations of "well-represented values": *entropy* ℓ-diversity and recursive (c,ℓ)-diversity. The former yields tighter privacy constraints, but is too restrictive for practical purposes. The latter is a more relaxed condition: an equivalence class G is ℓ-diverse if $f_1 < c(f_\ell + f_{\ell+1} + .. + f_m)$, where c is a constant, f_i is the number of occurrences of the i^{th} most frequent value of SA in G, and m is the number of distinct values in SA. In order for an ℓ-diverse partitioning to exist, the original table T must itself satisfy the aforesaid condition, referred to as the Eligibility Condition (EG).

In practice, the privacy threat to a certain database record is expressed as the probability of associating an anonymized record with a certain value $s \in SA$; we denote this breach probability by P_{br}. Given an equivalence class G,

$$P_{br} = occ^G_{max}/|G|, \tag{5}$$

where occ^G_{max} is the maximum number of occurrences over all SA values in G. Since P_{br} is directly relevant to the privacy of records, it is desirable to have an ℓ-diversity formulation that can be linked to P_{br}. We therefore adopt the following definition from Xiao and Tao [2006a].

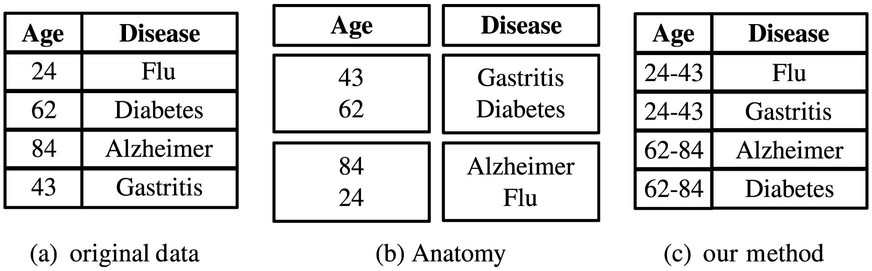
Definition 4 (*ℓ-Diversity*). An equivalence class G has the ℓ-diversity property, if the probability of associating a record in G with any particular sensitive attribute value[7] is at most $1/\ell$.

The privacy-constrained or direct ℓ-diversification problem is formally defined as:

Problem 2 *(Privacy-Constrained ℓ-Diversification).* Given a table T, a quasi-identifier Q_T, a sensitive attribute SA, and a privacy bound expressed as the degree of diversity ℓ, determine a partitioning $\mathcal{P}$ of T such that each equivalence class $G \in \mathcal{P}$ satisfies the ℓ-diversity property and $\mathcal{IL}(\mathcal{P})$ is minimized.

Machanavajjhala et al. [2006] implement ℓ-diversity on top of Incognito and suggest that any k-anonymization technique can be adapted for ℓ-diversification. However, as we demonstrated in the example of Figure 2, k-anonymity techniques may result in unacceptable information loss, due to the requirement of diverse SA values. Anatomy [Xiao and Tao 2006a] is an ℓ-diversity-specific method. It hashes records into buckets according to the SA value, and builds partitions by randomly selecting ℓ records from distinct buckets; the complexity is $O(|T|)$. Anatomy has two drawbacks: (i) it releases the *exact* quasi-identifiers of records. While this does not violate the ℓ-diversity requirement, it confirms that a particular individual is included in the data. Consider, for instance, a dataset containing quasi-identifiers of convicted persons and their crime. Although Anatomy hides the exact crime, an attacker can still conclude that a specific person has been convicted. Therefore, Anatomy is not suitable for applications that require protection against record linkage. (ii) anatomy does not consider the extent of each partition in the Q_T space, hence information loss may be high. Consider the medical dataset in Figure 5(a), with quasi-identifier Age and sensitive attribute $Disease$. Note that, in the original data only patients more than 80 years old can suffer from Alzheimer's. Assume $\ell = 2$: Anatomy may randomly choose to group together records 1 with 3 and

[7] Under this definition, the eligibility condition requires that at most $|T|/\ell$ tuples in the original table T have the same SA value.



Age	Disease
24	Flu
62	Diabetes
84	Alzheimer
43	Gastritis

(a) original data

Age	Disease
43 62	Gastritis Diabetes
84 24	Alzheimer Flu

(b) Anatomy

Age	Disease
24-43	Flu
24-43	Gastritis
62-84	Alzheimer
62-84	Diabetes

(c) our method

Fig. 5. Privacy-constrained ℓ-diversification example. In the microdata only patients over 80 years old can suffer from Alzheimer's.

2 with 4, resulting in the anatomized table of Figure 5(b). The published data suggests that Alzheimer's is equally probable for young and old persons alike: for instance, the average age for Alzheimer's patients that can be inferred from the anatomized data is 54. In contrast, our approach obtains the generalized table in Figure 5(c), and implies that Alzheimer's is only possible for elderly patients (the average age in the group that includes Alzheimer's is 73).

2.4 Other Privacy-Constrained Anonymization Approaches

Like Anatomy, the work of Zhang et al. [2007] publishes the exact Q_T. It focuses on *SA*s with numerical values and deals with situations where these values are similar; its drawbacks are analogous to Anatomy's. Another recent work [Li et al. 2007] proposes a new privacy paradigm called t-closeness, which dictates that the table-wise distribution of *SA* values should be reproduced within each anonymized group. No specific technique is proposed; instead, it is suggested to modify existing k-anonymization techniques. However, this is expected to face the same drawbacks as the application of k-anonymization techniques to ℓ-diversification. Xiao and Tao [2007] propose m-invariance, a privacy paradigm that addresses correlation attacks among multiple versions of data released at different timestamps. Yet another model is described in Xiao and Tao [2006b], where each record in the table has an individual privacy constraint. However, in order to enforce privacy, *SA* values must also be generalized. Kifer and Gehrke [2006] propose a method for publishing anonymized marginals, in addition to microdata. Marginals are summaries of the original table that may improve accuracy. Anonymizing the marginals is orthogonal to anonymizing the microdata.

2.5 Accuracy-Constrained Problems

Previous research has focused exclusively on the privacy-constrained problem. In this section we introduce the accuracy-constrained anonymization problem in the context of both k-anonymity and the ℓ-diversity paradigms. Accuracy-constrained anonymization maximizes privacy for a given bound E of acceptable information loss per group.

The privacy metric $\mathcal{PM}$ of an entire partitioning $\mathcal{P}$ is as good as the lowest privacy achieved by any group in $\mathcal{P}$. In the context of k-anonymity, the privacy metric of group G is defined as group cardinality $|G|$. On the other hand, in the context of ℓ-diversity, the privacy metric of group G is defined as $1/P_{br}$, where P_{br} (Eq. (5)) is the probability of associating an individual with a specific *SA* value. Formally,

$$\mathcal{PM}_{div}(\mathcal{P}) = \min_{G \in \mathcal{P}} \mathcal{PM}_{div}(G) = \min_{G \in \mathcal{P}} \frac{|G|}{occ^{G}_{max}}. \tag{6}$$

The accuracy-constrained problem is defined next.

Problem 3 *(Accuracy-Constrained Problem).* Given a table T, a quasi-identifier Q_T, a privacy metric $\mathcal{PM}$, and an information loss bound E, determine a partitioning $\mathcal{P}$ of T such that every group satisfies the bound E and $\mathcal{PM}(\mathcal{P})$ is maximized.

Note that the nature of the dual problem dictates the enforcement of the accuracy bound for maximum-loss-per-group metrics, such as $\mathcal{IL}_\infty$, or maximum *NCP* over all groups.

3. OPTIMAL 1D PRIVACY-CONSTRAINED K-ANONYMIZATION

In this section we present an optimal solution to the direct k-anonymization problem for 1D quasi-identifiers. Although the problem is NP-hard in the general case [Byun et al. 2007], we show that the complexity is linear in the size of the input for 1D quasi-identifiers. In Section 6 we will use the 1D solution as an heuristic in multiple dimensions.

Let $\mathcal{R} = \{r_i\}_{1 \leq i \leq N}$ be the set of records in table T, where $N = |T|$. $\mathcal{R}$ is a *totally ordered* set according to the 1D quasi-identifier Q_T. Our goal is to compute a partitioning of $\mathcal{R}$ that minimizes $\mathcal{IL}$ and satisfies the k-anonymity property.

An algorithm that computes the 1D optimal k-anonymous partitioning of $\mathcal{R}$ needs only to consider groups with records that are consecutive in the Q_T space. This results immediately from the fact that if two groups with at least k records each overlap, we can swap records between them such that the number of records in each group remains the same and the overlap is eliminated, without increasing $\mathcal{IL}$. Thus, the optimal k-anonymization solution that we propose holds for any superadditive information loss metric.

LEMMA 1. *Let $\mathcal{P}$ be the optimal k-anonymous partitioning of a set $\mathcal{R}$ according to $\mathcal{IL}$. Then $\mathcal{P}$ does not contain groups of more than $2k - 1$ records.*

PROOF. Assume that a group G in $\mathcal{P}$ contains more than $2k - 1$ records. We split G into two groups G_1 and G_2 of at least k records each, such that $G = G_1 \cup G_2$, $G_1 \cap G_2 = \emptyset$. Since $\mathcal{IL}$ is superadditive, $\mathcal{IL}(\mathcal{P}) \geq \mathcal{IL}((\mathcal{P} \setminus \{G\}) \cup \{G_1\} \cup \{G_2\})$; hence information loss cannot increase. Therefore the optimal partitioning does not need to contain groups of cardinality larger than $2k - 1$. □

For the sake of showing a specific metric calculation, we present our solution in the context of *GCP*. The privacy-constrained 1D k-anonymization problem can be solved with dynamic programming as follows: Let $Opt(i)$ be the information loss of the optimal partitioning achieved for the prefix subset of the first i records of $\mathcal{R}$; and $Opt_I([b, e]) = (e - b + 1) \cdot NCP(\{r_b, \ldots, r_e\})$ be the information

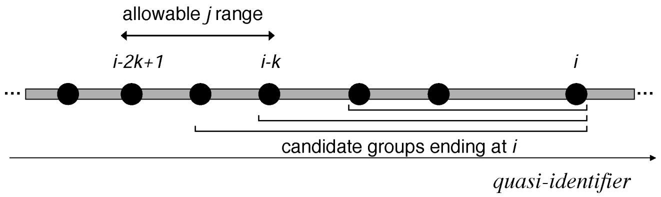


Fig. 6. Optimal 1D privacy-constrained k-anonymization example, $k = 3$.

```
Optimal 1D Privacy-Constrained k-anonymization
Input: set $\mathcal{R}$ in ascending order of 1-D $Q_T$
1.  for $i := k$ to $2k - 1$
2.    $Opt(i) = Opt_I([1, i])$
3.    $prev(i) = NIL$ /* used to reconstruct solution*/
4.  for $i := 2k$ to $N$
5.      for $j := \max\{k, i - 2k + 1\}$ to $i - k$
6.        $Opt(i) = \min_j \{Opt(j) + Opt_I([j + 1, i])\}$
7.        $prev(i) = j$ value that minimizes $Opt(i)$
8.  $i = N$ /* output $k$-anonymized groups */
9.  while $(prev(i) \neq NIL)$
10.   output group with boundaries $[prev(i) + 1, i]$
11.   $i = prev(i)$
12. output group $[1, i]$
```

Fig. 7. Optimal 1D privacy-constrained k-anonymization.

loss of the group containing all records in the interval $\{r_b, \dots, r_e\}$. Then

$$Opt(i) = \min_{i-2k<j\leq i-k} (Opt(j) + Opt_I([j + 1, i])).$$

This recursive scheme selects the best out of all suffixes of $\mathcal{R}$ to create the last group. Figure 6 shows an example of determining $Opt(i)$, with $k = 3$. Every group should contain between k and $2k - 1$ records, therefore there are three candidate groups ending at i, of cardinalities 3, 4, and 5, respectively. Furthermore, the last record j of the previous group must be in the interval $[i - 2k + 1, i - k]$, namely $[i - 5, i - 3]$.

Figure 7 shows the pseudocode of the algorithm. For the first group in $\mathcal{R}$, namely $k \leq i \leq 2k - 1$, the value of $Opt(i)$ is determined directly, and is equal to the NCP of the first i records times i. Then, the computation proceeds with increasing i, $2k \leq i \leq N$. The optimal solution for all j-prefixes of $\mathcal{R}$, where $j < i$, has been computed in advance. In addition to the best Opt value at record i, the algorithm needs to maintain the particular j value for which $Opt(i)$ was obtained, in order to reconstruct the solution after the tabulation is completed. The auxiliary table $prev$ serves this purpose (line 7). The algorithm generates an optimal partitioning $\mathcal{P}$, and the information loss of the partitioning is $GCP(\mathcal{P}) = Opt(N)/N$.

Complexity analysis. The algorithm ranges through $O(k)$ values of j for $O(N)$ values of i. Since $Opt_I([j + 1, i])$ can be computed in $O(1)$ (using the distance between the group's first and last records), the time complexity is $O(k \cdot N)$. The dynamic programming arrays Opt and $prev$ both have N entries; however, we only need to access a constant fraction $O(k)$ of the arrays at any time, yielding a constant space complexity $O(k)$. After the computation ends, we must scan the *prev* array (lines 9-11) one more time to output the solution. The overall I/O overhead is linear to N.

4. 1D PRIVACY-CONSTRAINED ℓ-DIVERSIfiCATION

In this section we study the privacy-constrained ℓ-diversification problem for 1D quasi-identifiers. In contrast to k-anonymity, optimal solutions within the ℓ-diversity model cannot be computed efficiently even in the 1D case. The inefficiency arises from the fact that the optimal partitioning may have to contain overlapping groups; therefore, numerous possible combinations must be examined. In this section, we study the properties of an optimal solution. Guided by these properties, we first develop a polynomial-time algorithm that finds the optimal solution. However, since the cost of the algorithm may be too high in practice, we also propose an efficient linear-time (in the size of the input) heuristic algorithm.

Our theoretical analysis holds for the $\mathcal{IL}$ information loss metrics in Eqs. (3) and (4). In Section 6, we discuss how our optimal solution can be extended as a heuristic for multidimensional quasi-identifiers with the *GCP* metric.

4.1 Properties of the Optimal Solution

Let $\mathcal{R} = \{r_i\}_{1 \le i \le N}$ be the set of records in the original table, and S the projection of $\mathcal{R}$ on the sensitive attribute (SA). Denote by $r_i.Q$ the 1D Q_T value of r_i and by $r_i.S$ the *SA* value of record r_i. Let $m = |S|$, that is, there are m distinct values of *SA*. For a pair of records r_i, r_j we denote $|r_i - r_j| = |r_i.Q - r_j.Q|$.

LEMMA 2. *Let $\mathcal{P}$ be an optimal ℓ-diverse partitioning of $\mathcal{R}$ according to the information loss metric $\mathcal{IL}$. Then $\mathcal{P}$ does not need to contain groups of more than $2\ell - 1$ records.*

PROOF. Assume there is a group G in the optimal solution such that $|G| \ge 2\ell$. Express the cardinality of G as $|G| = c \cdot \ell + r$, where c is an integer, $c \ge 2$, $0 \le r < \ell$. Since G is ℓ-diverse, according to Definition 4 every *SA* value in G can occur at most c times. There are at most ℓ values in G with c occurrences. We remove from G the ℓ records with the most frequent *SA* values in G, and create group G'. By construction, G' is ℓ-diverse. Let $G'' = G \setminus G'$. Any sensitive attribute value in G'' can occur at most $c-1$ times and $|G''| = (c-1)\ell + r$. Hence, G'' is ℓ-diverse. Furthermore, since $\mathcal{IL}$ is superadditive, $\mathcal{IL}(\mathcal{P}) \ge \mathcal{IL}((\mathcal{P} \setminus \{G\}) \cup \{G'\} \cup \{G''\})$. Splitting G'' recursively, we obtain a partitioning with equal or lower information loss compared to $\mathcal{P}$, and cardinality of each group between ℓ and $2\ell - 1$. □

COROLLARY 1 (VALUE SINGULARITY PROPERTY). *In an optimal ℓ-diverse partitioning $\mathcal{P}$, every group $G \in \mathcal{P}$ contains at most one occurrence for any SA value $s_j \in S$.*

PROOF. Assume an optimal solution $\mathcal{P}$ and $G \in \mathcal{P}$ such that s_j appears twice in G. Since $|G| \le 2\ell - 1$, it results that G is not ℓ-diverse, namely, a contradiction. □

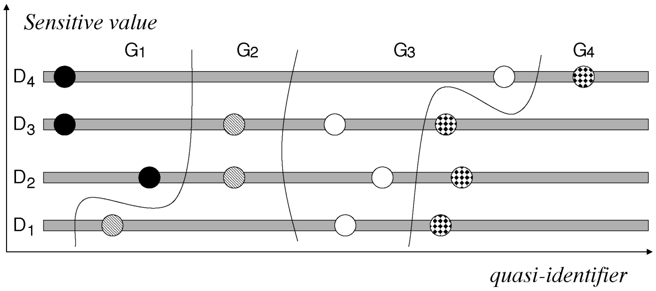


Fig. 8. Sensitive value domains.

Since there are only m distinct *SA* values, we conclude that $|G| \leq min(2\ell - 1, m)$.

$\mathcal{R}$ is a totally ordered set according to Q_T, and each record in $\mathcal{R}$ belongs to exactly one ℓ-diverse group. According to this order, we refer to the first and last record in group G_i as the begin (i.e., b_i) and, respectively, the end (i.e., e_i) record of G_i. We refer to b_i and e_i as border elements. In the optimal solution, there exists a total order of both begin and end records of the set of groups. However, unlike the case of k-anonymity, a group need not contain only consecutive records.

Let domain $\mathcal{D}_q = \{r_i \in \mathcal{R} | r_i.S = s_q\}, 1 \leq q \leq m$, that is, $\mathcal{D}_q$ contains all tuples whose *SA* value is s_q. Figure 8 depicts the domains $\mathcal{D}_q$ for a 3-diverse partitioning of $\mathcal{R}$, where $m = 4$. Note that the total order in the quasi-identifier space induces a total order for each of the domains $\mathcal{D}_q$.

The following lemma shows that the order of groups in each value domain $\mathcal{D}_q$ is the same.

LEMMA 3 (GROUP ORDER PROPERTY). *There exists an optimal ℓ-diverse partitioning $\mathcal{P}$ of $\mathcal{R}$, producing $|\mathcal{P}|$ groups $G_1, G_2, \ldots G_{|\mathcal{P}|}$, such that the order of sets $\{G_1^q, G_2^q, \ldots, G_{|\mathcal{P}|}^q\}$, defined for the groups in $\mathcal{P}$ as they appear along each domain $\mathcal{D}_q$, $G_i = \cup_q G_i^q$, $1 \leq i \leq |\mathcal{P}|$, is consistent across all domains $\mathcal{D}_q$, $1 \leq q \leq m$ (except for the fact that some groups may not be represented in each domain).*

PROOF. Assume an optimal solution in which there exist records $r_i \in G_i^q$ and $r_j \in G_j^q$ such that $r_i.Q < r_j.Q$, and records $t_i \in G_i^p$ and $t_j \in G_j^p$ such that $t_j.Q < t_i.Q$. Then, for all possible relative orderings in the 1D Q_T, $|r_i - t_j| + |r_j - t_i| \leq |r_i - t_i| + |r_j - t_j|$. Let $G'_i = G_i \setminus \{t_i\} \cup \{t_j\}$ and $G'_j = G_j \setminus \{t_j\} \cup \{t_i\}$. It results that $\mathcal{IL}_1(G'_i) + \mathcal{IL}_1(G'_j) \leq \mathcal{IL}_1(G_i) + \mathcal{IL}_1(G_j)$, that is, $\mathcal{IL}_1(\mathcal{P})$ cannot increase by exchanging t_j and t_i (a similar reasoning applies to the $\mathcal{IL}_\infty$ metric). Since t_i and t_j have the same *SA* value, the ℓ-diversity of the partitioning is not affected by the exchange. The same reasoning can be applied for all remaining pairs of records that violate a given order. It follows that the order of the partitions in the newly constructed optimal partitioning is consistent across all domains $\mathcal{D}_q$, $1 \leq q \leq m$. □

We write $G_i \prec G_j$ to denote that G_i precedes G_j in the partial order defined over optimal partitioning $\mathcal{P}$. As a consequence of Lemma 3, in order to find an optimal solution, we can build groups by assigning records from each domain in

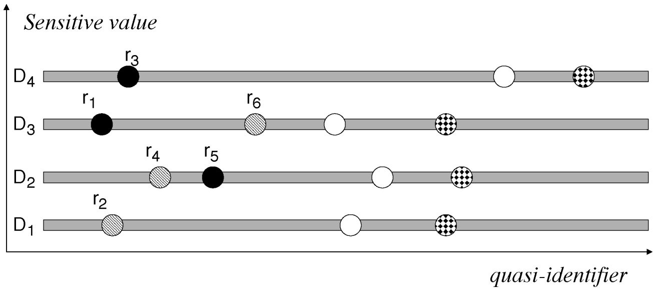


Fig. 9. Group order violation.

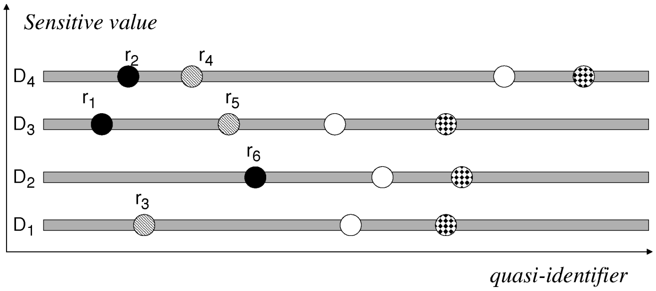


Fig. 10. Border order is violated, although group order is satisfied.

order. This prunes significantly the search space of the solution. Figure 9 shows an example where the group order property is violated. Let $G_1 = \{r_1, r_3, r_5\}$ and $G_2 = \{r_2, r_4, r_6\}$. G_1 precedes G_2 in the D_3 domain, while the opposite occurs for D_2. However, the optimal solution is $G'_1 = \{r_1, r_2, r_4\}$, $G'_2 = \{r_3, r_5, r_6\}$, and $G'_1 \prec G'_2$.

The following lemma states that the group ordering extends to the begin and end records of groups.

Lemma 4 (Border Order Property). *There exists an optimal ℓ-diverse partitioning $\mathcal{P}$ of $\mathcal{R}$, producing $|\mathcal{P}|$ groups $G_1, G_2, \ldots G_{|\mathcal{P}|}$ with begin records $b_1, b_2, \ldots, b_{|\mathcal{P}|}$ and end records $e_1, e_2, \ldots, e_{|\mathcal{P}|}$, such that the begin and end sets obey the same order as the groups $\{G_1, G_2, \ldots, G_{|\mathcal{P}|}\}$ they belong to, that is, if $G_i \prec G_j$, then $b_i.Q < b_j.Q$ and $e_i.Q < e_j.Q$.*

Proof. The proof is similar to that of Lemma 3. □

Lemma 4 further reduces the search space by limiting the choices of records for the currently built group based on the begin and end records of the previously built group. Figure 10 shows an example where the border order property is violated (although the group order property is satisfied). Let $G_1 = \{r_1, r_2, r_6\}$ and $G_2 = \{r_3, r_4, r_5\}$. b_1 (i.e., r_1) precedes b_2 (i.e., r_3), but e_1 (i.e., r_6) succeeds e_2 (i.e., r_5). The solution is not optimal; in the optimal case, $G'_1 = \{r_1, r_2, r_3\}$, $G'_2 = \{r_4, r_5, r_6\}$, $b'_1 < b'_2$, and $e'_1 < e'_2$.

Lemma 5 (Cover Property). *There exists an optimal ℓ-diverse partitioning $\mathcal{P}$ of $\mathcal{R}$ with the following property: $\forall G_i$, $G_j \in \mathcal{P}$ such that $G_i \prec G_j$, and $\nexists G_k$:*

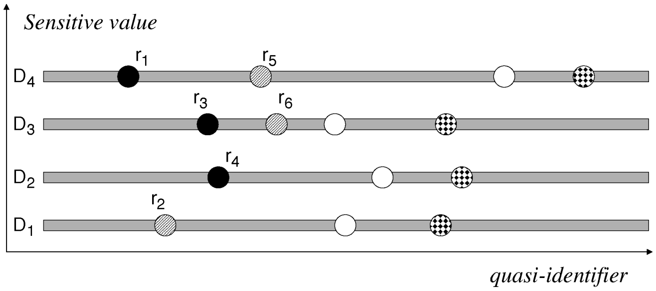


Fig. 11. Cover property violation.

$G_i \prec G_k \prec G_j$, if there exists a pair of records $r \in G_i, t \in G_j$, such that $r.Q > t.Q$, then there is either a record $r' \in G_j$ of the same sensitive value as r (where, according to Lemma 3, $r'.Q > r.Q$) or a record $t' \in G_i$ of the same sensitive value as t (where, according to Lemma 3, $t'.Q < t.Q$), or both.

PROOF. Assume there are records $r \in G_i$, $t \in G_j$ such that $r.Q > t.Q$, and there is neither $r' \in G_j$ with the same *SA* value as r, nor $t' \in G_i$ with the same *SA* value as t. Then we can swap r and t between G_i and G_j without compromising ℓ-diversity. Furthermore, since $b_i.Q \le b_j.Q \le t.Q \le r.Q \le e_i.Q \le e_j.Q$, it follows that the swap does not increase the information loss. Hence, we obtain an optimal solution where the condition specified in the lemma is satisfied. □

The intuition behind the cover property is that if record r can be added to any of two groups G_1 and G_2, then it should be added to the group that is closer to r in the Q_T space. Figure 11 shows an example where the cover property is violated: Consider partially completed groups $G_1 = \{r_1, r_3\}$ and $G_2 = \{r_5, r_6\}$. If r_2 is assigned to G_2 and r_4 to G_1, the cover property does not hold; in an optimal solution, r_4 must belong to G_2 and r_2 to G_1.

Definition 5 (*Group Boundaries*). The end boundary of $G_i = \cup_q G_i^q$ is the vector $\mathbf{e}_i = \{e_1^i, e_2^i, \ldots e_m^i\}$, where e_q^i is either the order of the record with largest Q_T of G_i in $\mathcal{D}_q$, if G_i includes a record with SA value s_q at all, or otherwise the order of the record with the largest Q_T value in $\mathcal{D}_q$ in any group $G_j \prec G_i$. We say that G_i ends at boundary $\mathbf{e}_i$. The begin boundary $\mathbf{b}_i = \{b_1^i, b_2^i, \ldots b_m^i\}$ is defined symmetrically. We call G_i the group between boundaries $\mathbf{b}_i$ and $\mathbf{e}_i$. A group's end record is the record with the largest Q_T value in the group end boundary, and is denoted by $\mathsf{enditem}(\mathbf{e}_i)$.

Intuitively, $\mathbf{e}_i$ marks the position of the last record of G_i in each domain D_q of a sensitive value s_q. If G_i does not contain a record with sensitive value s_q, then e_q is equal to the corresponding e_q in the previous group G_j ($G_j \prec G_i$). For instance, in Figure 11, let $G_1 = \{r_1, r_2, r_3\}$ and $G_2 = \{r_4, r_5, r_6\}$. The end boundary of G_2 is $\mathbf{e}_2 = \{1, 1, 2, 2\}$, since records r_4, r_6, and r_5 that belong to G_2 have positions 1, 2, and 2 in their respective domains D_2, D_3, and D_4. Furthermore, since $G_1 \prec G_2$, the value of $\mathbf{e}_2$ in D_1 is 1, namely, the position of r_2 from G_1. As an immediate result of Lemma 4, if two groups are ordered as $G_j \prec$

G_i, then the same order is enforced for their corresponding end boundaries. In other words, even if groups overlap in the Q_T space, their boundaries defined over the D_q domains do not overlap.

4.2 An Optimal 1D Algorithm

We introduce a dynamic programming algorithm that determines the optimal solution to the 1D privacy-constrained ℓ-diversification problem under the group extent-based information loss metrics we focus on. The following lemma gives the number of possibilities in choosing a group boundary, and provides upper bounds that are later used in the complexity analysis.

LEMMA 6. *Given an interval I that includes c records drawing their sensitive values from a set of m different values s_q, $1 \le q \le m$, the worst-case number of ways B_m^c in which we can construct a group boundary within I is*

$$B_m^c = \begin{cases} O(\lceil \frac{c}{m} \rceil^m) & , m \le c \\ O(2^c) & , m > c. \end{cases} \tag{7}$$

PROOF. Let I_q be the subset of records of value s_q in I and let $c_q = |I_q|$, $1 \le q \le m$. Then $I = \cup_q I_q$ and $c = \sum_{q=1}^m c_q$. We can choose exactly one record from I_q in c_q ways. In combination, we can choose at most one record in each value domain from I in $Prod = \prod_{q=1}^m (c_q + 1)$ ways (the $+1$ term represents the case where the boundary is placed before the first I record in domain D_q). If $m \le c$, then $Prod$ is maximized when $\forall q$, $c_q = \lceil \frac{c}{m} \rceil$ or $c_q = \lfloor \frac{c}{m} \rfloor$. In that case, $B_m^c = O(\lceil \frac{c}{m} \rceil^m)$. Otherwise, if $m > c$, $Prod$ is maximized when $\forall q$, $c_q = 1$, that is, each value s_q is represented by exactly one record. Then $B_m^c = O(2^c)$. □

Definition 6 (**a**-*Prefix*). Given a boundary **a**, the **a**-prefix of $\mathcal{R}$ is the subset of $\mathcal{R}$ that includes all records which precede or match the record of **a** in their value domain: $\{r \in \mathcal{R} | \exists q : r \in \mathcal{D}_q \text{ such that } r \le a_q\}$.

Given a boundary $\mathbf{a} = \{a_1, a_2, \ldots a_m\}$, let $\mathcal{IL}(\mathbf{a})$ be the information loss of the optimal ℓ-diverse partitioning achieved for the **a**-prefix of $\mathcal{R}$. Given two boundaries **a**, **b**, such that $\mathbf{b} \prec \mathbf{a}$, we define $\mathsf{IL}_I(\mathbf{b}, \mathbf{a})$ as the (immediate) information loss of the group between the boundaries **b** and **a**. Similarly, we use $\mathsf{PM}_I(\mathbf{b}, \mathbf{a})$ to denote the privacy of the group between the boundaries **b** and **a**, measured according to Eq. (6). Based on these definitions, the following recursive dynamic programming formulation determines the optimal ℓ-diverse partitioning of $\mathcal{R}$. We have

$$\mathcal{IL}(\mathbf{a}) = \min_{\mathbf{b} \prec \mathbf{a}, \mathsf{PM}_I(\mathbf{b}, \mathbf{a}) \ge \ell} \{\mathcal{F}\{\mathcal{IL}(\mathbf{b}), \mathsf{IL}_I(\mathbf{b}, \mathbf{a})\}\}, \tag{8}$$

where $\mathcal{F}$ can be either *sum* (if $\mathcal{IL}_1$ is used) or *max* (for $\mathcal{IL}_\infty$). This recursive scheme, based on Lemmata 3 and 4, selects the best out of all possible options of groups ending at each allowed end boundary **a**. We group allowed end boundaries **a** based on their end record $a_q = \mathsf{enditem}(\mathbf{a})$. There are $O(N)$ possible end records a_q, and, according to Lemma 6, for each $a_q = r_i$, there are B_m^i possible end boundaries **a**, since the rest of the records (besides a_q) in the boundary must be chosen from the predecessors of r_i in the 1D space. Besides, for each end boundary **a** corresponding to end record a_q, we should establish all allowed

```
Optimal 1D Privacy-Constrained $\ell$-diversification
Input: extent bound $E$, record set $\mathcal{R} = [r_1, \ldots, r_N]$
Output: privacy-optimal partitioning $\mathcal{P}$ of $\mathcal{R}$ under extent bound $E$
    /*$\mathcal{IL}[1, \ldots, N][1, \ldots, N^m]$ stores information loss for end-record/boundary pairs */
    /*$OptimalBoundary[1, \ldots, N][1, \ldots, N^m]$ stores tabulated boundaries */
    for $i = 1$ to $N$
        $\mathcal{A}$ = set of all possible boundaries ending at $r_i$
        for $a = 1$ to $|\mathcal{A}|$
            $\mathbf{a}$ = next boundary in $\mathcal{A}$
            $\mathcal{B}$ = set of all possible predecessor boundaries $\mathbf{b}$ of $\mathbf{a}$
                    s.t. at most one value in each domain lies between $\mathbf{b}$ and $\mathbf{a}$
            for $b = 1$ to $|\mathcal{B}|$
                $\mathbf{b}$ = next boundary in $\mathcal{B}$
                $j$ = index of rightmost record in $\mathbf{b}$
                if $(\mathsf{PM_I}(\mathbf{b}, \mathbf{a}) \geq \ell$ and $\mathcal{IL}[i][a] > \mathcal{F}\{\mathcal{IL}[j][b], \mathsf{IL_I}(\mathbf{b}, \mathbf{a})\})$
                    $\mathcal{IL}[i][a] = \mathcal{F}\{\mathcal{IL}[j][b], \mathsf{IL_I}(\mathbf{b}, \mathbf{a})\}$
                    $OptimalBoundary[i][a] = \mathbf{b}$
    $i = N$ /* output solution */
    choose $a$ s.t. $\mathcal{IL}[N][a]$ is the minimum in its column
    while $(i > 0)$
        $\mathbf{b} = OptimalBoundary[i][a]$
        output group between $\mathbf{b}$ (exclusive) and $\mathbf{a}$ (inclusive)
        $i$ = index of rightmost record in $\mathbf{b}$
        $\mathbf{a} = \mathbf{b}$
```

Fig. 12. Optimal 1D privacy-constrained ℓ-diversification pseudocode.

begin boundaries **b**, that is, all possible groups with a_q as end record which have at most one record in each value domain $\mathcal{D}_q$ (according to Corollary 1). Then the chosen **b**-boundary record in each domain $\mathcal{D}_q$ contains the predecessor of the last assigned record in its respective domain.

Figure 12 presents the pseudocode of the proposed algorithm.

Complexity analysis. The algorithm needs to maintain an entry for each record r_i, storing all allowed end boundaries **a** such that $\mathsf{enditem}(\mathsf{a}) = r_i$. Hence, its space complexity is $O(N^m)$. For each **a**, it has to iterate through all eligible begin boundaries (equivalently, end boundaries of an immediately preceding group) **b**. Due to the value singularity property (Corollary 1), once a boundary is set, we only have two choices for each of the m sensitive values (either to include the respective boundary record in the group or not), hence at most $O(2^m)$ choices. Each choice requires the computation of the information loss $\mathsf{IL}_I(\mathbf{b}, \mathbf{a})$, incurring cost $O(1)$ for a 1D quasi-identifier, since the group's extent is straightforwardly computed from its first and last records. The worst-case time complexity is $O(2^m N^m)$. Although the algorithm is polynomial in the input size, it can be prohibitively costly in practice. Next, we discuss an efficient heuristic.

4.3 An Efficient 1D Heuristic for ℓ-Diversification

We present an heuristic 1D ℓ-diversification algorithm. Our heuristic is inspired from the theoretical analysis of Section 4.1, but its applicability is *not* limited to the group extent-based metrics $\mathcal{IL}$. As we will show in Section 7, our heuristic yields good results with a variety of information loss metrics.

Given a sorted input, the algorithm exhibits time and I/O cost linear in the input size. The heuristic guarantees that if the original table satisfies the eligibility condition (EG, see Section 2.3) for a given ℓ value, then a solution will be found, although it may not be an optimal one.

First, the records are sorted according to their Q_T value and are assigned to m domains $D_{1 \leq q \leq m}$, based on their sensitive attribute value. Subsequently, following the results from Corollary 1 and Lemmata 3 and 4, the group formation phase attempts to form groups having between ℓ and m records with distinct *SA* values. Let $\mathbf{e} = \{e_1, e_2, \ldots e_m\}$ be the end boundary of the previously formed group. We denote by frontier of the search the set $\{r_q \in D_q | 1 \leq q \leq m\}$, such that each r_q is the successor of e_q in its respective domain. Initially, the frontier consists of the first record in each domain D_q.

The heuristic consists of two steps: the greedy step and the fall-back step. In the greedy step, it assigns to the current group G those ℓ records on the frontier with the lowest Q_T values, and checks the eligibility condition EG for the remaining records. If EG is satisfied, then G is closed, the frontier is advanced beyond the records in G, and the algorithm starts building the next group. Otherwise, out of the remaining unassigned records on the frontier, the record with the lowest Q_T is added to G, and EG is checked again. The process continues until EG is satisfied, or all m records on the frontier are in G.

If EG is still not satisfied, the records in G are rolled back, and the following fall-back strategy step is executed: ℓ of the records on the frontier with *SA* values which are the most frequent among the unassigned records are added to G (in case of ties, the record with the lowest Q_T is chosen). If EG is not satisfied, the record with the $(\ell + 1)^{th}$ most frequent value is added, and so forth, up to $m - 1$ (the case where all m records on the frontier are chosen has been considered in the greedy step). It is guaranteed that by picking the most frequent records, EG is eventually satisfied [Xiao and Tao 2006a]; therefore, a solution can be found. We emphasize that the need to execute the fall-back step for the current group does not imply that it will be necessary for the next one. The fall-back step may be necessary for Q_T regions with significant variance in density of records among distinct *SA* domains.

Figure 13 shows the pseudocode of the heuristic algorithm. To evaluate EG, we maintain a counter remaining with the number of unassigned records, and a histogram H with the distribution of the *SA* values of the remaining records. Upon each record assignment, remaining and H are updated. H contains m elements; hence the cost of updating H and evaluating EG is $O(m)$ (the cost can be reduced to $O(\log m)$ using a priority queue for histogram H).

The presented heuristic will finalize the current group G if it is able to find $count \leq m$ records such that EG holds. However, in some cases, this approach may generate groups with large extent. Consider the example in Figure 14, where $\ell = 3$. After picking the first three records, the algorithm closes G at boundary 1, and r_4 is grouped with r_{5-7}. However, if it were grouped with r_{1-3} (boundary 2), the extent of the partitioning (hence, the resulting information loss) would be considerably smaller.

To minimize this effect, we implement the following optimization: After G is formed (e.g., $\{r_1, r_2, r_3\}$ in Figure 14), we inspect records r_A and r_B on the

```
Heuristic 1D Privacy-Constrained ℓ-diversification
Input: set $\mathcal{R} = \{r_i\}_{1 \le i \le N}$ in ascending order of 1D $Q_T$
split sorted records in $m$ buckets based on $SA$ value
$H[i]$ = number of records in bucket $i$
$remaining = N$
frontier $\mathcal{F}$ = {set of first record in each bucket}
while ($remaining > 0$)
    $count = \ell$
    do /*greedy step*/
        $G$ = {set of $count$ records of $\mathcal{F}$ with lowest $Q_T$}
        $count{+}{+}$
    until ($EG$ holds or $count > m$)
    if ($EG$ does not hold) /*fall-back step*/
        $count = \ell$
        do
            $G$ = {set of $count$ records in $\mathcal{F}$ with max $H$ value}
            $count{+}{+}$
        until ($EG$ holds)
    close $G$, update $H$ and advance $\mathcal{F}$
    $remaining = remaining - count + 1$
output $\ell$-diverse groups
```

Fig. 13. Heuristic 1D privacy-constrained ℓ-diversification.

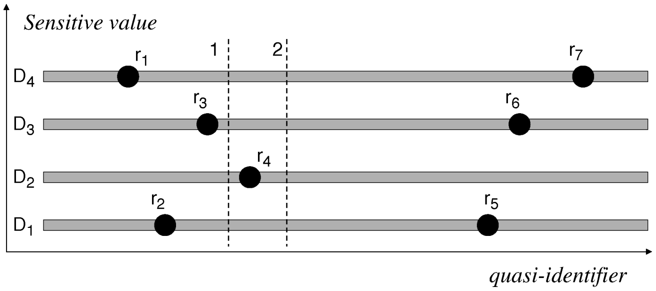


Fig. 14. Heuristic optimization, $\ell = 3$.

frontier with the 1^{st} and, respectively, ℓ^{th} lowest Q_T value (i.e., $r_A \equiv r_4, r_B \equiv r_6$ in the example). The extent of the group that contains $r_A, .., r_B$ is a lower bound for the extent of the group that will contain r_A. If the distance from r_A to the leftmost record in G (e.g., $|r_4 - r_1|$) is smaller than the distance from r_A to r_B (e.g., $|r_4 - r_6|$), and there is not already a record with $r_A.S$ in G (e.g., no record from D_2 in G), we add r_A to G, subject to EG being satisfied for the set of remaining records. In the running example, the two obtained groups are $\{r_1, r_2, r_3, r_4\}$ and $\{r_5, r_6, r_7\}$. This optimization aims to reduce the information loss of ℓ-diverse groups, and has complexity $O(m)$. The overall cost of the heuristic is $O(m \cdot N)$.

5. 1D ACCURACY-CONSTRAINED PROBLEMS (DUAL PROBLEMS)

In this section, we study the accuracy-constrained problem (Problem 2.5), focusing on the more difficult scenario of accuracy-constrained ℓ-diversification. The dual k-anonymization problem can be easily solved by collapsing the multidomain representation discussed in Section 4, and placing a constraint on group extent only. The optimal 1D solution for dual k-anonymization is similar to the DP formulation in Section 3, and has complexity $O(N^2)$, where N is the dataset size. In the following, we present the solution for accuracy-constrained ℓ-diversification with 1D quasi-identifiers.

We preserve the multidomain data representation, the notations, and terminology introduced in Section 4.1. As stated in Section 2.5, the information loss bound E is expressed as a maximum-loss-per-group function. In our theoretical analysis, we consider the information loss metric $\mathcal{IL}_\infty$ of Eq. (4). The extension to the *NCP* metric is straightforward, since *NCP* is the normalized version of $\mathcal{IL}_\infty$. We say that a partitioning $\mathcal{P}$ is *E-bounded* iff $\mathcal{IL}_\infty(\mathcal{P}) \leq E$. Privacy is measured according to the $\mathcal{PM}$ metric (Eq. (6)).

Some of the general properties of the optimal solution for the direct problem also hold for the dual problem. The group order property (Lemma 3) and the border order property (Lemma 4) apply without change. However, the value singularity property (Corollary 1) does not hold anymore, because the solution does not rely on a fixed (integer) privacy bound. Instead, given an accuracy bound, we aim at maximizing privacy. Increasing the number of records in a group can enhance its privacy, according to the monotonicity property of ℓ-diversity [Machanavajjhala et al. 2006], which states that given groups G_1 and G_2 with respective privacy metrics $\mathcal{PM}(G_1)$ and $\mathcal{PM}(G_2)$, then $\mathcal{PM}(G_1 \cup G_2) \geq \min\{\mathcal{PM}(G_1), \mathcal{PM}(G_2)\}$. In other words, by merging two groups of records, the resulting privacy can only increase.

The following lemma shows that although groups in the optimal solution may contain multiple records with the same *SA* value, these records are consecutive within each *SA* domain.

LEMMA 7 (CONSECUTIVITY PROPERTY). *Any E-bounded partitioning $\mathcal{P}$ of $\mathcal{R}$ can be substituted by another E-bounded partitioning $\mathcal{P}'$ such that $\mathcal{PM}(\mathcal{P}) \leq \mathcal{PM}(\mathcal{P}')$ and each projection $\mathcal{P}'_q$ of $\mathcal{P}'$ onto a domain $\mathcal{D}_q$, $1 \leq q \leq m$, contains only groups of records that are* consecutive *in $\mathcal{D}_q$. Formally, $\forall G \in \mathcal{P}'$, if group G contains records $r_i \in \mathcal{D}_q$ and $r_k \in \mathcal{D}_q$, $r_i < r_k$, then it also contains every record $r_j \in \mathcal{D}_q$ such that $r_i < r_j < r_k$.*

PROOF. Assume a partitioning $\mathcal{P}$ such that a group $G \in \mathcal{P}$, with begin record b and end record e, contains records $r_i, r_k \in \mathcal{D}_q$, but there exists a record $r_j \in \mathcal{D}_q$ such that $r_i < r_j < r_k$ and $r_j \notin G$. Assume that $r_j \in G'$, where $G' \in \mathcal{P}$, with begin record b' and end record e'. Since r_i, r_j, r_k share the same sensitive value, the $\mathcal{PM}$ of the groups that contain them is not modified by exchanging a pair of these records between these groups. In all cases, there exists such an exchange that enforces the consecutivity of G and maintains the privacy of the partitioning. The same reasoning applies to all nonconsecutive group records and all value domains. By induction, it follows that there exists a partitioning $\mathcal{P}'$ of no worse privacy than $\mathcal{P}$, whose projection $\mathcal{P}'_q$ onto each domain $\mathcal{D}_q$ contains only groups of records which are *consecutive* in $\mathcal{D}_q$. □

The practical consequence of the consecutivity property is that an algorithm striving to establish an optimal partitioning for the dual problem needs to

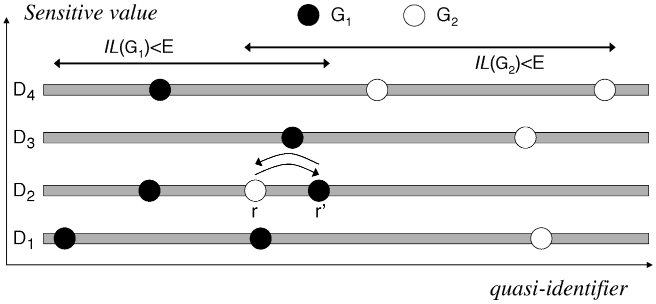


Fig. 15. Consecutivity property violation.

include in the same group records with the same *SA* value s_q which are consecutive in their domain $\mathcal{D}_q$. In other words, there need not be overlaps between partitions within each sensitive value domain $\mathcal{D}_q$. It follows that an ordered set $\{G_1^q, G_2^q, \ldots, G_{|\mathcal{P}|}^q\}$ is defined by the groups formed by a partitioning $\mathcal{P}$ of $\mathcal{R}$ for each domain $\mathcal{D}_q$, where $G_i = \cup_q G_i^q$, $1 \le i \le |\mathcal{P}|$.

Figure 15 shows an example with two E-bounded groups, G_1 and G_2, and records $r \in G_1$ and $r' \in G_2$ which violate the consecutivity property. By performing an exchange of the two records, the extent of neither G_1 nor G_2 can be enlarged, whereas the privacy metric of the two groups is unchanged. Hence, an E-bounded partitioning of no worse privacy than the initial one is obtained, such that the consecutivity property is satisfied.

5.1 An Optimal 1D Algorithm

We propose a dynamic programming algorithm that finds the optimal solution for the 1D dual problem. The following lemma bounds the number of possible group formations in an interval of size E.

LEMMA 8. *Given an interval I of extent E which includes c records drawing their sensitive values from a set of m different values s_q, $1 \le q \le m$, the worst-case number of ways C_m^c to form a group with records from I is*

$$C_m^c = \begin{cases} O\left(\left(\frac{\lceil \frac{c}{m} \rceil^2}{2}\right)^m\right) & , m \le c \\ O(2^c) & , m > c. \end{cases}$$

PROOF. Let I_q be the subset of records of $\mathcal{D}_q$ in I and let $c_q = |I_q|$, $1 \le q \le m$. Then $I = \cup_q I_q$ and $c = \sum_{q=1}^m c_q$. We can choose i consecutive records from I_q in $j = c_q + 1 - i$ ways, $0 \le i \le c_q$, where the case $i = 0$ stands for the choice of a boundary record for a formed group which does not include records in $\mathcal{D}_q$. In total, we can choose from 0 to c_q consecutive records (note that choosing 0 records implies choosing just a boundary position; such choices also count) from I_q in $\sum_{j=1}^{c_q+1} j = \frac{(c_q+1)(c_q+2)}{2}$ ways. In combination, we can form a group with records from I in $C_m^c = \prod_{q=1}^m \frac{(c_q+1)(c_q+2)}{2}$ ways. If $m \le c$, then C_m^c is maximized when $\forall q, c_q = \lceil \frac{c}{m} \rceil$, or $c_q = \lfloor \frac{c}{m} \rfloor$. In that case, $C_m^c = O((\lceil \frac{c}{m} \rceil^2/2)^m)$. Otherwise, if $m > c$, then C_m^c is maximized when $\forall q, c_q = 1$, that is, when each represented value s_q in I is represented by exactly one record. For each record there are two options: to include the record in the group, or to put the group boundary before it (still counted as a boundary within interval I). Hence $C_m^c = O(2^c)$. □

Let c be the maximum number of consecutive records of $\mathcal{R}$ found in an interval of extent E. Then the optimal algorithm needs to examine groups of at most c records in the worst case. Given a boundary $\mathbf{a} = \{a_1, a_2, \dots a_m\}$, let $\mathcal{PM}(\mathbf{a})$ be the optimal (i.e., maximum) privacy achieved by a partitioning of the $\mathbf{a}$-prefix of $\mathcal{R}$, allowing groups of maximum extent E. We use the notation $\mathsf{PM}_I(\mathbf{b}, \mathbf{a})$ to denote the privacy of the group between the boundaries $\mathbf{b}$ and $\mathbf{a}$. Similarly, we denote by $\mathsf{E}(\mathbf{b}, \mathbf{a})$ the extent of the group between $\mathbf{b}$ and $\mathbf{a}$. The following recursive dynamic programming scheme computes the optimal value of $\mathcal{PM}(\mathbf{a})$.

$$\mathcal{PM}(\mathbf{a}) = \max_{\mathbf{b} \prec \mathbf{a} | \mathsf{E}(\mathbf{b},\mathbf{a}) \le E,} \{\min\{\mathcal{PM}(\mathbf{b}), \mathsf{PM}_I(\mathbf{b}, \mathbf{a})\}\} \tag{9}$$

This recursive scheme is based on the group order property (Lemma 3). It selects the best out of all possible groups ending at each allowed end boundary $\mathbf{a}$; that group has to be delimited by a start boundary $\mathbf{b}$. For a given $\mathbf{a}$, the eligible choices of $\mathbf{b}$ are those for which the group between $\mathbf{b}$ and $\mathbf{a}$ satisfies the bound E. The appropriate optimal solution for the $\mathbf{b}$-prefix of $\mathcal{R}$, $\mathcal{PM}(\mathbf{b})$, must be calculated in advance. In our implementation, we index allowed end boundaries $\mathbf{a}$ based on their end record a_q. Each out of N records in $\mathcal{R}$ can be an end record a_q. Thereafter, each end record $a_q = \mathsf{enditem}(\mathbf{a})$ (recall Definition 5) defines an E-extent interval $I_{a_q}^E$ ending at a_q. According to Lemma 6, for each a_q, there are $O(B_m^c)$ possible end boundaries $\mathbf{a}$ within $I_{a_q}^E$, where $c = |I_{a_q}^E|$. Then, for each end boundary $\mathbf{a}$ corresponding to end record a_q, we should establish all allowed start boundaries $\mathbf{b}$ within $I_{a_q}^E$. In other words, for each end record a_q, we should establish all possible valid groups within $I_{a_q}^E$ having a_q as end record. According to Lemma 8, there are C_m^c such possible groups. The final result is $\mathcal{PM}(\mathbf{e})$, where $\mathbf{e}$ is the end boundary such that $\mathbf{e}$-prefix $= \mathcal{R}$.

Figure 16 presents the pseudocode of the proposed algorithm.

Complexity analysis. The algorithm maintains a table which stores at entry i the $\mathcal{PM}(\mathbf{a})$ value for every allowed boundary $\mathbf{a}$ such that $\mathsf{enditem}(\mathbf{a}) = r_i$; hence the space complexity is $O(B_m^c \cdot N)$, $c = \max_i |I_{r_i}^E|$. For each $\mathbf{a}$, it ranges through all eligible start boundaries $\mathbf{b}$, for which the group between $\mathbf{b}$ and $\mathbf{a}$ satisfies the bound E. That makes a total of $O(C_m^c)$ choices for group formation between $\mathbf{b}$ and $\mathbf{a}$ for each r_i. Besides, each choice of $\mathbf{b}$ (i.e., group formation) requires the computation of $\mathsf{PM}_I(\mathbf{b}, \mathbf{a})$ in $O(c)$ time, which is absorbed by the C_m^c factor. Hence the total worst-case time complexity is $O(C_m^c \cdot N)$.

5.2 Using the Dual Problem to Solve Privacy-Constrained ℓ-Diversification

As mentioned in Section 4.1, the 1D *direct* optimal algorithm for the privacy-constrained problem has $O(2^m N^m)$ complexity. However, the solution to the dual problem can be used to solve the direct problem more efficiently, because in practice $c \ll N$. Given the privacy constraint ℓ, we choose an initial arbitrary upper bound E_{init}, for instance, half of the maximum 1D extent in the quasi-identifier attribute. Then, we use E_{init} as the seed for a binary search procedure

```
Optimal 1D Accuracy-Constrained $\ell$-diversification
Input: extent bound $E$, record set $\mathcal{R} = [r_1, \ldots, r_N]$
Output: privacy-optimal partitioning $\mathcal{P}$ of $\mathcal{R}$ under extent bound $E$
    /*$\mathcal{PM}[1, \ldots, N][1, \ldots, B_m^c]$ stores privacy values for end-record/boundary pairs */
    /*$OptimalBoundary[1, \ldots, N][1, \ldots, B_m^c]$ stores tabulated boundaries */
for $i = 1$ to $N$
    $\mathcal{A}$ = set of boundaries ending at $r_i$ in $I_{r_i}^E$
    for $a = 1$ to $|\mathcal{A}|$
        $\mathbf{a}$ = next boundary in $\mathcal{A}$
        $\mathcal{B}$ = set of predecessor boundaries to $\mathbf{a}$ in $I_{r_i}^E$
        for $b = 1$ to $|\mathcal{B}|$
            $\mathbf{b}$ = next boundary in $\mathcal{B}$
            $j$ = index of rightmost record in $\mathbf{b}$
            if $(\mathcal{PM}[i][a] < \min\{\mathcal{PM}[j][b], \mathsf{PM_l}(\mathbf{b}, \mathbf{a})\})$
                $\mathcal{PM}[i][a] = \min\{\mathcal{PM}[j][b], \mathsf{PM_l}(\mathbf{b}, \mathbf{a})\}$
                $OptimalBoundary[i][a] = \mathbf{b}$
$i = N$ /* output solution */
choose $a$ s.t. $\mathcal{PM}[N][a]$ is the minimum in its column
while $(i > 0)$
    $\mathbf{b} = OptimalBoundary[i][a]$
    output group between $\mathbf{b}$ (exclusive) and $\mathbf{a}$ (inclusive)
    $i$ = index of rightmost record in $\mathbf{b}$
    $\mathbf{a} = \mathbf{b}$
```

Fig. 16. Optimal 1D accuracy-constrained ℓ-diversification pseudocode.

that solves the dual problem, with changing bound E. The search stops when two values E_0, E_0' are found, such that the privacy obtained with E_0 is at least ℓ, the privacy obtained with E_0' is less than ℓ, and $|E_0 - E_0'| < \epsilon$, an arbitrary-small constant.

Complexity analysis. The time complexity of the binary search procedure is $O(C_m^{\bar{c}} \cdot N \cdot \log E_0)$, where E_0 is the final, optimal extent bound and $\bar{c}$ the maximum value of c encountered during the search. Accordingly, the space complexity is $O(B_m^{\bar{c}} \cdot N)$.

5.3 An Efficient 1D Heuristic for the Dual Problem

Similarly to the direct problem, the optimal 1D algorithm for the dual one can incur a high overhead. Although it is polynomial to the input size N, the exponential factor m (cardinality of *SA* domain) can lead to prohibitive cost. For this reason, we propose an efficient heuristic (linear in N) which is guided by the properties of the optimal solution, but reduces considerably the search space.

As dictated by Lemma 7, only records that are consecutive in the multi-domain representation need to be considered for inclusion in one group. In addition, our heuristic uses the monotonicity property of ℓ-diversity [Machanavajjhala et al. 2006] mentioned earlier, and for the given accuracy bound E, attempts to increase the cardinality of each group, therefore improving its privacy metric.

Based on these two guidelines, our heuristic traverses the record set $\mathcal{R}$ (ordered by the 1D quasi-identifier) in a greedy fashion, and adds records to the

```
Heuristic 1D Accuracy-Constrained $\ell$-diversification
Input: set $\mathcal{R} = \{r_i\}_{1 \le i \le N}$ in ascending order of 1D $Q_T$
1.  $G_{prev} = \emptyset$, $i = 1$
2.  while $(\mathcal{IL}_\infty(G_{prev} \cup \{r_i\}) \le E)$ /*build first group*/
3.    $G_{prev} = G_{prev} \cup \{r_i\}$
4.    $i++$
5.  while $(i \le N)$ /*while unassigned records exist*/
6.    $G_{crt} = \{r_i\}$, $i++$
7.    while $(\mathcal{IL}_\infty(G_{crt} \cup \{r_i\}) \le E)$
8.      $G_{crt} = G_{crt} \cup \{r_i\}$
9.      $i++$
10.   $adjust\_privacy(G_{prev}, G_{crt})$
11.   output $G_{prev}$
12.   $G_{prev} = G_{crt}$
13. output $G_{prev}$

$adjust\_privacy(G_1, G_2)$
14. do
15.   success = $move\_record(G_1, G_2)$ or $move\_record(G_2, G_1)$
16. while(success)

$move\_record(src, dest)$
17. for $i = 1$ to $m$ do
18.   $r$ =record with $i^{th}$ most frequent SA value in $src$ which is closest to $dest$
19.   $src' = src \setminus \{r\}$, $dest' = dest \cup \{r\}$
20.   if $(\min\{\mathcal{PM}(src), \mathcal{PM}(dest)\} < \min\{\mathcal{PM}(src'), \mathcal{PM}(dest')\}$
                          and $\mathcal{IL}_\infty(dest') \le E)$
21.     $src = src'$, $dest = dest'$
22.     update $\mathcal{IL}_\infty(src)$ and $\mathcal{IL}_\infty(dest)$
23.     return $true$
24. return $false$
```

Fig. 17. Heuristic 1D accuracy-constrained ℓ-diversification.

current group G_{crt} as long as $\mathcal{IL}_\infty(G_{crt})$ does not exceed E. Once G_{crt} can no longer be enlarged, it is stored as candidate group G_{prev} (but not output yet), and a new group G_{crt} is built with the records that follow G_{prev}. When G_{crt} can no longer be enlarged, a privacy adjustment phase is performed between G_{prev} and G_{crt}, with the purpose of maximizing

$$\min\{\mathcal{PM}(G_{prev}), \mathcal{PM}(G_{crt})\}.$$

After readjustment, G_{prev} is output, G_{crt} becomes G_{prev}, and the process continues with the remaining records.

Figure 17 shows the pseudocode for the 1D dual ℓ-diversification heuristic. The main routine (lines 1–13) assembles candidate groups and invokes *adjust_privacy* for each pair of consecutive groups. The *adjust_privacy* routine consists of redistribution of records among the input groups G_1 and G_2. Intuitively, G_1 attempts to move some of its records to G_2, and vice versa. For each group, the heuristic attempts to transfer away the records with the most frequent *SA* value first, since removing them has the most significant impact

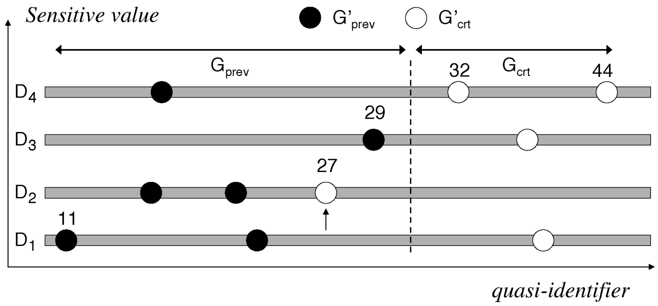


Fig. 18. Accuracy-constrained heuristic example.

in increasing the group $\mathcal{PM}$. The readjustment phase continues until no such transfer can be done. The *move_record(src,dest)* routine performs the actual record transfer, provided that the minimum privacy is not decreased, and the resulting $\mathcal{IL}_\infty(dest)$ does not exceed the accuracy bound (line 20). If such a transfer is allowed for a certain *SA* value *s*, out of all *s*-valued records in *src*, the one that causes the smallest enlargement to *dest* is chosen (this can be either the first or last *s*-valued record in *src*, depending on the relative 1D order of groups *src* and *dest*). In order to avoid situations when records with a particular *SA* value are swapped back-and-forth between the two groups, we allow the transfer of any *SA* value to occur in one direction only, during the re-adjustment.[8] This restriction does not affect the process of increasing the minimum privacy among groups, since if a *SA* value *s* is very frequent in G_1, but less frequent in G_2, then a transfer of an *s*-valued record from G_2 to G_1 will be invalidated by the test in line 20 (should the transfer be attempted from G_2 to G_1 first).

Figure 18 shows an example; the number on top of a record signifies the 1D quasi-identifier value. Assume $E = 20$. Group G_{prev} contains all records to the left of the dotted line, which signifies the limit (value $31 = 11+20$) beyond which G_{prev} cannot be extended. All remaining records belong to G_{crt}. The privacy of the two groups is $\mathcal{PM}(G_{prev}) = 7/3$ (because there are at most three records of the same *SA* value in G_{prev}, and $|G_{prev}| = 7$) and $\mathcal{PM}(G_{crt}) = 4/2 = 2$. During the readjustment phase, the record with 1D value 27, which has the most frequent *SA* value in G_{prev}, is moved to G_{crt}. The transfer is allowed, since the resulting extent of G_{crt} is $44 - 27 < 20$, resulting in new groups G'_{prev} and G'_{crt} (highlighted with distinct colors), such that $\mathcal{PM}(G'_{prev}) = 6/2 = 3$ and $\mathcal{PM}(G'_{crt}) = 5/2$. The overall privacy is thus increased from 2 to 2.5.

The *move_record* routine, which performs the bulk of the work, does not need to access all records in the group to check the eligibility of a transfer. In each group, records are partitioned into buckets, one for each *SA* value; within each bucket, records are ordered according to the 1D quasi-identifier. A record count is maintained for each bucket, resulting into a histogram over all *SA* values. With this data structure, determining the resulting privacy of a

[8] For brevity, this detail is not included in the pseudocode, but it can be efficiently implemented by keeping a bitmap with one entry for each *SA* value, and setting the corresponding bit every time a transfer is performed.

potential transfer costs[9] $O(m)$. Furthermore, determining the particular record to move (i.e., the closest one to *dest*) costs $O(1)$ (it is either the first or last in its bucket, depending on whether *dest* precedes or follows *src* in the 1D order).

Each record belongs to exactly one group and, for each pair of consecutive candidate groups, a record can be moved at most once (due to the unidirectional transfer constraint). Note that the transfer of a record cannot be propagated over two or more groups. For instance, if a record is initially in G_1 and it is transferred to G_2 in the privacy adjustment phase, it cannot be subsequently moved to G_3 when adjustment is performed among groups G_2 and G_3, because the E bound for G_3 would be exceeded. This is a consequence of the greedy group formation process, which keeps adding records to the current candidate group as long as bound E is satisfied. Checking whether a transfer is allowed (line 20) takes $O(1)$; updating the $\mathcal{IL}_\infty$ of the *src* and *dest* groups (line 22) also costs $O(1)$ (due to the 1D quasi-identifier). Therefore, the overall cost of the heuristic is $O(m \cdot N)$.

6. GENERAL MULTIDIMENSIONAL CASE

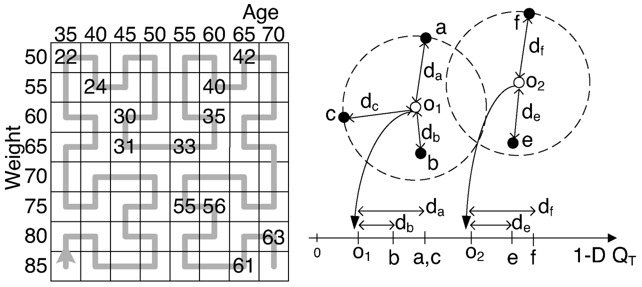
In this section we extend our 1D k-anonymization and ℓ-diversification algorithms to multidimensional quasi-identifiers. Let Q_T be a quasi-identifier with d attributes (i.e., d dimensions). We map the d-dimensional Q_T to one dimension and execute our 1D algorithms on the transformed data, while adapting them to compute information loss in the multidimensional space. Recall that both optimal k-anonymization and ℓ-diversification are NP-hard [Meyerson and Williams 2004; Machanavajjhala et al. 2006] in the multidimensional case. The solutions we obtain through mapping are not optimal; however, due to the good locality properties of the space mapping techniques, information loss is low, as we demonstrate experimentally in Section 7. In the following, we measure the information loss of each k-anonymous or ℓ-diverse group using *NCP*, and the information loss over the entire partitioning using *GCP* (see Section 2).

We employ two well-known space mapping techniques: the Hilbert space filling curve and iDistance [Zhang et al. 2005]. The Hilbert curve is a continuous fractal which maps each region of the space to an integer. With high probability, if two points are close in the multidimensional space, they will also be close in the Hilbert transformation [Moon et al. 2001]. Figure 19(a), for instance, shows the transformation from 2D to 1D for the 8 × 8 grid of the example in Section 1; the granularity of the regions can be arbitrarily small. The dataset is totally ordered with respect to the 1D Hilbert value.

iDistance is optimized for nearest-neighbor queries. In iDistance, a random sample of the data is first clustered around a fixed number of center points. The cluster centers are ordered according to any method (e.g., Hilbert ordering). Each data point is then assigned to its closest cluster center according to Euclidean distance. The 1D value of a point p is the sum of the 1D value of its cluster center C, plus the distance from p to C (see Figure 19(b)).

Regardless of the technique, in order to perform the data mapping, each attribute value must be assigned to a number. For numerical attributes, we can

[9] As mentioned in Section 4.3, this bound can be improved to $O(\log m)$ by using a priority queue.



(a) Hilbert curve (b) iDistance

Fig. 19. Multidimensional to 1D mappings.

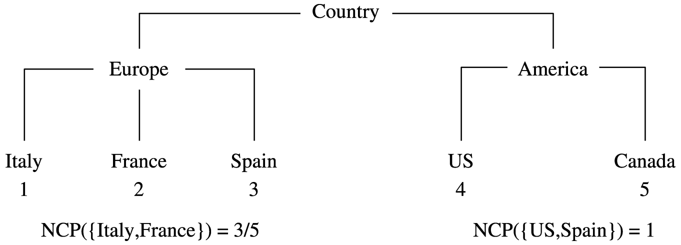


Fig. 20. Categorical attribute mapping.

use the attribute value directly; furthermore, the semantic distance between two numeric attribute values can be measured as the difference between the two values. For categorical attributes and their associated taxonomy tree, we adopt the labeling approach of Bayardo and Agrawal [2005] and LeFevre et al. [2005], where each attribute value is assigned to a distinct integer according to the in-order traversal of the taxonomy tree. If an equivalence class spans across different subtrees, it is penalized according to *NCP*. Figure 20 shows an example, where *NCP*({Italy, France}) $= 3/5$ because their common ancestor is Europe (which has 3 leaves) and there are 5 leaves in the entire Country domain. Also, NCP({US, Spain}) $= 1$ (i.e., maximum information loss), because their common ancestor is the entire Country domain. The mapping is performed only with respect to Q_T; the sensitive attribute is not included in the mapping.

The overhead of the Hilbert mapping is $O(d)$ per record, hence the method is efficient. For iDistance, the mapping involves the additional overhead of finding the cl cluster centers for a random sample of the data. After selecting the centers, the overhead of mapping is $O(cl)$ per record. Our 1D k-anonymization and ℓ-diversification algorithms require the input to be sorted according to 1D value; the cost is $O(N \log N)$. Assuming a sorted input, our methods need to scan the data only once; therefore the I/O cost is linear. Next, we discuss some further issues about the extension of our 1D algorithms to d dimensions.

6.1 Privacy-Constrained k-Anonymization

The k-anonymization dynamic programming algorithm builds two tables: (i) the main table with N entries, which stores at entry i the cost of the optimal solution for the first i records, and (ii) the auxiliary table that stores the base-case cost (i.e., *NCP*) for each sequence of consecutive k to $2k - 1$ records. Since the tabulation proceeds from left to right, at each step we need to look back at most $2k - 1$ entries; therefore, we do not need more than a constant fraction of the tables in main memory. If the tables do not fit in main memory, we need to store and then read them from the disk once; the I/O cost is $O(N)$.

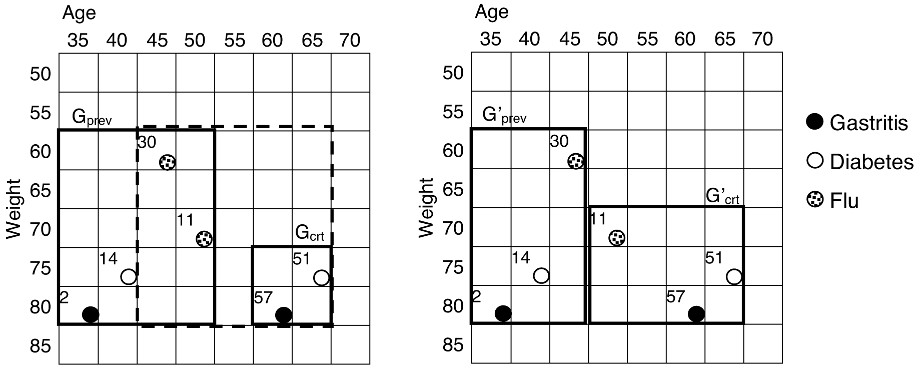
The time required to compute the *NCP* for a sequence of records is linear to the sequence length. Since the sequences are in the form $[r_{i-2k+2}, r_i] \ldots [r_{i-k+1}, r_i]$, we optimize this process as follows: For each sequence $[r_a, r_b]$, we use the already computed cost for the sequence $[r_a, r_{b-1}]$, and check if r_b increases the cost. The check needs constant time, if we maintain the Minimum Bounding Rectangle (MBR) of each sequence. This reduces the computational cost for the auxiliary table from $O(k^2Nd)$ to $O(kNd)$, where the d factor corresponds to updating the MBR and recomputing the *NCP*. To improve execution time, we also implement more time-efficient versions of our algorithms, HilbFast and iDistFast, which calculate the cost of each sequence by its extent in the 1D space. This variation relies on the assumption that records in close proximity in the multidimensional space are also likely to be close in the 1D space. Specifically, let r^{1D} denote the mapped 1D value of the quasi-identifier of record r. We approximate the cost of each sequence $[r_a, r_b]$ with its 1D extent $|r_b^{1D} - r_a^{1D}|$. The 1D extent can be computed in $O(1)$, regardless of the sequence length. Therefore, there is no need to maintain an auxiliary table at all. The computational complexity of k-anonymization is reduced by a factor of d, to $O(kN)$. Later in Section 7.1, we investigate through experiments the trade-off between information loss and computational overhead when the *NCP* is approximated with the 1D cost.

6.2 Privacy-Constrained ℓ-Diversification

Our heuristic ℓ-diversification algorithm presented in Section 4.3 performs a preprocessing step in which it partitions the input into m buckets, one for each value of the sensitive attribute. Combined with the sorting of mapped 1D data, the preprocessing step costs $O(NlogN)$. Since tabulation is not needed, the space requirement of the algorithm is $O(m)$ (i.e., constant in practice), as we only need to access the frontier of the search at each step and look back at most one group. The *NCP* computation for each ℓ-diverse group formation is $O(d \cdot m)$, and the overall complexity of the heuristic is $O(d \cdot m \cdot N)$.

6.3 Accuracy-Constrained ℓ-Diversification (Dual Problem)

The complexity of the dual ℓ-diversification heuristic from Section 5.3 increases in the case of multidimensional quasi-identifiers. The main factor causing the increase is finding the closest record to the dest group in line 18 of the *move_record* routine (Figure 17). Consider the example in Figure 21(a), and assume $E = 0.5$, measured according to *NCP*. During the candidate group formation phase, records are added to G_{prev}, according to the 1D order, as long as the accuracy bound is not exceeded: As a result, $G_{prev} = \{2, 11, 14, 30\}$ and $G_{crt} = \{51, 57\}$ (records are identified by their 1D mapped value). We have



(a) before readjustment (b) after readjustment

Fig. 21. Accuracy-constrained heuristic: extension to multidimensional quasi-identifiers.

$NCP(G_{prev}) = 0.5$, $NCP(G_{crt}) = 0.14$, and a privacy metric value of $\mathcal{PM} = 2$. In the privacy readjustment phase, the heuristic will attempt to move one of the records with *SA* value *Flu* from G_{prev} to G_{crt}, since *Flu* occurs most frequently in G_{prev}. Note that trying to move the record which is closest to G_{crt} in the mapped 1D value, namely record 30, will determine the group marked with dashed line in Figure 21(a), having $NCP = 0.57 > E$, hence the transfer will be disallowed. Nevertheless, if we choose to move the *Flu* record which is closest to G_{crt} in the multidimensional Q_T space, namely record 11, we obtain groups G'_{prev} and G'_{crt} in Figure 21(b), with $NCP(G'_{prev}) = 0.43$, $NCP(G_{crt}) = 0.36$, and a privacy metric value of $\mathcal{PM} = 3$. Therefore, by considering the distance in the multidimensional space in line 18 of the $move\_record$ routine, we can considerably improve the privacy of the partitioning. However, finding the closest record in the multidimensional space costs $O(d \cdot |src|)$, which in the worst case is $O(d \cdot N)$, as opposed to $O(1)$ in the 1D space.

Furthermore, updating the *NCP* for the candidate groups after privacy adjustment (line 22 in Figure 17) is also more expensive. By storing separately the extents of the groups in each dimension, we can compute the new *NCP* of the *dest* group (the one which is enlarged) in $O(d)$ time. However, determining for the *src* group whether the removed record reduces the group extent requires $O(d \cdot \log |src|)$, if sorted lists of records' coordinates are separately maintained for each of the d dimensions. Hence, due to the changes in lines 18 and 22, the cost of $move\_record$ increases from $O(m)$ to $O(m \cdot (d \cdot N + d \cdot \log N)) = O(m \cdot d \cdot N)$.

Finally, the property discussed in Section 5.3 that a record transfer cannot be propagated over two or more groups no longer holds in the multidimensional space. However, in our implementation we explicitly disallow this sort of transfer propagation (which is unlikely to occur in practice anyway). The resulting overall complexity of the heuristic is $O(m \cdot d \cdot N^2)$. Although the worst-case complexity is quadratic in N, we show in Section 7.3 that the dual heuristic is very efficient in practice.

Table II. CENSUS Dataset Characteristics

Attribute	Cardinality	Type
Age	79	Numerical
Gender	2	Hierarchical (2)
Education Level	17	Numerical
Marital Status	6	Hierarchical (3)
Race	9	Hierarchical (2)
Work Class	10	Hierarchical (4)
Country	83	Hierarchical (3)
Occupation	50	Sensitive Value
Salary Class	50	Sensitive Value

Table III. Experimental Parameter Values

Parameter Notation	Description	Values
k	anonymity degree	10,20,**50**,100
ℓ	diversity degree	2,3,4,**5**,6,7,8,9,10,11,12,13
d	Q_T dimensionality	2,3,4,5,6,**7**
N	data size	50000,100000,**200000**,400000
E	accuracy bound	0.2,0.3,**0.4**,0.5,0.6

7. EXPERIMENTAL EVALUATION

In this section, we evaluate our techniques against the existing state-of-the-art. All algorithms are implemented in C++ and the experiments were run on an Intel Xeon 2.8 GHz machine with 2.5GB of RAM and Linux OS.

Our workload consists of the CENSUS[10] dataset, containing information of 500,000 persons. The schema is summarized in Table II. There are nine attributes; the first seven represent the quasi-identifier Q_T, whereas the last two (i.e., Occupation and Salary) are the sensitive attributes (*SA*) (for brevity, we only include in our evaluation the *Occupation* attribute). Two of the Q_T attributes are numerical and the rest categorical; the number of levels in the taxonomy trees is shown in parentheses. We generate input tables with 50,000 to 400,000 records, by randomly selecting tuples from the entire dataset.

For the sake of comparison with previous work, we use mainly the *GCP* metric (Section 2) to measure information loss. Recall that the values of *GCP* are in the range [0, 1], and 0 is the best score (i.e., no information loss). Still, in some experiments we also include results for other metrics, to emphasize the versatility of our methods. Table III summarizes the parameter values used in the experiments; the default values are typeset in boldface.

7.1 Privacy-Constrained k-Anonymization

In the following experiments, we compare our 1D optimal privacy-constrained k-anonymization algorithm against the existing state-of-the-art techniques: the multidimensional (Mondrian) k-anonymity [LeFevre et al. 2006a], and the TopDownstering-based technique [Xu et al. 2006] (see Section 2 for details). For our optimal 1D algorithm, we consider both the Hilbert and iDistance mappings (for the latter, we set $cl = 512$ reference points, chosen from a fraction of

[10] http://www.ipums.org/.

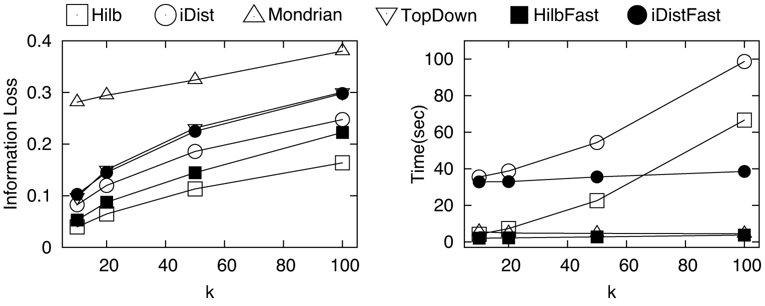


Fig. 22. Privacy-constrained k-anonymization, variable k.

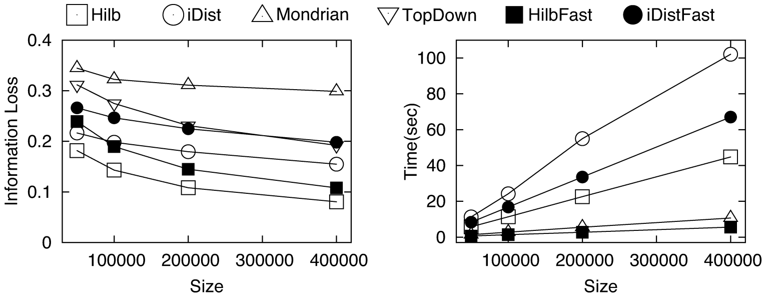


Fig. 23. Privacy-constrained k-anonymization, variable N.

10% of the data). For each of the two mappings, we consider two versions: (i) in the base version (i.e., *Hilb* and *iDist*), partitioning is guided by accurate cost estimation at the original multidimensional space. As discussed in Section 6, the amortized complexity for calculating the cost is $O(d)$, where d is the dimensionality of Q_T. (ii) in the faster variants *HilbFast* and *iDistFast* (see Section 6), the algorithm estimates the cost at the 1D space in $O(1)$ time. Since this is only an estimation of the real cost, the resulting information loss is expected to be higher.

In our first experiment, we vary anonymity degree k. Figure 22 presents the results. Both Hilb and iDist achieved lower information loss compared to TopDown and Mondrian, in all cases. In terms of execution time, Mondrian was faster. However, given the superior quality of the results, we believe that the running time of Hilb would be acceptable in practice (it was 60 sec. in the worst case). iDist is a little slower than Hilb, due to the initial phase of selecting the reference points. Both Hilb and iDist execution times include the data mapping and sorting phase. We also include the fast implementations of our algorithms in the graph. HilbFast is better than TopDown and Mondrian in terms of information loss. It is also very fast, achieving the same running time as Mondrian. iDistFast is similar to TopDown in terms of information loss; however it is much faster. The execution time of TopDown is around 2 hours, considerably longer than the other methods, so we do not include it in the graph.

Next, we vary dataset cardinality N: Figure 23 shows the results. All methods manage to reduce information loss when the size of the input increases. This is because the data density becomes higher and the probability of finding good

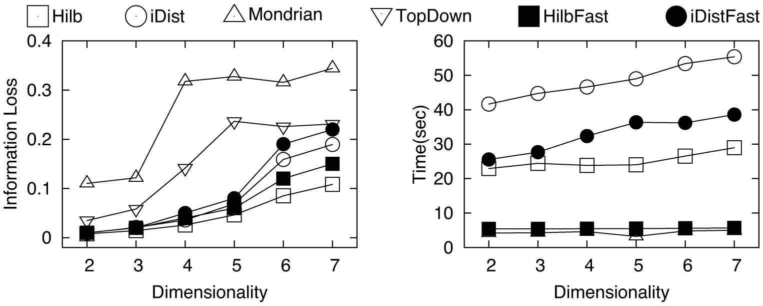


Fig. 24. Privacy-constrained k-anonymization, variable Q_T dimensionality.

partitions increases. Hilb and iDist are better than Mondrian and TopDown in all cases. As expected, the running time increases with the input size. Hilb needs only 40 sec. to anonymize 400,000 records, when $k = 50$. The execution time of TopDown (not included in the graph) is considerably higher: it ranges from 8 min. for $50,000$ records to 6 hours for $400,000$ records.

In Figure 24 we vary the dimensionality d of the quasi-identifier by projecting the original 7D data to fewer dimensions. Since Hilb and iDist are optimal for $d = 1$, for low dimensionality their information loss is close to 0 (note that the information loss of the optimal solution is typically greater than 0 due to generalization). Interestingly, for larger dimensionality, Hilb outperforms its competitors by a larger factor; therefore Hilb is suitable for real-life high-dimensional data. The running time is affected only slightly by dimensionality. Our methods face a small overhead due to the calculation of the cost of each partition in the multidimensional space.

7.2 Privacy-Constrained ℓ-Diversification

We compare our linear 1D heuristic for privacy-constrained ℓ-diversification against an ℓ-diverse variation of Mondrian, which uses the original median split heuristic and checks for each partition whether the ℓ-diversity property is satisfied. We defer the comparison against Anatomy [Xiao and Tao 2006a] until Section 7.4, since Anatomy does not use generalization and the *GCP* metric would penalize the method unfairly.

In Figure 25 we vary the value of ℓ. Hilb is the best in terms of information loss, followed closely by iDist. The execution time of Hilb is very low (roughly 5 sec.) and similar to Mondrian. iDist is slower, due to the initial mapping phase.

Next (Figure 26) we vary dataset cardinality N. As N increases, so does the data density; therefore, information loss decreases slightly for both Hilb and iDist. In terms of execution time, Hilb and Mondrian have similar performance, but Hilb is far superior in terms of information loss. Note that in all experiments the entire input table fits in the main memory. If the input table is larger than the main memory, the I/O cost of Mondrian will be much larger, since it needs to scan the input at each split. In contrast, our methods require a single scan of the input (excluding the sorting phase). Also observe that Mondrian may exhibit unpredictable, nonmonotonic behavior with respect to ℓ or N. The reason is that for particular inputs, the ℓ-diversity property cannot be satisfied by any split.

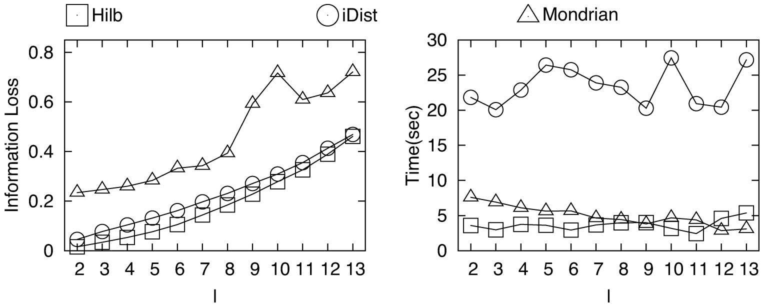


Fig. 25. Privacy-constrained ℓ-diversification, variable ℓ.

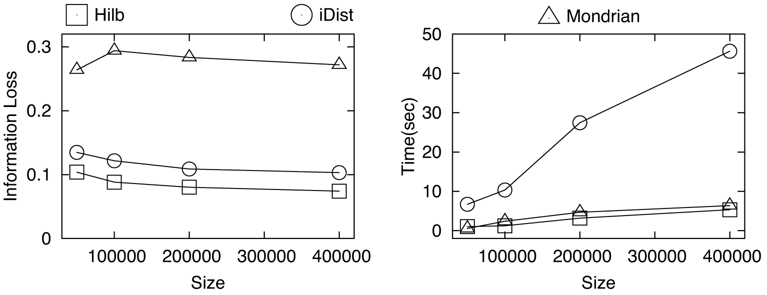


Fig. 26. Privacy-constrained ℓ-diversification, variable N.

In Figure 27 we vary the dimensionality d of Q_T. Hilb and iDist clearly outperform Mondrian. Observe that Mondrian deteriorates sharply as d increases. Also note that the execution time is virtually unaffected by dimensionality.

In order to evaluate the performance of our heuristic for various data distributions, we also include an experiment with synthetic data. Our goal is to show the behavior of the proposed algorithms when certain correlation patterns exist between quasi-identifier and SA attributes. Intuitively, if there is no such correlation (i.e., SA values are randomly distributed in the Q_T space), ℓ-diversification is easier to solve, because sufficient records with distinct SA can be found in close proximity to each other. On the other hand, if there is strong correlation among Q_T and SA, ℓ-diverse groups need to grow larger in the Q_T space in order to fulfill the privacy requirement. The worst-case scenario is when there is both high correlation and the Q_T are randomly distributed in the data space, because in this case groups will span large regions of the Q_T space.

We generate a 5D dataset, with a 4D Q_T randomly distributed in the data space. Both Q_T and SA have numerical values between 0 and 9, and the SA has a linear dependence on Q_T as follows. We have

$$SA = \begin{cases} (\beta_1 \times A_1 + \beta_2 \times A_2 + \beta_3 \times A_3 + \beta_4 \times A_4) \mod 10 & , if\ rand() < \alpha \\ 10 \times rand() & , otherwise \end{cases}$$

where $A_1 \ldots A_4$ are the Q_T values, β_i are coefficients of a linear function in $A_1 \ldots A_4$, and $rand()$ returns a random value in [0, 1). Parameter α controls the degree of correlation among Q_T and SA: A small value (e.g., close to 0) means

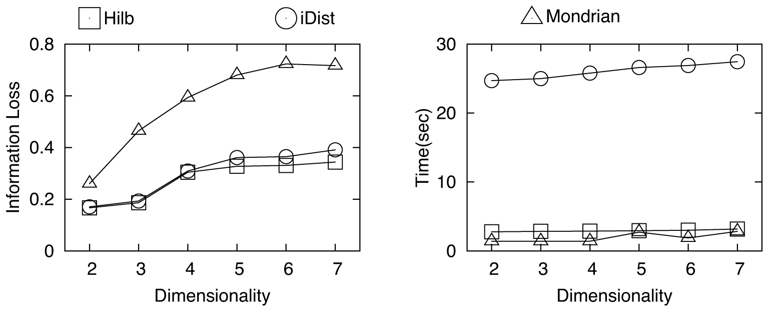


Fig. 27. Privacy-constrained ℓ-diversification, variable Q_T dimensionality.

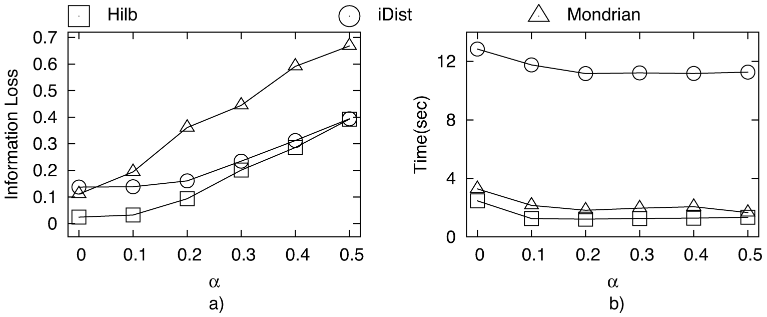


Fig. 28. Privacy-constrained ℓ-diversification, synthetic data.

there is little correlation, whereas $\alpha = 1$ signifies complete correlation among Q_T and *SA*.

We vary α between 0 and 0.5. Figure 28(a) shows the information loss results: Mondrian performs well for 0 correlation. In fact, this scenario favors Mondrian the most, since Q_T is also randomly distributed, and balanced splits are possible. iDist performs slightly worse than Mondrian because it is designed for skewed data, where clustering is efficient. As correlation grows, the information loss of Mondrian deteriorates quickly: Due to its constraint that group extents should not overlap, only few splits of the dataset can be performed. Both Hilb and iDist outperform Mondrian by a large margin, and their performance becomes similar as correlation increases, because diverse records are situated far apart and the accuracy of the mapping becomes less influential in group formation.

Figure 28(b) shows execution time: All methods require a longer execution time for the 0 correlation case. In the case of Mondrian, more splits are possible, and the algorithm executes more iterations. For Hilb and iDist, the heuristic has more flexibility to form groups, and therefore more choices are considered. As correlation increases, there are fewer choices to form groups and execution time decreases. Similarly, for Mondrian, only few splits are completed and the algorithm terminates faster.

Finally, we evaluate the performance of our algorithms with respect to information loss metrics other than *GCP*: Specifically, we consider: (i) weighted maximum *NCP* (i.e., the *NCP* of the group with largest extent times group

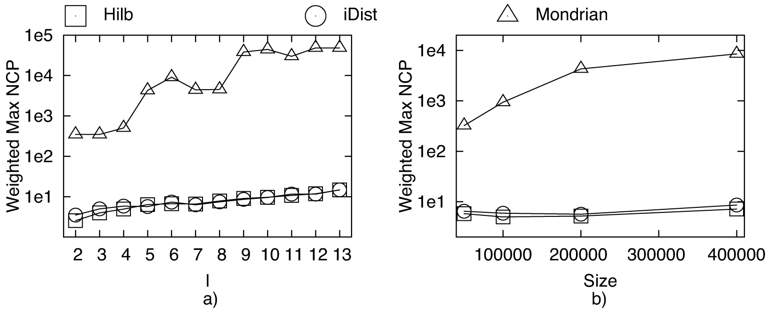


Fig. 29. Privacy-constrained ℓ-diversification, weighted maximum NCP metric.

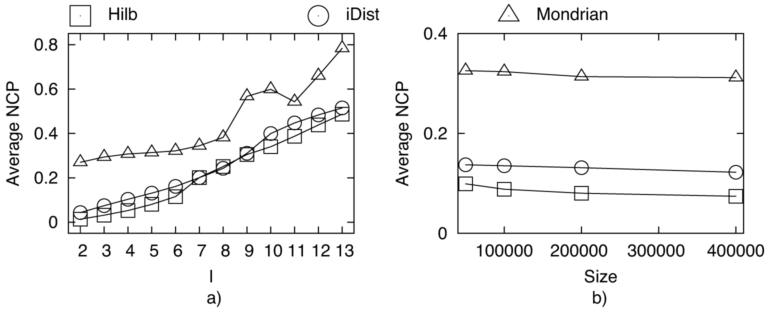


Fig. 30. Privacy-constrained ℓ-diversification, average NCP metric.

cardinality), (ii) average *NCP*, and (iii) the average group volume weighted by cardinality. The group volume is determined similarly to *NCP* in Eq. (1), except that we compute the product of extents over all attributes, instead of sum (for those attributes which have a unique value within a group, we measure the extent as 1/AttributeCardinality, to avoid multiplication by 0). Note that metric (i) is a maximum-loss-per-group metric, like $\mathcal{IL}_\infty$, but weighted by cardinality, like *GCP*, whereas metric (ii) is a normalized version of $\mathcal{IL}_1$. For this experiment, we use the CENSUS dataset with all seven quasi-identifier attributes.

Figure 29 shows the results for the weighted maximum *NCP* metric. Our methods are clearly superior to Mondrian in all cases. The information loss of Mondrian is several orders of magnitude higher because it generates groups with both large extent, and large number of records. Note that for varying dataset size (Figure 29(b)) a different trend is observed than for the *GCP* metric: The information loss first decreases (when dataset size grows from $50k$ to $100k$), but then exhibits an increasing trend. This is the result of two factors: On one hand, data density increases, which leads to lower information loss. On the other hand, increasing the number of records introduces more outliers in the data. Since here we measure the extent of the maximum group (and not average), the latter factor prevails for larger sizes, and the resulting information loss is higher.

In Figure 30, we present the results for the average *NCP* metric (i.e., normalized $\mathcal{IL}_1$), belonging to the family of metrics that our theoretical analysis

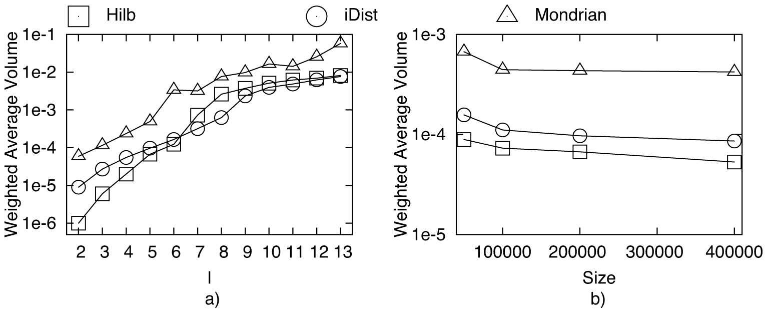


Fig. 31. Privacy-constrained ℓ-diversification, weighted average volume metric.

directly applies to. The measured trends are similar to those observed for the *GCP* metric (Figures 25 and 26).

In Figure 31 we present the results for the weighted average volume metric. Again, our methods outperform Mondrian by a considerable margin. Note that in Figure 31(a), iDist is better than Hilb in some cases. This is a consequence of the fact that iDist performs a clustering-based mapping of quasi-identifiers. Since the optimization objective in iDist is cluster volume, iDist is more suitable for the volume metric than Hilbert, whose mapping is more influenced by the Manhattan distance between records.

7.3 Accuracy-Constrained ℓ-Diversification (Dual Problem)

In this section we evaluate the heuristic for the accuracy-constrained (*AC*) problem presented in Section 5.3. Since there is no other existing algorithm that solves the dual problem, we compare *AC* against the IPCSD algorithm introduced in Figure 3. IPCSD is based on the solution to the direct problem, and performs a binary search to find the maximum value of ℓ for which the accuracy bound E is not exceeded. We use IPCSD in conjunction with both the 1D direct heuristic from Section 4.3 (IPCSD-1D), as well as Mondrian [LeFevre et al. 2006a] (IPCSD-Mondrian). Mondrian can handle fractional values of ℓ (recall the example in Figure 4). On the other hand, IPCSD-1D can only handle integer values of ℓ; nevertheless, in some of the graphs, the privacy value appears as a fractional number. The reason is the following: Our dataset has 7 dimensions for the Q_T; if an experiment uses fewer than 7 of them, there are multiple ways to project the original data. We run the experiments with all possible projections, and report the average privacy value, which is not necessarily an integer. We measure the value of E according to *NCP*, with values between 0 and 1. For the 1D methods, we consider only the Hilbert transformation.

In Figure 32 we vary E and measure the privacy metric *PM* for a quasi-identifier with dimensionality $d = 3$. *AC* always outperforms both IPCSD methods in terms of privacy (i.e., PM metric), whereas IPCSD-1D obtains the lowest privacy, due to its restriction of allowing only one *SA* value per group. *AC* is also faster in terms of execution time (recall that IPCSD includes a multiple-stage binary search phase, which increases the computational overhead). For *AC*, the processing cost increases with E, since the cardinality of candidate groups grows, and so does the cost of the privacy adjustment phase.

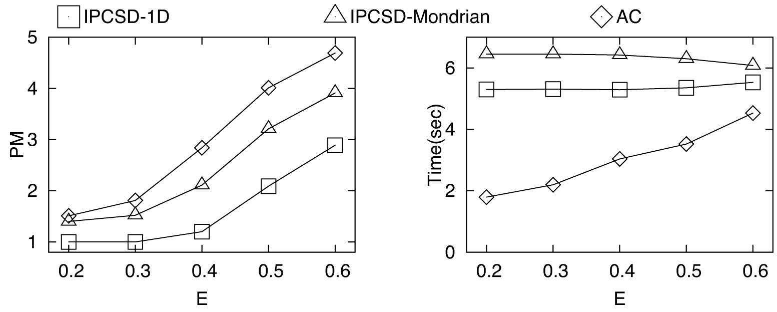


Fig. 32. Accuracy-constrained ℓ-diversity, variable accuracy bound E.

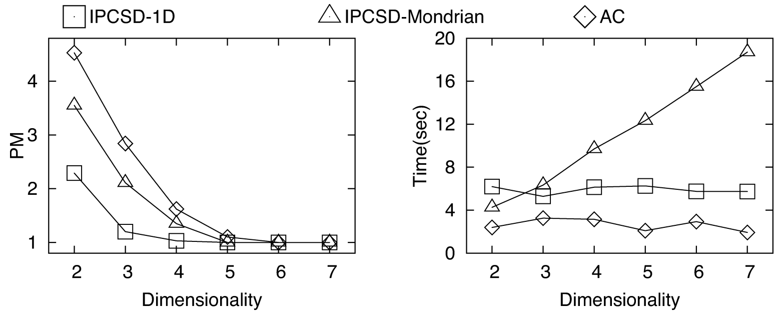


Fig. 33. Accuracy-constrained ℓ-diversity, variable Q_T dimensionality.

In Figure 33 we evaluate the effect of varying quasi-identifier dimensionality. The privacy obtained by AC is considerably better at lower dimensionality. However, as d grows, the difference between AC and the IPCSD flavors attenuates, due to the dimensionality curse: As d increases, fewer records can be located within an extent that does not exceed E. AC remains more efficient in terms of computational overhead. Note that, despite their worst-case complexity which is linear in d (as discussed in Section 6), both 1D methods (AC and IPCSD-1D) do not exhibit an increase in execution time when d grows, because they work in the mapped 1D space. Hence, the worst-case complexity is not observed in practice. On the other hand, IPCSD-Mondrian needs to examine candidate splits in all dimensions, therefore its overhead is linear in d.

In Figure 34, we vary the number of records N, $d = 3$. AC has the lowest execution time for most of the considered N range, while it also maintains its advantage in terms of privacy.

7.4 Precision of Data Analysis Queries

In addition to the general-purpose GCP metric, in this section we employ a realistic query workload, as suggested by LeFevre et al. [2006b]. We compare the privacy-constrained ℓ-diversity versions of Hilb and iDist against Anatomy and ℓ-diverse Mondrian. Anonymized data can be used to extract statistics and assist decision-making. Since these are typical OLAP operations, our workload consists of the following type of aggregation queries.

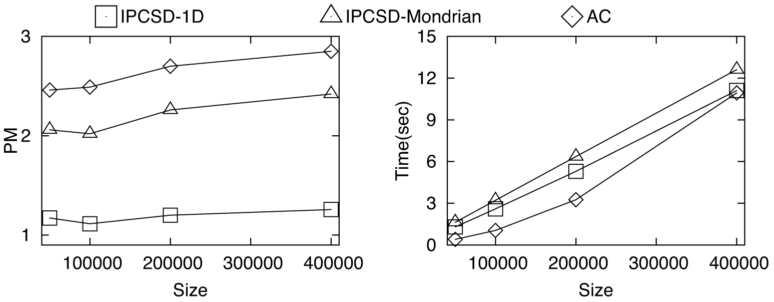


Fig. 34. Accuracy-constrained ℓ-diversity, variable N.

```
SELECT QT1, QT2,..., QTi, COUNT(*)
FROM data
WHERE SA = val
GROUP BY QT1, QT2,..., QTi
```

Each QT_i is an attribute of the quasi-identifier (e.g., *Age*, *Gender*), whereas *SA* is a sensitive attribute (e.g., Occupation). The OLAP datacube [Harinarayan et al. 1996] consists of all group-bys for all possible combinations of the quasi-identifier attributes. Interdependencies among group-bys are captured by the datacube lattice. Level i of the lattice corresponds to all group-bys over exactly i attributes (the higher the level, the finer the granularity of the group-by). We represent the cube as a multidimensional array; the cells that do not correspond to a result tuple of the aforesaid query are set to 0.

We use the CENSUS dataset and compute the entire datacube for (i) the original microdata (P cube) and (ii) the anonymized tables (Q cube). Obviously, Q is an estimation of P. Each cell of Q is computed as follows: For Anatomy, which does not use generalization, the estimation is straightforward since the exact quasi-identifier and the probability of an *SA* value for a specific record are given. For the generalization-based methods, we take into account the intersection of the query with each group, assuming a uniform distribution of records within the group.

Ideally, the values of all cells in cube Q should be equal to the values in the corresponding cells of P. Several methods exist to measure similarity. Xiao and Tao [2006a] use the relative error: $RE = |P_C - Q_C|/P_C$, where P_C and Q_C are values of a cell in P and Q, respectively. However, this metric is undefined for $P_C = 0$. In our experiments we use KL-Divergence (KLD), which has been acknowledged as a representative metric in the data anonymization literature [Kifer and Gehrke 2006]. P and Q are modeled as multidimensional probability distribution functions. The estimation error is defined as

$$KLD(P, Q) = \sum_{\forall cell C} P_C \log \frac{P_C}{Q_C}.$$

In the best case (i.e., identical P, Q), $KLD = 0$.

In Figure 35(a), we show the query precision for varying ℓ at level 2 of the datacube lattice (i.e., all group-bys with two attributes). For small ℓ, Hilb and iDist clearly outperform the competitors. Hilb is two orders of magnitude better

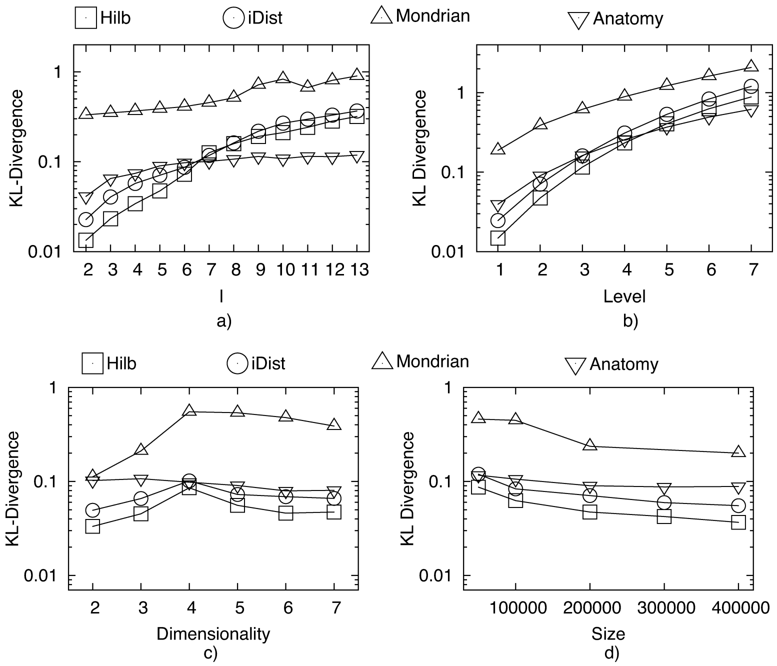


Fig. 35. OLAP query precision results: generalization-based methods vs Anatomy.

than Mondrian, and one order of magnitude better than Anatomy, despite the fact that Anatomy is *not* using generalization but publishes the exact quasi-identifier. As ℓ increases, the extent of the anonymized groups grows accordingly in all dimensions. This is a clear disadvantage for all generalization methods; however, even for larger ℓ values, our methods outperform Mondrian by an order of magnitude, and their precision is only marginally worse than Anatomy.

In Figure 35(b) we show the query precision for different levels of the OLAP lattice. Hilb and iDist are better than Mondrian by up to an order of magnitude, and also outperform Anatomy. Hilb and iDist are better at lower levels of the lattice (i.e., coarse-grained aggregation), since the extent of the anonymized groups is likely to be completely included in the query range. For finer granularity, Anatomy performs equally well as our methods, since it is favored by small query ranges.

In Figure 35(c) we focus on level 2 of the lattice, and vary the dimensionality d of the quasi-identifier. Lower dimensionality results to more compact ℓ-diverse groups, which improves precision. However, since the group-by level is kept constant, a lower quasi-identifier dimensionality also results in a smaller extent (i.e., finer granularity) of the queries, which decreases query answering precision. Depending on the domains of the quasi-identifier attributes, any of the two effects may become significant. This is why there is an increasing trend until $d = 4$ and a decreasing trend afterwards. Hilb and iDist maintain an advantage over the competitors. Observe that Anatomy is not affected by dimensionality, since it does not use generalization.

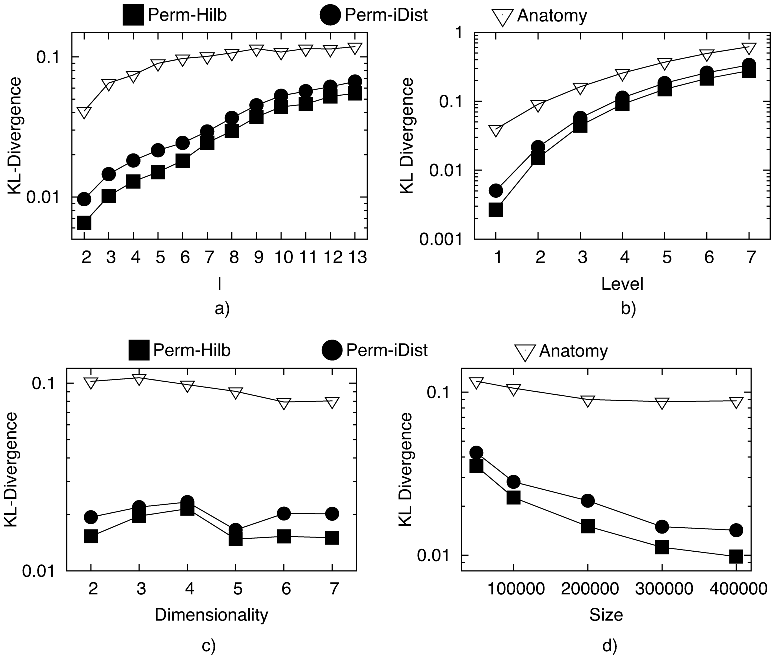


Fig. 36. OLAP query precision results: Perm-Hilb and Perm-iDist vs. Anatomy.

In Figure 35(d) we vary the size of the input N, at level 2 of the lattice. Since the extent of the queries is constant but the density of data in the quasi-identifier space increases, precision increases with N.

Note that our group formation algorithms are orthogonal to the publication format chosen. If a certain application does not require generalization of quasi-identifiers, we can adopt the permutation-based publication format of Anatomy, which enhances the precision of query answering. In the next experiment, we evaluate the query precision of the proposed methods in conjunction with permutation-based publishing (Perm-Hilb and Perm-iDist), against Anatomy. Figure 36(a)–(d) shows that our methods clearly outperform Anatomy in all cases, due to their superior group formation heuristic which accounts for proximity in the quasi-identifier space. Perm-Hilb and Perm-iDist maintain the trends observed for their generalization-based counterparts Hilb and iDist (Figure 35), but they always out-perform Anatomy by up to one order of magnitude.

7.5 Discussion

We demonstrated that for k-anonymization, our algorithms are superior to existing techniques in terms of information loss. Hilb is the best, but is a bit slower than Mondrian. If speed is essential, HilbFast can be used. It is as fast as Mondrian and its quality is only slightly worse that Hilb.

For privacy-constrained ℓ-diversification, Hilb is the clear winner. It is by far superior in terms of information loss and precision for real queries; it is also as fast as its competitors. Interestingly, Hilb out-performs Anatomy in most cases, although Anatomy implements a less secure model, by publishing the exact quasi-identifiers. This happens because Anatomy ignores the distance of the tuples in the Q_T space (see Section 2.3).

iDist also performed well, but slightly worse than Hilb. We used iDist mainly to demonstrate the versatility of our framework. For specific applications, other multidimensional to 1D mappings may be more appropriate. Any such mapping can be used in our framework.

Our privacy-constrained solutions scale well with the input size, since the computational complexity is linear, the required memory is constant, and only one scan of the data is necessary (provided the dataset is sorted).

In practice, the information loss incurred by ℓ-diversification is dependent on the data distribution. In particular, if the data is skewed, there will be a high correlation among quasi-identifier and SA values. In this case, to satisfy the diversity requirement, groups need to span larger regions of the quasi-identifier space. The experimental evaluation shows that the proposed methods clearly outperform competitor techniques, which do not handle data skewness well.

Lastly, our accuracy-constrained heuristic for the dual problem (i.e., *AC*) is the first to appear in the literature. Compared to the iterative methods based on the direct solution (i.e., *IPCSD*), *AC* achieves superior privacy with faster execution time.

8. CONCLUSIONS

In this article, we developed a framework for solving the privacy-constrained and accuracy-constrained data anonymization problems. Our approach relies on mapping the multidimensional quasi-identifiers to one dimension. We identified a set of properties for the optimal 1D solutions. Guided by these properties, we developed efficient heuristics at the 1D space. We used two popular transformations, namely the Hilbert curve and iDistance, to solve the multidimensional problems through 1D mapping; other transformations can easily be incorporated in our framework. The experiments demonstrate that our methods clearly outperform the existing approaches in terms of execution time and information loss. Moreover, our algorithms are efficient, therefore they are applicable to large datasets.

In the future we plan to extend our framework to other privacy paradigms, such as t-closeness and m-invariance. Furthermore, we intend to study the privacy- and accuracy-constrained problems for data streams. Streaming data poses two additional challenges: First, not all data are available from the beginning; instead, new data arrive continuously. Second, the data have expiration deadlines; therefore, it is crucial to minimize the computational overhead of anonymization algorithms.

REFERENCES

Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., and Zhu, A. 2006. Achieving anonymity via clustering. In *Proceedings of ACM Conference on Principles of Database Systems (PODS)*, 153–162.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. 2005. Approximation algorithms for k-anonymity. *J. Privacy Technol.* Article number: 20051120001.

BAYARDO, R. J. AND AGRAWAL, R. 2005. Data privacy through optimal k-anonymization. In *Proceedings of the International Conference on Data Engineering (ICDE)*, 217–228.

BYUN, J.-W., KAMRA, A., BERTINO, E., AND LI, N. 2007. Efficient k -anonymization using clustering techniques. In *Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA)*, 188–200.

FROOMKIN, A. 2000. The death of privacy. *Stanford Law Rev. 52,* 5, 1461–1543.

GHINITA, G., KARRAS, P., KALNIS, P., AND MAMOULIS, N. 2007. Fast data anonymization with low information loss. In *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 758–769.

HARINARAYAN, V., RAJARAMAN, A., AND ULLMAN, J. D. 1996. Implementing data cubes efficiently. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 205–216.

IWUCHUKWU, T. AND NAUGHTON, J. F. 2007. k-Anonymization as spatial indexing: toward scalable and incremental anonymization. In *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 746–757.

IYENGAR, V. S. 2002. Transforming data to satisfy privacy constraints. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, 279–288.

KIFER, D. AND GEHRKE, J. 2006. Injecting utility into anonymized datasets. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 217–228.

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2005. Incognito: Efficient full-domain k-anonymity. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 49–60.

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2006a. Mondrian multidimensional k-anonymity. In *Proceedings of International Conference on Data Engineering (ICDE)*.

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2006b. Workload-aware anonymization. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, 277–286.

LI, N., LI, T., AND VENKATASUBRAMANIAN, S. 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of International Conference on Data Engineering (ICDE)*, 106–115.

MACHANAVAJJHALA, A., GEHRKE, J., KIFER, D., AND VENKITASUBRAMANIAM, M. 2006. lDiversity: Privacy beyond k-anonymity. In *Proceedings of International Conference on Data Engineering (ICDE)*.

MEYERSON, A. AND WILLIAMS, R. 2004. On the complexity of optimal k-anonymity. In *Proceedings of ACM Conference on Principles of Database Systems (PODS)*, 223–228.

MOON, B., JAGADISH, H., AND FALOUTSOS, C. 2001. Analysis of the clustering properties of the Hilbert space-filling curve. *IEEE Trans. Knowl. Data Eng. 13,* 1, 124–141.

MUTHUKRISHNAN, S. AND SUEL, T. 2005. Approximation algorithms for array partitioning problems. *J. Algo. 54,* 1, 85–104.

PARK, H. AND SHIM, K. 2007. Approximate algorithms for k-anonymity. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 67–78.

SAMARATI, P. 2001. Protecting respondents' identities in micro-data release. *IEEE Trans. Knowl. Data Eng. 13,* 6, 1010–1027.

SWEENEY, L. 2002. k-Anonymity: A model for protecting privacy. *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst. 10,* 5, 557–570.

WONG, R., FU, A., PEI, J., WANG, K., WAN, S., AND LO., C. 2006. Multidimensional k-anonymization by linear clustering using space-filling curves. Tech. rep. 2006-27, Simon Fraser University. March.

XIAO, X. AND TAO, Y. 2006a. Anatomy: Simple and effective privacy preservation. In *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 139–150.

XIAO, X. AND TAO, Y. 2006b. Personalized privacy preservation. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 229–240.

XIAO, X. AND TAO, Y. 2007. m-Invariance: Towards privacy preserving re-publication of dynamic datasets. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 689–700.

XU, J., WANG, W., PEI, J., WANG, X., SHI, B., AND FU, A. 2006. Utility-based anonymization using local recoding. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–790.

ZHANG, Q., KOUDAS, N., SRIVASTAVA, D., AND YU, T. 2007. Aggregate query answering on anonymized tables. In *Proceedings of International Conference on Data Engineering (ICDE)*, 116–125.

ZHANG, R., KALNIS, P., OOI, B. C., AND TAN, K.-L. 2005. Generalized multidimensional data mapping and query processing. *ACM Trans. Datab. Syst. 30,* 3, 661–697.

Received March 2008; revised September 2008; accepted December 2008