Towards Application-Oriented Data Anonymization*†

Li Xiong‡ Kumudhavalli Rangachari§

Abstract

Data anonymization is of increasing importance for allowing sharing of individual data for a variety of data analysis and mining applications. Most of existing work on data anonymization optimizes the anonymization in terms of data utility typically through one-size-fits-all measures such as data discernibility. Our primary viewpoint in this paper is that each target application may have a unique need of the data and the best way of measuring data utility is based on the analysis task for which the anonymized data will ultimately be used. We take a top-down analysis of typical application scenarios and derive application-oriented anonymization criteria. We propose a prioritized anonymization scheme where we prioritize the attributes for anonymization based on how important and critical they are to the application needs. Finally, we present preliminary results that show the benefits of our approach.

1 Introduction

Data privacy and identity protection is a very important issue in this day and age when huge databases containing a population's information need to be stored and distributed for research or other purposes. For example, the National Cancer Institute initiated the Shared Pathology Informatics Network (SPIN)[1] for researchers throughout the country to share pathology-based data sets annotated with clinical information to discover and validate new diagnostic tests and therapies, and ultimately to improve patient care. However, individually identifiable health information is protected under the Health Insurance Portability and Accountability Act (HIPAA)[2]. The data have to be sufficiently anonymized before being shared over the network.

These scenarios can be generalized into the problem of privacy preserving data publishing where a data custodian needs to distribute an anonymized view of the data that does not contain individually identifiable information to data recipient(s) for various data analysis and mining tasks. Privacy preserving data publishing has been extensively studied in recent years and a few principles have been proposed that serve as criteria for judging whether a published dataset provides sufficient privacy protection [40, 34, 43, 3, 32, 53, 35, 37]. Notably, the earliest principle, k-anonymity [40], requires a set of k records (entities) to be indistinguishable from each other based on a quasi-identifier set, and its extension, l-diversity [34], requires every group to contain at least l well-represented sensitive values. A large body of work contributes to transforming a dataset to meet a privacy principle (dominantly k-anonymity) using techniques such as generalization, suppression (removal), permutation and swapping of certain data values while minimizing certain cost metrics [20, 50, 36, 9, 2, 17, 10, 59, 29, 30, 31, 49, 27, 51, 58].

Most of these methods aim to optimize the data utility measured through a one-size-fitsall cost metric such as general discernibilty or information loss. Few works have considered targeted applications like classification and regression [21, 50, 17, 31] but do not model other kinds of applications nor provide a systematic or adaptive approach for handling various needs.

Contributions. Our primary viewpoint in this paper is that each target application may have a unique need of the data and the best way of measuring data utility is based on the analysis task for which the anonymized data will ultimately be used. We aim to adapt existing methods by incorporating the application needs into the anonymization process, thereby increasing its utility to the target applications.

The paper makes a number of contributions. First, we take a top-down analysis of potential application scenarios and devise models and schemes to represent application requirements in terms of relative attribute

*P3DM'08, April 26, 2008, Atlanta, Georgia, USA.

†This research is partially supported by an Emory URC grant and an Emory ITSC grant.

‡Dept. of Math & Computer Science, Emory University

§Dept. of Math & Computer Science, Emory University

[1]Shared Pathology Informatics Network (SPIN). http://www.cancerdiagnosis.nci.nih.gov/spin/

[2]Health Insurance Portability and Accountability Act (HIPAA). http://www.hhs.gov/ocr/hipaa/. State law or institutional policy may be considered as well.

importance that can be specified by users or learned from targeted analysis and mining tasks. Second, we propose a prioritized anonymization scheme where we prioritize the attributes for anonymization based on how important and critical they are to the application needs. We devise a prioritized cost metric that allows users to assign different weights to different attributes and adapt existing generalization-based anonymization methods in order to produce an optimized view for the user applications. Finally, we present preliminary results that show the benefits of our approach.

2 Related Work

Our research is inspired and informed by a number of related areas. We discuss them briefly below.

Privacy Preserving Access Control and Statistical Databases. Previous work on multilevel secure relational databases [22] provides many valuable insights for designing a fine-grained secure data model. Hippocratic databases [7, 28, 5] incorporate privacy protection within relational database systems. Byun et al. presented a comprehensive approach for privacy preserving access control based on the notion of purpose [14]. While these mechanisms enable multilevel access of sensitive information through access control at a granularity level up to a single attribute value for a single tuple, micro-views of the data are desired where even a single value of a tuple attribute may have different views [13]. Research in statistical databases has focused on enabling queries on aggregate information (e.g. sum, count) from a database without revealing individual records [1]. The techniques developed have focused only on aggregate queries and relational data types.

Privacy Preserving Data Mining. One data sharing model is the mining-as-a-service model, in which individual data owners submit the data to a data collector for mining or a data custodian outsources mining to an untrusted service provider. The main approach is random perturbation that transforms data by adding random noise in a principled way [8, 48]. The main notion of privacy studied in this context is data uncertainty as versus individual identifiability. There are studies focusing on specific mining tasks such as decision tree [8, 12], association rule mining [39, 15, 16], and on disclosure analysis [26, 19, 42, 12]. A main advantage of data anonymization as opposed to data perturbation is that the released data remain "truthful", though at a coarse level of granularity. This allows various analysis to be carried out using the data, including selection.

Another related area is distributed privacy preserving data sharing and mining that deals with data sharing for specific tasks across multiple data sources in a distributed manner [33, 44, 23, 25, 46, 56, 45, 4, 6, 47, 24, 54, 11, 55]. The main goal is to ensure data is not disclosed among participating parties. Common approaches include data approach that involves data perturbation and protocol approach that applies random-response techniques.

Data Anonymization The work in this paper has its closest roots in data anonymization that provides a micro-view of the data while preserving privacy of individuals. The work in this area can be classified into a number of categories. The first one aims at devising generalization principles in that a generalized table is considered privacy preserving if it satisfies a *generalization principle* [40, 34, 43, 3, 32, 53, 35, 37]. Recent work[52] also considered *personalized anonymity* to guarantee minimum generalization for every individual in the dataset. Another large body of work contributes to the algorithms for transforming a dataset to one that meets a generalization principle and minimizes certain quality metrics. Several hardness results [36, 2] show that computing the optimal generalized table is NP-hard and the result suffers severe information loss when the number of quasi-identifier attributes are high. Optimal solutions [9, 29] enumerate all possible generalized relations with certain constraints using heuristics to prune the search space. Greedy solutions [20, 50, 17, 10, 59, 30, 31, 49] are proposed to obtain a suboptimal solution much faster. A few works are suggesting new approaches in addition to generalization, such as releasing marginals [27], anatomy technique [51], and permutation technique [58], to improve the utility of the published dataset. Another thread of research is focused on disclosure analysis [35]. A few works considered targeted classification and regression applications [20, 50, 17, 31].

Our work builds on top of the existing generalization principles and anonymization techniques and aims to adapt existing solutions for application-oriented anonymization that provides an optimal view for targeted applications.

3 Privacy Model

Among the many identifiability based privacy principles, k-anonymity [41] and its extension l-diversity [34]

are the two most widely accepted and serve as the basis for many others, and hence, will be used in our discussions and illustrations. Our work is orthogonal to these privacy principles. Below we introduce some terminologies and illustrate the basic ideas behind these principles.

In defining anonymization, attributes of a given relational table T, are characterized into three types. *Unique identifiers* are attributes that identify individuals. Known identifiers are typically removed entirely from released micro-data. *Quasi-identifier set* is a minimal set of attributes (X_1, ..., X_d) that can be joined with external information to re-identify individual records. We assume that a quasi-identifier is recognized based on domain knowledge. *Sensitive attributes* are those attributes that an adversary should not be permitted to uniquely associate their values with a unique identifier.

Table 1: Illustration of Anonymization: Original Data and Anonymized Data

Name	Age	Gender	Zipcode	Diagnosis
Henry	25	Male	53710	Influenza
Irene	28	Female	53712	Lymphoma
Dan	28	Male	53711	Bronchitis
Erica	26	Female	53712	Influenza

Original Data

Name	Age	Gender	Zipcode	Disease
*	[25 − 28]	Male	[53710-53711]	Influenza
*	[25 − 28]	Female	53712	Lymphoma
*	[25 − 28]	Male	[53710-53711]	Bronchitis
*	[25 − 28]	Female	53712	Influenza

Anonymized Data

Table 1 illustrates an original relational table of personal information. Among the attributes, $Name$ is considered as an identifier, (Age, $Gender$, $Zipcode$) is considered as a quasi-identifer set, and $Diagnosis$ is considered as a sensitive attribute. The k-anonymity model provides an intuitive requirement for privacy in stipulating that no individual record should be uniquely identifiable from a group of k with respect to the quasi-identifier set. The set of all tuples in T containing identical values for the quasi-identifier set X_1, ..., X_d is referred to as an *Equivalence Class*. T is *k-anonymous* with respect to X_1, ..., X_d if every tuple is in an equivalence class of size at least k. A *k-anonymization* of T is a transformation or generalization of the data T such that the transformation is k-anonymous. The l-diversity model provides a natural extension to incorporate a nominal sensitive attribute S. It requires that each equivalence class also contains at least l well-represented distinct values for S. Typical techniques to transform a dataset to satisfy k-anonymity include data generalization, data suppression, and data swapping. Table 1 also illustrates one possible anonymization with respect to a quasi-identifier set ($Age, Gender, Zipcode$) using data generalization that satisfies 2-anonymity and 2-diversity.

4 Application-Oriented Anonymization

Our key hypothesis is that by considering important application requirements, the data anonymization process will achieve a better tradeoff between general data utility and application-specific data utility. We first take a top-down analysis of typical application scenarios and analyze what requirements and implications they pose to the anonymization process. We then present our prioritized optimization metric and anonymization techniques that aim to prioritize the anonymization for individual attributes based on how important they are to target applications.

4.1 Anonymization Goals There are different types of target applications for sharing anonymized data including: 1) query applications supporting ad-hoc queries, 2) applications with a specific mining task such as classification or clustering, and 3) exploratory applications without a specific mining task. We consider two typical scenarios of these applications on anonymized medical data and analyze their implications on the anonymization algorithms.

Scenario 1. Disease-specific public health study. In this study, researchers select a subpopulation of certain health condition (e.g. $Diagnosis =$ "$Lymphoma$"), and study their geographic and demographic distribution, reaction to certain treatment, or survival rate. An example is to identify geographical patterns for the health condition that may be associated with features of the geographic environment.

Scenario 2. Demographic / population study. In this study, researchers may want to study a certain demographic subpopulation (e.g. $Gender = Male$ and $Age > 50$), and perform exploratory analysis or learn classification models based on demographic information and clinical symptoms to predict diagnosis.

The data analysis for the mentioned applications is typically conducted in two steps: 1) subpopulation identification through a selection predicate, and 2) analysis on the identified subpopulation including mining tasks such

as clustering or classification of the population with respect to certain class labels. Given such a two-step process, we identify two requirements for optimizing the anonymization for applications: 1) maximize precision and recall of subpopulation identification, and 2) maximize quality of the analysis.

We first categorize the attributes with respect to the applications on the anonymized data and then explain how the application requirement and optimization goal transform to concrete criteria for application-oriented anonymization. Given an anonymized relational table, each attribute can be characterized by one of the following types with respect to the target applications.

- *Selection attributes* are those attributes used to identify a subpopulation (e.g. *Diagnosis* in Scenario 1 and *Gender* and *Age* in Scenario 2).
- *Feature attributes* are those attributes used to perform analysis such as classifying or clustering data (e.g. *Zipcode* in Scenario 1 for geographic location based analysis).
- *Target attributes* are the class label or attributes for which the classification or prediction are trying to predict (e.g. *Diagnosis* in Scenario 2). Target attributes are not applicable for unsupervised learning tasks such as clustering.

Given the above categorization and the goals in optimizing anonymization for target applications, we derive a set of generalization criteria for the different types of attributes in our anonymization model.

- *Discernibility of selection attributes or predicates.* If a selection attribute is part of the quasi-identifier set and is subject to generalization, it may result in an imprecise query selection. For example, if the *Age* attribute is generalized into ranges of $[0-40]$ and $[40\ above]$, the selection predicate $Age > 50$ in Scenario 2 will result in an imprecise subpopulation. In order to maximize the precision of the population identification, the generalization of the selection attributes should be minimized or adapted to the selection predicates so that the discernibility of selection attributes or predicates are maximized.
- *Discernibility of feature attributes.* For most mining tasks, the anonymized dataset needs to maintain as much information about feature attributes as possible, in order to derive accurate classification models or achieve high quality clustering. As a result, the discernibility of feature attributes needs to be maximized in order to increase data utility.
- *Homogeneity of target attributes.* For classification tasks, an additional criterion is to produce homogeneous partitions or equivalence classes of class labels. The few works specializing on optimizing anonymization for classification applications [21, 50, 17, 31] are mainly focused on this objective. However, it is important to note that if the class label is a sensitive attribute, this criterion is conflicting with the goal of l-diversity and other principles that attempts to achieve a guaranteed level of diversity in sensitive attributes and the question certainly warrants further investigation to achieve best tradeoff.

4.2 Attribute Priorities Based on the above discussion and considering the variety of applications, the first idea we explored is to represent the application requirements using a list of attribute and weight pairs where each attribute is associated with a priority weight based on how important it is to the target applications. We envision that these priority weights can be either explicitly specified by users or implicitly learned by the system based on a set of sample queries and analysis. If the target applications can be fully specified by the users with feature attributes, target attributes, or selection attributes, they can be assigned a higher weight than other attributes in the quasi-identifer set. For instance, in Scenario 1, the attribute-weight list can be represented as $(Age, 0)$, $(Gender, 0)$, $(Zipcode, 1)$ where *Zipcode* is the feature attribute for the location-based study.

Alternatively, the attribute priorities can be learned implicitly from sample queries and analysis. For example, statistics can be collected from query loads on attribute frequencies for projection and selection. In many cases, the attributes in the SELECT clause (projection) correspond to feature attributes while attributes in the WHERE clause (selection) correspond to the selection attributes. The more frequently an attribute is queried, the more important it is to the application, and the less it should be generalized. Attributes can be then ordered by their frequencies where the weight is a normalized frequency. Another interesting idea is to use a min-term predicate set derived from query load and use that in the anonymization process similar to the data fragmentation techniques in distributed databases. This is on our future research agenda.

4.3 Anonymization Metric Before we can devise algorithms to optimize the solution for the application, we first need to define the optimization objective or the cost function. When the query and analysis semantics are known, a suitable metric for the subpopulation identification process is the *Precision* of the relevant subpopulation similar to the precision of relevant documents in Information Retrieval. Note that a generalized dataset will often produce a larger result set than the original table does with respect to a set of predicates consisting of quasi-identifiers. This is similar to the imprecision metric defined in [31]. For analysis tasks, appropriate metrics for specific analysis tasks should be used as the ultimate optimization goal. This includes accuracy for classification applications and intra-cluster similarity and inter-cluster dissimilarity for clustering applications. The majority metric [25] is a class-aware metric introduced to optimize a dataset for classification applications.

When the query and analysis semantics are not specified, we need a general metric that measures the data utility. Intuitively, the anonymization process should generalize the original data as little as is necessary to satisfy the given privacy principle. There are mainly three cost metrics that have been used in the literature [38], namely, general loss metric, majority metric, and discernibility metric. Among the three, the *discernibility metric*, denoted by C_{DM}, is most commonly used and is defined based on the size of equivalence classes E:

$$C_{DM} = \sum_m |E^m|^2 \tag{4.1}$$

To facilitate the application-oriented anonymization, we devise a prioritized cost metric that allows users to incorporate attribute priorities in order to achieve more granularity for more important attributes. Given a quasi-identifier X_i, let $|E^m_{X_i}|$ denote the size of the mth equivalent class with respect to X_i, let $weight_i$ denote attribute priority associated with attribute X_i, the metric is defined as follows:

$$C_{WDM} = \sum_i weight_i * \sum_m |E^m_{X_i}|^2 \tag{4.2}$$

Consider our example Scenario 1, if given an anonymized dataset such as in Table 1, the discernibility of equivalent classes along attribute *Zipcode* will be penalized more than the other two attributes because of the importance of geographic location. This metric corresponds well with our weighted attributed list representation of the application requirements. It provides a general judgement of the anonymization for exploratory analysis when there is some knowledge about attribute importance in applications but not sufficient knowledge about specific subpopulation or applications.

4.4 Anonymization A large number of algorithms have been developed for privacy preserving data anonymization. They can be roughly classified into top-down and bottom-up approaches and single dimensional and multidimensional approaches. Most of the techniques take a greedy approach and rely on certain heuristics at each step or iteration for selecting an attribute for partitioning (top-down) or generalization (bottom-up). In this study, we adapt the greedy top-down Mondrian multidimensional approach [30] and investigate heuristics for adapting it based on our prioritized optimization metric. It is on our future research agenda to explore various anonymization approaches and investigate systematic ways for adapting them towards application-oriented anonymization.

The Mondrian algorithm (based on k-anonymity principle) uses greedy recursive partitioning of the (multidimensional) quasi-identifer domain space. In order to obtain approximately uniform partition occupancy, it recursively chooses the split attribute with the largest normalized range of values, referred to as *spread*, and (for continuous or ordinal attributes) partitions the data around the median value of the split attribute. This process is repeated until no allowable split remains, meaning that a particular region cannot be further divided without violating the anonymity constraint, or constraints imposed by value generalization hierarchies.

The key of the algorithm is to select the best attribute for splitting (partitioning) during each iteration. In addition to using the spread (range) of the values of each attribute i, denoted as $spread_i$, in the original algorithm, our approach explores additional metrics.

Attribute priority. Since our main generalization criteria is to maximize the discernibility of important attributes including selection attributes, feature attributes and class attributes for target applications, we use the attribute priority weight for attribute i, denoted by $weight_i$, as an important selection criteria. Attributes with a larger weight will be selected for partitioning so that important attributes will have a more

precise view in the anonymized data.

Information gain. When target applications are well specified a priori, another important generalization criterion for classification applications is to maximize the homogeneity of class attributes within each equivalence class. This is reminiscent of decision tree construction where each path of the decision tree leads to a homogeneous group of class labels [18]. Similarly, information gain can be used as a scoring metric for selecting the best attribute for partitioning in order to produce equivalence classes of homogeneous class labels. The information gain for a given attribute i, denoted by $infogain_i$, is computed as the weighted entropy of the resultant partitions based on the split of attribute i:

$$infogain_i = \sum_{P'} \left(\frac{|P'|}{|P|} \sum_{c \in D_c} -p(c|P')logp(c|P')\right) \quad (4.3)$$

where P denotes the current partition, P' denotes the set of resultant partitions of the iteration, $p(c|P')$ is the fraction of tuples in P' with class label c, and D_c is the domain of the class variable c.

The attribute selection criteria for each iteration selects the best attribute based on an overall scoring metric determined by an aggregation of the above metrics. In this study, we use a linear combination of the individual metrics, denoted by O_i for attribute i:

$$O_i = \sum_j (w_j * metric_i^j) / \sum_j w_j \quad (4.4)$$

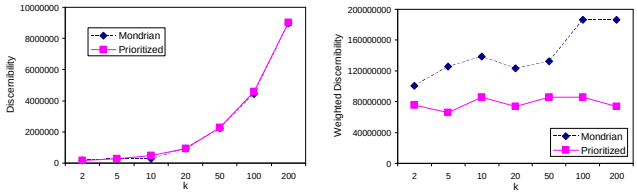
where $metric_i^j \in \{spread_i, infogain_i, weight_i\}$, and w_j is the weight of the individual metric j ($w_j >= 0$).

5 Experiments

We performed a set of preliminary experiments evaluating our approach. The main questions we would like to answer are: 1) does the prioritized anonymization metric (weighted discernibility metric) correlate with good data utility from applications point of view? 2) does the prioritized anonymization scheme provide better data utility than general approaches?

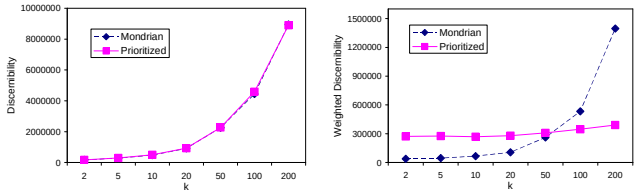
We implemented a prioritized anonymization algorithm based on the Mondrian algorithm [30]. It uses a combined heuristic of the spread and attribute priorities (without information gain) and aims to minimize the prioritized cost metric (instead of the general discernibility metric). We conducted two sets of experiments for exploratory and classification applications respectively.

5.1 Exploratory Applications For exploratory applications, we used the Adults dataset from UC Irvine Machine Learning Repository configured as in [30]. We considered a simple application scenario that requires precise information on a single demographic attribute (*Age* and *Sex* respectively) and hence it is assigned with a higher weight than other attributes in the experiment. The dataset were anonymized using the Mondrian and prioritized approach respectively and we compare the weighted discernibility as well as general discernibility of the two anonymized datasets.



Discernibility Weighted Discernibility

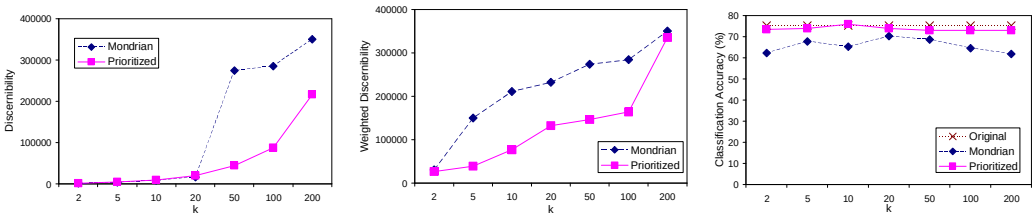
Figure 1: Adult Dataset (*Sex*-Prioritized)



Discernibility Weighted Discernibility

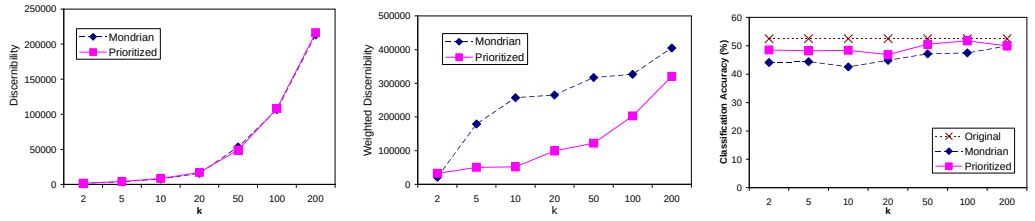
Figure 2: Adult Dataset (*Age*-Prioritized)

Figure 1 and 2 compare the prioritized approach and the Mondrian approach in terms of general discernibility and weighted discernibility with respect to different value of k for *Sex*-prioritized and *Age*-prioritized anonymization respectively. We observe that even though the prioritized approach has a comparable general discernibility with the Mondrian, it achieves a much improved weighted discernibility in both cases, which is directly correlated with the user-desired data utility (i.e. having a more fine-grained view for *Age* attribute or *Sex* attribute for exploratory query or mining purposes).



(a) Discernibility (b) Weighted Discernibility (c) Classification Accuracy

Figure 3: Japanese Credit Screening Dataset - Classification



(a) Discernibility (b) Weighted Discernibility (c) Prediction Accuracy

Figure 4: Japanese Credit Screening Dataset - Prediction (A3)

5.2 Classification Applications For classification applications, we used the Japanese Credit Screening dataset, also from the UCI Machine Learning Repository. The dataset consists of 653 instances, 15 attributes and a 2-valued class attribute ($A16$) that corresponds to a positive/negative (+/-) credit. The missing valued instances were removed and the experiments were carried out considering only the continuous attributes ($A2, A3, A8, A11, A14$ and $A15$). The dataset was anonymized using the prioritized approach and the Mondrian approach and the resultant anonymized data as well as the original data were used for classification and prediction. The Weka implementation of the simple Naive-Bayes classifier was used for the classification, with 10 fold cross-validation for classification accuracy determination.

For classification, the class attribute was recoded as 1.0/0.0. Different feature attributes were selected and given varying weights (both arbitrary or assuming user knowledge) to examine their effect on classification accuracy. For prediction, attributes other than the class attribute were recoded into ranges using equi-width[3] approach. A target attribute is selected as the prediction attribute and the rest of the attributes are anonymized and used to predict the target attribute.

We assume the users have some domain knowledge of which attributes will be used as feature attributes for their classification and we then assigned higher priority weights for these attributes. In addition, we also experimented with a set of single-attribute classification by selecting one feature attribute each time and assigned weights for the attributes based on their classification accuracy. The results are similar and we report the first set of results below.

Figure 3(a) and 3(b) compare the prioritized and Mondrian approach in terms of general discernibility and weighted discernibility of the anonymized dataset respectively. Figure 3(c) compares the anonymized datasets as well as the original dataset in terms of accuracy for the class attribute. Similarly, Figure 4 presents the results for prediction of attribute $A3$. We observe that the prioritized approach performs better than the Mondrian for both classification and prediction in terms of accuracy and achieves a comparable accuracy as the original dataset. In addition, a comparison of the discernibility metrics and the classification accuracy shows that the weighted discernibility metric corresponds well to the application-oriented data utility, i.e. the classification accuracy.

6 Conclusion and Discussions

We presented an application-oriented approach for data anonymization that takes into account the relative attribute importance for target applications. We derived

[3]Equal spread ranges for the recoded attributes.

a set of generalization criteria for application-oriented data anonymization and presented a prioritized generalization approach that aims to minimize the prioritized cost metric. Our initial results show that the prioritized anonymization metric correlates well with application-oriented data utility and the prioritized approach achieves better data utility than general approaches from application point of view.

There are a few items on our research agenda. First, the presented anonymization technique uses a special generalization algorithm and a simple weighted heuristic. We will study different heuristics and generalize the result to more advanced privacy principles and anonymization approaches. Second, while it is not always possible for users to specify the attribute priorities before hand, we will study how to automatically learn attribute priorities from sample queries and mining tasks and further devise models and presentations that allow application requirements to be incorporated. In addition, a more in-depth and longer-term issue that we will investigate is the notion of priorities, in particular, the interaction between what data owners perceive and what the data users (applications) perceive. Finally, it is important to note that there are inference implications of releasing multiple anonymized views where multiple data users may collude and combine their views to breach data privacy. While there is work beginning investigating the inference problem [57], the direction certainly warrants further research.

References

[1] N. R. Adams and J. C. Wortman. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 21(4), 1989.

[2] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[3] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In *PODS*, pages 153–162, 2006.

[4] G. Aggarwal, N. Mishra, and B. Pinkas. Secure computation of the kth ranked element. In *IACR Conference on Eurocrypt*, 2004.

[5] R. Agrawal, P. Bird, T. Grandison, J. Kieman, S. Logan, and W. Rjaibi. Extending relational database systems to automatically enforce privacy policies. In *21st ICDE*, 2005.

[6] R. Agrawal, A. Evfimievski, and R. Srikant. Information sharing across private databases. In *SIGMOD*, 2003.

[7] R. Agrawal, J. Kieman, R. Srikant, and Y. Xu. Hippocratic databases. In *VLDB*, 2002.

[8] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 439–450. ACM Press, May 2000.

[9] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE '05: Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

[10] E. Bertino, B. Ooi, Y. Yang, and R. H. Deng. Privacy and ownership preserving of outsourced medical data. In *ICDE*, 2005.

[11] S. S. Bhowmick, L. Gruenwald, M. Iwaihara, and S. Chatvichienchai. Private-iye: A framework for privacy preserving data integration. In *ICDE Workshops*, page 91, 2006.

[12] S. Bu, L. V. S. Lakshmanan, R. T. Ng, and G. Ramesh. Preservation of patterns and input-output privacy. In *ICDE*, pages 696–705, 2007.

[13] J. Byun and E. Bertino. Micro-views, or on how to protect privacy while enhancing data usability - concept and challenges. *SIGMOD Record*, 35(1), 2006.

[14] J.-W. Byun, E. Bertino, and N. Li. Purpose based access control of complex data for privacy protection. In *ACM Symposium on Access Control Models and Technologies (SACMAT)*, 2005.

[15] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222, 2003.

[16] A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. *Inf. Syst.*, 29(4):343–364, 2004.

[17] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *Proc. of the 21st IEEE International Conference on Data Engineering (ICDE 2005)*, pages 205–216, Tokyo, Japan, April 2005.

[18] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition, 2006.

[19] Z. Huang, W. Du, and B. Chen. Deriving private information from randomized data. In *SIGMOD Conference*, pages 37–48, 2005.

[20] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.

[21] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *SIGKDD*, 2002.

[22] S. Jajodia and R. Sandhu. Toward a multilevel secure relational data model. In *ACM SIGMOD*, 1991.

[23] M. Kantarcioglu and C. Clifton. Privacy preserving data mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(9), 2004.

[24] M. Kantarcioglu and C. Clifton. Privacy preserving k-nn classifier. In *ICDE*, 2005.

[25] M. Kantarcoglu and J. Vaidya. Privacy preserving naive bayes classifier for horizontally partitioned data.

In *IEEE ICDM Workshop on Privacy Preserving Data Mining*, 2003.

[26] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *ICDM*, pages 99–106, 2003.

[27] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD Conference*, pages 217–228, 2006.

[28] K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, and D. DeWitt. Limiting disclosure in hippocratic databases. In *30th International Conference on Very Large Data Bases*, 2004.

[29] K. LeFevre, D. Dewitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *ACM SIGMOD International Conference on Management of Data*, 2005.

[30] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *IEEE ICDE*, 2006.

[31] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *SIGKDD*, 2006.

[32] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *To appear in International Conference on Data Engineering (ICDE)*, 2007.

[33] Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of Cryptology*, 15(3), 2002.

[34] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, page 24, 2006.

[35] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE*, pages 126–135, 2007.

[36] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, pages 223–228, 2004.

[37] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD Conference*, pages 665–676, 2007.

[38] M. E. Nergiz and C. Clifton. Thoughts on k-anonymization. In *ICDE Workshops*, page 96, 2006.

[39] S. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *VLDB*, pages 682–693, 2002.

[40] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[41] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, fuzziness and knowledge-based systems*, 10(5), 2002.

[42] Z. Teng and W. Du. Comparisons of k-anonymization and randomization schemes under linking attacks. In *ICDM*, pages 1091–1096, 2006.

[43] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *ICDE Workshops*, page 94, 2006.

[44] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *ACM SIGKDD*, 2002.

[45] J. vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *SIGKDD*, 2003.

[46] J. Vaidya and C. Clifton. Privacy preserving nave bayes classifier for vertically partitioned data. In *ACM SIGKDD*, 2003.

[47] J. Vaidya and C. Clifton. Privacy-preserving top-k queries. In *ICDE*, 2005.

[48] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, 33(1), 2004.

[49] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *ACM SIGKDD*, 2006.

[50] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: a data mining solution to privacy protection. In *Proc. of the 4th IEEE International Conference on Data Mining (ICDM 2004)*, November 2004.

[51] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150, 2006.

[52] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, 2006.

[53] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *SIGMOD Conference*, pages 689–700, 2007.

[54] L. Xiong, S. Chitti, and L. Liu. Topk queries across multiple private databases. In *25th International Conference on Distributed Computing Systems (ICDCS 2005)*, 2005.

[55] L. Xiong, S. Chitti, and L. Liu. Mining multiple private databases using a knn classifier. In *ACM Symposium of Applied Computing (SAC)*, pages 435–440, 2007.

[56] Z. Yang, S. Zhong, and R. N. Wright. Privacy-preserving classification of customer data without loss of accuracy. In *SIAM SDM*, 2005.

[57] C. Yao, X. S. Wang, and S. Jajodia. Checking for k-anonymity violation by views. In *VLDB*, 2005.

[58] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *ICDE*, pages 116–125, 2007.

[59] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing k-anonymization of customer data. In *PODS*, 2005.