Technical Report

TR-2007-028

Weighted approach to application oriented data anonymization

by

Li Xiong, K. Rangachari

Mathematics and Computer Science

EMORY UNIVERSITY

Weighted Approach to Application Oriented Data Anonymization

Dr. Li Xiong Kumudha Rangachari
Dept. of Math & Computer Science
Emory University

ABSTRACT

This paper summarizes the effort to observe the effectiveness of a weighted approach to data anonymization keeping in mind the applications which the anonymized data would cater to, thereby increasing its usefulness to the application.

Categories and Subject Descriptors

H.2.7 [Database Administration]: *Security, integrity, and protection*

H.2.8 [Database Applications]: *Data mining*

General Terms

Identity protection, k-anonymization, Classification, Query-load, Performance

Keywords
k-anonymization, Weighted k-anonymization, adaptive anonymization

1. Introduction

1.1 The Anonymity Problem

Data privacy and identity protection is a very important issue in this day and age when huge databases containing a population's information can be stored and distributed for research or other purposes. However, such data sharing has been stymied by restrictions and concerns about the privacy of individuals. The U.S. Government Accounting Office (GAO) issued a homeland security study that found that the poor information sharing efforts might cause critical clues of impending terrorist attacks to go unnoticed [20]. Government access to privately held data with personal information remains a vexing problem. In the healthcare domain, the United States President's Information Technology Advisory Committee (PITAC) released a report in June 2004 entitled "Revolutionizing Health Care through Information Technology" and one of the key challenges identified is to ensure security, privacy, and interoperability for information sharing. An example initiative is the Shared Pathology Informatics Network (SPIN) initiated by The National Cancer Institute initiated 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). It is necessary for each institution to de-identify or anonymize the data before having it accessible by 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 to a shared network or individual institutions and researchers (*data recipients*) that does not contain individually identifiable information.

1.2 The k-anonymity model

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 [49, 40, 53, 5, 38, 62, 41, 43]. Notably, the earliest principle, k-anonymity [49], requires that a set of k records (entities) to be indistinguishable from each other based on a quasi-identifier set. Later principles remedy the problems of k-anonymity, such as l-diversity [40] and t-closeness [38] which requires the distribution of sensitive values in each group to be analogous to the distribution and m-invariance [62] that protects data re-publishing. 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 quality metrics [25, 60, 42, 12, 3, 19, 13, 70, 35, 36, 37, 59, 32, 61, 69].

1.3 Issues with general discernibilty & optimization:

Though these methods to anonymize data are widely gaining popularity and performing better, their aim remains making the data optimally anonymized as is possible, measured through discernibilty (Section 3) and information loss. Each target application may have a unique need of the data. Few works have considered targeted applications like classification and regression [24, 60, 19, 37] but do not model other kinds of applications nor provide a systematic or adaptive approach for handling various needs. We aim to better the existing methods by incorporating the

application for which the data will be used into the anonymization process, thereby increasing its utility to the target application.

1.4 Adaptive Anonymization metrics & Techniques:

Through this paper we propose that the best way of measuring data utility is based on the analysis task for which the data will ultimately be used. In consultation with domain experts, we set out to take a top-down analysis of various potential applications and devise models and schemes to represent important application requirements and develop techniques to incorporate that knowledge in the anonymization. The resulting techniques, we show, will allow data providers to incorporate a set of requirements specified by users or learned from sample queries and analysis in the anonymization algorithms to produce an optimized view for the users.

2. Related Work

The proposed research is inspired and informed by a number of existing ideas. We discuss the relationship of the proposed research with the state of knowledge in the field.

Privacy Preserving Access Control. Previous work on multilevel secure relational databases [26] provides many valuable insights for designing a fine-grained secure data model. In a multilevel relational database, each piece of information is classified into a security level, and every user is assigned a security clearance. Based on the access class, the system ensures that each user gains access to only the data for which she has proper clearance with no information flow from a higher security level to a lower security level. Hippocratic databases [8, 34, 6] 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 [27]. Purpose information associated with a given data element specifies the intended use of the data element. 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, we need to go beyond this level. In order to maximize the potential data utility, micro-views of the data are desired where even a single value of a tuple attribute may have different views [16].

Statistical Databases. 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 approaches can be broadly classified into data perturbation, and query restriction. Data perturbation involves either altering the input databases, or altering query results returned. Query restriction includes schemes that check for possible privacy breaches by keeping audit trails and controlling overlap of successive aggregate queries. The techniques developed have focused only on aggregate queries and relational data types. In addition, the inference implications of releasing one or more analysis results on the original data are not well-understood.

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 un-trusted service provider. The main approach is random perturbation that transforms data by adding random noise in a principled way [9, 58]. There are studies on specific mining tasks such as decision tree [9, 15], association rule mining [46, 17, 18] and disclosure analysis [31, 23, 51].

Distributed Privacy Preserving Data Sharing. 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 [39, 54, 28, 30, 56, 68, 55, 4, 7, 57, 29] including several that the PI has developed recently [63, 64, 67, 65]. 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. There are also recent works towards privacy preserving data integration [14]. Privacy preserving data publishing, as discussed below, deals with a different client/server setting (data provider/data recipient). In addition, a main advantage of generalization based anonymization as opposed to data perturbation is that the released data remain ”truthful”, though at a coarse level of granularity. This allows various analyses to be carried out using the data, including selection.

Privacy Preserving Data Publishing. The literature on centralized privacy preserving data publishing that provides a micro-view of the data while preserving privacy of individuals 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*[62]. Notably, the earliest principle, k-anonymity [49], requires that a set of k records (entities) to be indistinguishable from each other based on a quasi-identifier set. Later principles remedy the problems of k-anonymity, such as l-diversity [40] and t-closeness [38] which requires the distribution of sensitive values in each group to be analogous to the distribution and m-invariance [62] that protects data re-publishing.

Application-based Anonymization: The literature provides a suite of anonymization algorithms that produce an anonymous view based on a target class of workloads, consisting of one or more data mining tasks, like

classification, regression and clustering as well as selection predicates [37, 68, 70, 64]

3. Preliminaries:

We first introduce some terminology to facilitate our discussion and formalize the k-anonymity [47] and l-diversity [40] principles that illustrate the ideas behind privacy preserving data publishing. 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 (X1,..., Xd) that can be joined with external information to re-identify individual records. We assume that a quasi-identifier is recognized based on the 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 an identifier, (Age, Gender, Zipcode) is considered a quasi-identifier set, and Diagnosis is considered 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 setX1, ..., Xd is referred to as *Equivalence class E*. T is k-*anonymous* with respect to X1, ..., Xd 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 3 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.

3.1 The Mondrian Multi-dimensional Method

The Mondrian algorithm (Fig 1) proposed for k-anonymization using multidimensional recoding [36]. The algorithm is based on a greedy recursive partitioning of the (multidimensional) quasi-identifier domain space. In order to obtain approximately uniform partition occupancy, the paper suggests recursively choosing the split attribute with the largest normalized range of values, and (for continuous or ordinal attributes) partitioning 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. This algorithm is referred to as the Median Mondrian and forms the basis for the proposed exploratory application adaptive anonymization.

```
Anonymize(tuples, attrs)
  if (no allowable split for tuples)
    return φ : t ∈ tuples → bounding region(tuples)
  else
    best ← Choose_Attribute(attrs, tuples)
    if continuous(best) or ordinal(best)
      threshold ← Choose_Threshold(best)
      lhs ← {t ∈ tuples : t.best ≤ threshold}
      rhs ← {t ∈ tuples : t.best > threshold}
      return Anonymize(rhs,attrs) ∪ Anonymize(lhs,attrs)
    else if nominal(best)
      recodings ← {}
      for each child v_i of root(best.hierarchy)
        tuples_i ← {t ∈ tuples : t.best ≼ v_i}
        attrs' ← replace root(best.hierarchy) with v_i in attrs
        recodings ← recodings ∪ Anonymize(tuples_i, attrs')
      return recodings
```

Fig. 1 : Mondrian Multidimensional algorithm

3.2 Performance Evaluation Metrics for General Anonymization:

There are many metrics that determine the quality in microdata released by assigning a penalty based on the amount of perturbation that the original data suffered due to the anonymization. The simplest kind of quality measure is based on the size of the equivalence classes E in T.

Intuitively, the discernability metric (C_{DM}), assigns to each tuple t in T a penalty, which is determined by the size of the
equivalence class containing t.

$$C_{DM} = \sum_{EquivClasses\ E} |E|^2$$

An alternative, is the normalized average equivalence class size metric (C_{AVG}).

$$C_{AVG} = (\frac{total\_records}{total\_equiv\_classes})/(k)$$

We propose a new metric for the adaptive anonymization techniques (defined in Section 4.2) that portrays the granularity of the feature attributes important to the target application.

4. The Adaptive Approach

The first goal of the approach is to gather important application requirements. It assumes structured (relational) data, and will involve adapting general anonymization techniques to several application scenarios. 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 begin by considering two example classes of applications.

• *Application 1. Disease-specific public health study:* In this study, researchers select a subpopulation using a selection predicate 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.

• *Application 2. Demographic / population study*. In this study, researchers may want to study a certain demographic population, such as males over 50, and learn classification models based on demographic information and clinical symptoms to predict diagnosis.

4.1 Application Requirements

The data analysis for the mentioned applications is typically conducted in two steps:

1) *Subpopulation Identification* through a selection predicate

2) *Analysis* on the identified subpopulation using selected features.

The analysis could include clustering of the population or classification of the population with respect to certain class labels. Given such a two-step process, we have two requirements for optimizing the anonymization for applications, namely, *maximize precision and recall of subpopulation identification* and *maximize quality of the analysis.*

We first categorize the attributes with respect to the analysis tasks on the anonymized data and then explain how the application requirement and optimization goal transforms to concrete criteria for adaptive anonymization. Given an anonymized relational table Ta, each is characterized by one of the following types:

• *Selection attributes* are those attributes used to identify a subpopulation (e.g. Diagnosis in Application 1 and Age in Application 2).
• *Feature attributes* are those attributes used to cluster data or to classify data with respect to a target class (e.g. Location in Application 1).
• *Target attributes* are those for which the classification or prediction is trying to predict the class label or attribute value (e.g. Diagnosis in Application 2). Target attributes are not applicable for unsupervised learning tasks such as clustering.

We envision that the application requirements can be either explicitly specified by users or implicitly learned by the system based on a set of sample queries and analysis. For the first approach, we plan to work with domain experts to devise intuitive and expressive representations that allow medical and outcomes researchers to efficiently and effectively specify their needs or constraints in demographic study needs. One possibility is to have researchers specify a list of feature attributes, target attributes, as well as selection attributes or predicates if the targeted applications can be fully specified. If the targeted applications are unknown or the analysis is rather exploratory, a more generalized form could be an *ordered list of attribute and weight pairs* where each attribute is associated with a priority weight and the attributes are sorted in a descending order of priority. In cases when feature attributes or selection attributes are known, they can be assigned a higher weight than other attributes in the quasi-identifer set. In Application 1, the priority list can be represented as (Age, 0), (Gender, 0), (Zipcode, 1) where Zipcode is the most important while Age and Gender are lesser important but equal to each other.

Alternatively, the constraints 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 [45].

4.2 The Weighted Discernibility 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 [11].

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 [37]. 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, let's consider the commonly used *discernibility metric* CDM based on the size of equivalence classes E:

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

Clearly it is not sufficient in capturing the application-specific quality. We propose a novel application-driven but application-independent metric *Weighted Discernibility Metric*. It extends the general discernibility metric by measuring the discernibility of data with respect to different attributes where generalization along important attributes such as selection attributes and feature attributes will be penalized more. Given a quasi-identifier Xi, let $|E^m_{X_i}|$ denote the size of the mth equivalent class with respect to Xi, let wi denote a penalty weight associated with attribute Xi, the metric is defined as follows:

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

In Application 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.

4.3 Optimization Techniques:

There are four target applications identified on which the adaptive anonymization algorithms will be based and tested. They are:

- Exploratory analysis applications
- Query applications
- Classification applications
- Clustering applications

And further the algorithms would try and learn any association rules that follow from adapting the database to the application.

A large number of algorithms have been developed for privacy preserving data publishing. 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). Based on our discussion on generalization criteria for different types of attributes, we will explore a number of advanced anonymization approaches and investigate heuristics for adapting them towards the application-driven optimization metrics. For example, the implementation for the experiments uses the greedy top-down Mondrian Multidimensional partitioning approach, and in any iteration, a combination of the following heuristics was used for attribute selection:

• *Information gain of the attribute*. Similar to decision tree classifier construction [22], information gain can be used as a scoring metric for selecting an attribute for generalization or specialization for the iteration in greedy approaches [36].
• *Precision of selection*. Precision of selection can be used as another metric to optimize the subpopulation identification. We will also investigate the possibility of using selection predicates directly for determining the splitting point in a top-down strategy.
• *Attribute weight*. Attributes can be selected based on their weight so that important attributes will have a more precise view in the anonymized data. One important complication to address here is that a strict enforcement of the attributed weights may not even be feasible given a privacy principle. So both strict and relaxed enforcement of the attribute weight will be studied.

5. Experiments

The experiment outline for the ideas stated in this paper consists of heuristic exercises in determining weights of feature attributes and testing the performance of the method in light of the heuristics, thereby refining the heuristic and iteratively accruing performance.

The initial experiments that have been performed so far follow minimal heuristic exploration for the weights and also use a small dataset to test the viability of the methods. For the initial experiments the scope of the applications considered was restricted to classification alone, though a number of experiments pertaining to exploratory analysis were also performed. The set-up and results obtained from these experiments are furnished below.

5.1 General Experimental Setup

The experiment set-up for each of the target applications chosen are described in this section:

(a) Exploratory analysis application:

The technique assumes, the weights associated with each of the attributes, is learnt either through domain experts or through consultations with the users of the data. The optimization heuristic used during the attribute selection of the anonymization method would be the attribute weight in combination with the underlying basis of the anonymization method (like spread or info-gain). In terms of metrics, the Weighted Discernibilty Metric (WDM) and the General Discernibility Metric (GDM) would be used for attribute selection.

To assess the performance of the anonymized data we use the same metric, as the target application is unknown. Any number of combinations of the prioritized weights and the selection criteria are sampled and the sample that produces the best metric values is the released view. Further the algorithm tries to learn any association rules that might be deduced from the analysis phase.

(b) Query applications:

In the case of querying applications, the technique involves studying the past query loads, if available, and consulting with application and domain experts that request the data, to determine the selection attributes and their weights. The selection of attributes during anonymization is based on attribute frequency derived from query loads which translate into the WDM, and the query precision metric which will be determined by subjecting the resultant anonymized table of the iteration to the sample query load.

The performance assessor for this technique would be the query precision metric. Note that the technique degenerates to the attribute weight heuristic-based technique when the query loads are not available.

(c) Classification applications:

The technique for classification applications expects knowledge of the feature attributes, the target classes of the classification, and the weights assigned to the attributes based on the knowledge. The attribute section in this case uses the WDM obtained considering the feature attributes alone and the homogeneity of the equivalence classes on the class attributes and the general discernability metric. The GDM helps in keeping the l-diversity in check which could be destabilized by the homogeneity requirement of the technique.

The classification accuracy assesses the performance of the anonymization algorithm and the anonymized data. This metric could also be a selection factor.

(d) Clustering applications:

The clustering applications require the technique to make use of feature attributes and their associated weights for attribute selection. The two other factors that determine attribute selection are the intra-cluster homogeneity and the inter-cluster heterogeneity of the resultant anonymized equivalence classes.

The cluster homogeneity and heterogeneity also act as performance measures for the technique.

(e) Association rules:

The techniques mentioned above could develop a machine-learning flavor by learning association rules that are hidden within the anonymized data. This might help other applications for which the view was not initially intended.

5.2 Results

The preliminary study based on the attribute weight heuristic involved the implementation of an adaptive version of the Mondrian algorithm (Section 4.2)[36] that takes into account user specified attribute weights in the attribute selection. We used the Adults dataset, which is the most common for evaluating anonymization methods. from UC Irvine Machine Learning Repository configured as in [36], for exploratory analysis.

The exploratory experiments were useful in determining which attributes when weighted would lower the weighted discernibility metric (WDM). At first the attributes were weighted one at a time, and the resulting anonymized dataset of the method at various k's were recorded. Of all the attributes in the Adult dataset considered, we observed that the Age and Sex attributes had the best values for the WDM as shown in Fig. X-1 and Fig. X-2 (below). The Sex attribute though, can never be used as a feature attribute, for lack of spread.

Fig. X-1:

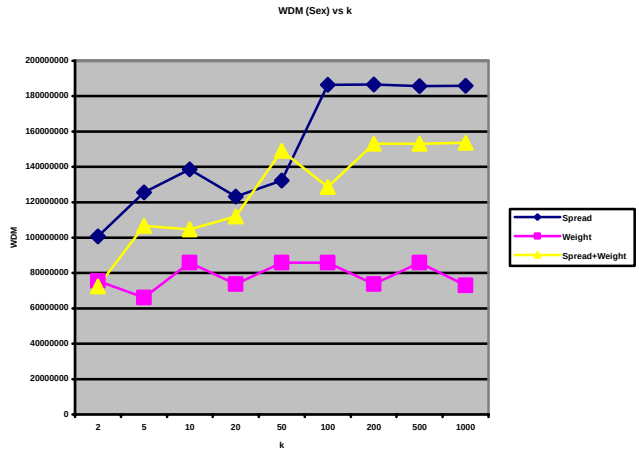
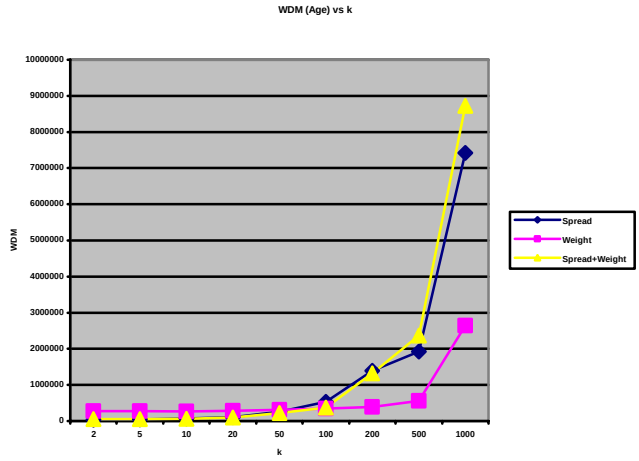
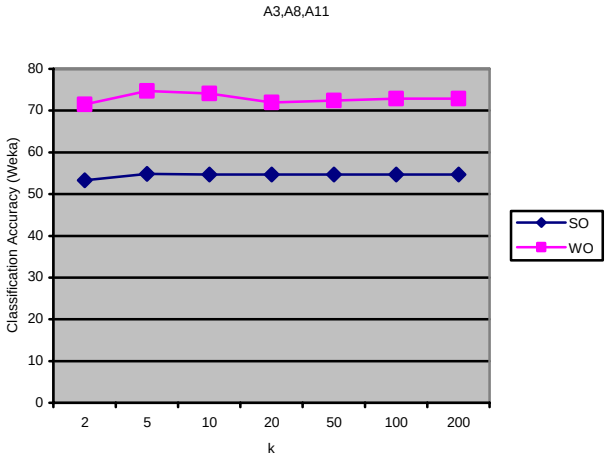


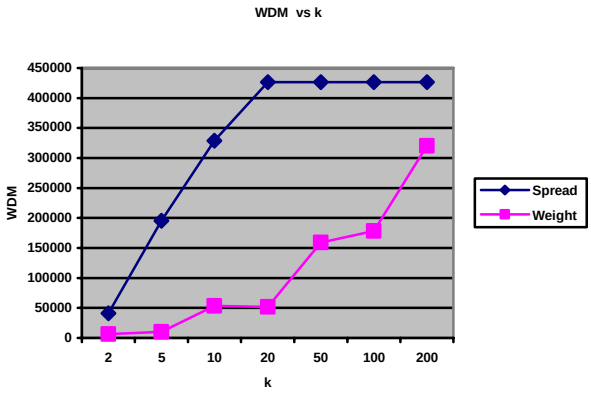
Fig X-2:



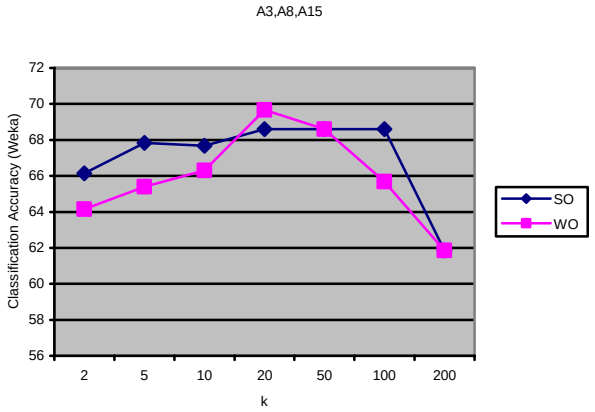
The classification experiments used the Japanese Credit Screening dataset, also from the UCI Machine Learning Repository. The dataset consists of 653 instances, 15 attributes and a 2-valued class attribute (A16) that corresponds to a positive/negative (+/-) credit. The missing valued instances were removed and the experiments were carried out considering only the continuous attributes (A2, A3, A8, A11, A14 and A15). Initially the impact of each of the attributes, when considered a feature attribute, on the classification was not known and so the weights were fixed arbitrarily. The resultant anonymized data was used to predict the classification attribute and other feature attributes as target attributes using Weka. The simple Naïve-Bayes classifier was used, with 10 fold cross-validation for classification accuracy determination. The conducted experiments were under two categories:

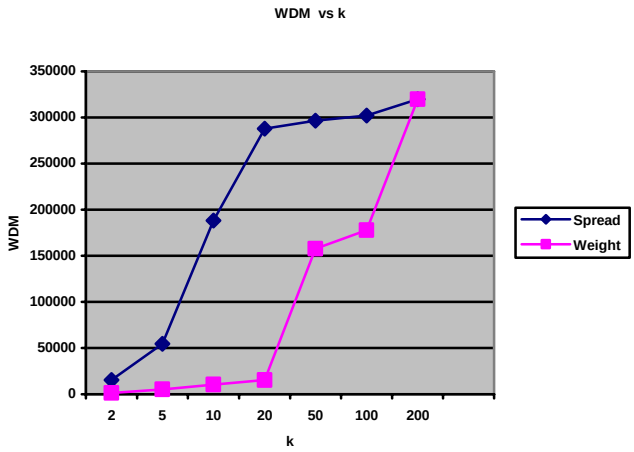
Case I: The class attribute was recoded as 1.0/0.0. The different feature attributes were selected and given varying weights (arbitrary) to examine the change in accuracy levels. The anonymized dataset was augmented with the recoded class attributes before classification. Only the feature attributes were used during classification. Finding the relation between original attribute values and the classification accuracy led to better results in almost all experiments, where those attributes that contributed to higher accuracy during classification were given larger weights during anonymization.





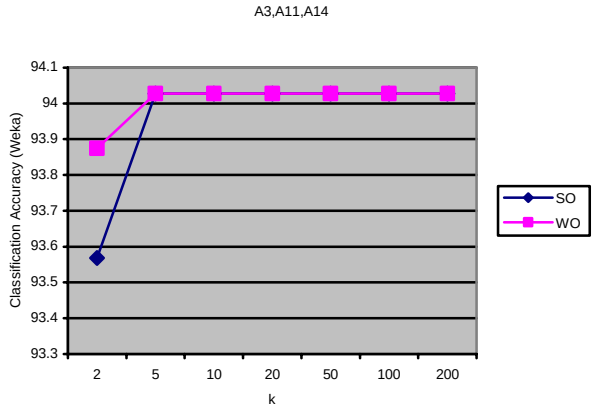
The Fig shows a significant increase in accuracy when the weight only approach was used compared to the spread only Mondrian approach. A3, A8 and A11 were considered feature attributes to classify the class attribute, A16. This Fig. is from the arbitrary equi-weighted approach.





The Fig shows an interesting graph obtained in the equi-weighted approach that shows the weight-only approach performing badly against the spread-only Mondrian.

Case II: The target attributes in this case were any attribute in the dataset, apart from the class attribute. The class attribute is eliminated from the dataset. The target attribute is recoded with equi-width ranges based on its original spread that its original value falls into. After weighted anonymization, the recoded attribute values are augmented with the data. Classification of the target attribute using the feature attributes alone were recorded and the accuracy improved for most attribute classification experiments, when ascending ranked attributes (based on resultant accuracy when using original values to classify target attribute) are assigned descending weights.



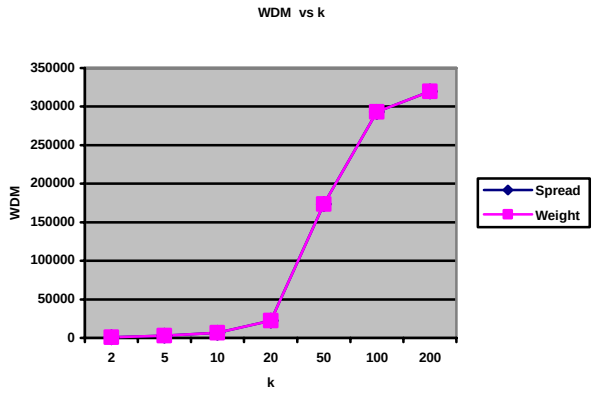
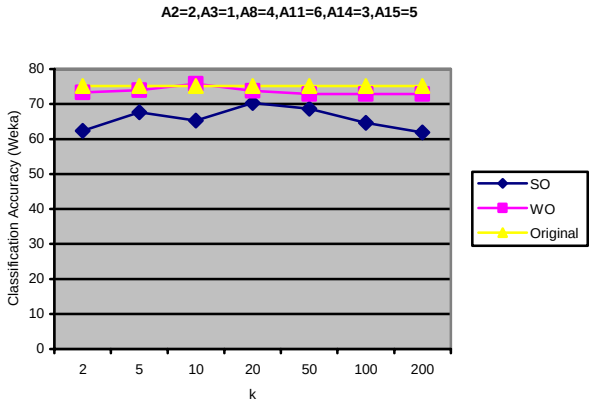


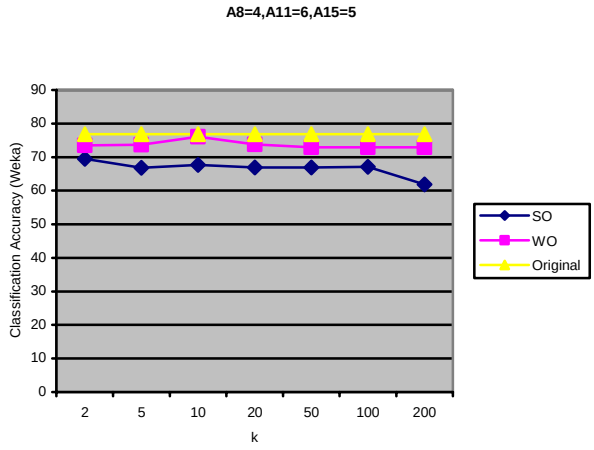
Fig above shows the classification accuracies obtained when target attribute is A8 and A3, A11 and A14 were feature attributes when anonymized using the spread-only and weight-only versions of the Mondrian approach.

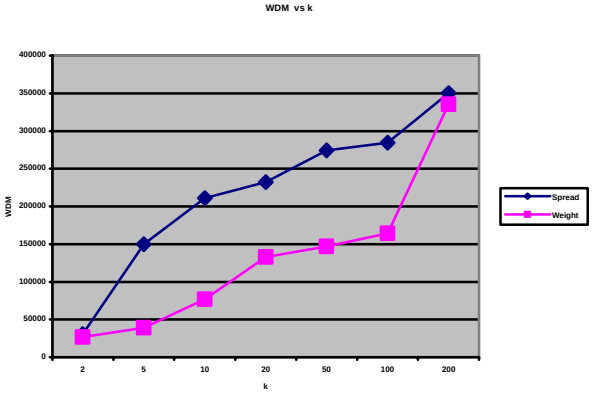
Fig. shows the performance of the weight only approach against the spread-only Mondrian on classifying attribute A2, using A3, A11 and A14 as feature attributes. This plot is interesting as it looks like the performance of each method seemingly varies on the k. But in fact, it is due to the fact that the attributes selected to attain the k on anonymization and their order are different.

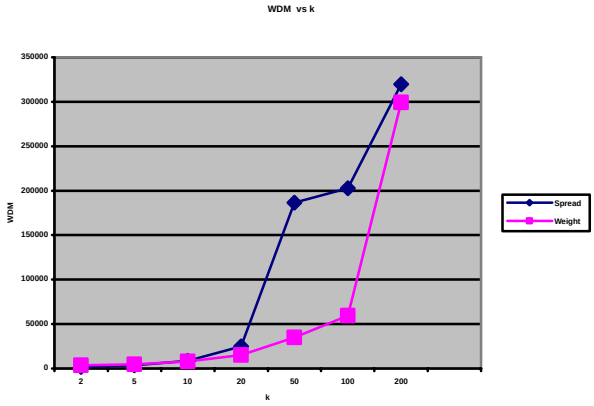
Case III: Other experiments were also performed to determine the subset of attributes that were to be weighted and their weights for the anonymization technique. One method was to use the single-value classification accuracy[1], after recoding the original dataset using the equi-depth idea, to determine the rank of each attribute and weight the top-3 attributes that gave the best accuracy on the target attribute classification in decreasing order. Some of the interesting outcomes of this experiment are as follows:

When the +/- class attribute was the target attribute, and all other attributes were feature attributes, the methods performed as below:

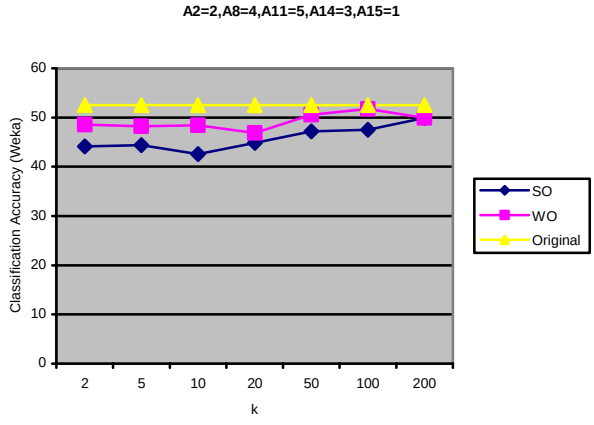


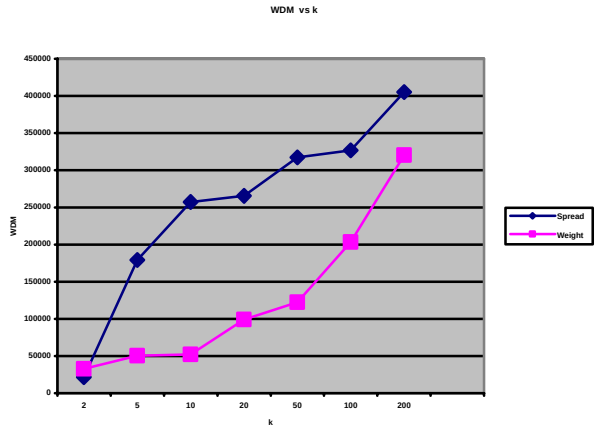




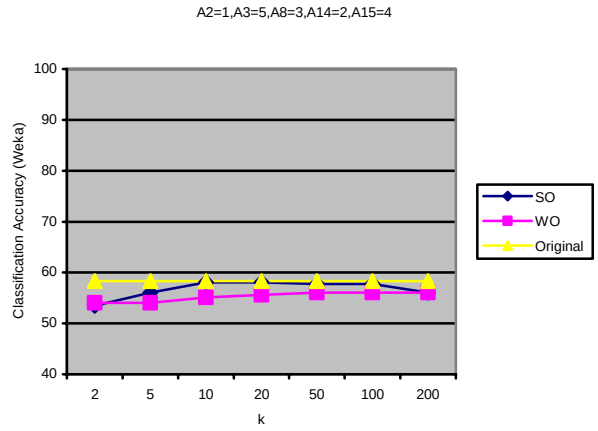


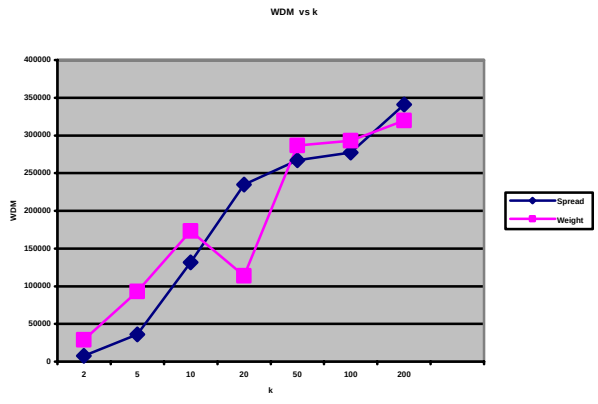
When the attributes that produced larger accuracies for the class attribute on single-attribute classification, the following graph depicts the results, which is as expected; i.e. the weight only performs better than spread only, and is as close to original accuracy as can be.





The graph above shows the performance curve of the weight-only and spread-only against the original recoded accuracies for different *k*s. The target attribute for this run was A3, and the feature attributes were all other attributes apart from the +/- class attribute, weighted based on their single-attribute classification rank. The indication is that weight-only method performs better under classification with almost equal accuracies as the original.





6. Future Work

The possibilities of the ideas mentioned above are vast and inexhaustible in the short period of time that has been devoted to them. But from the results, the existence of and the need for better ways of anonymizing data considering application semantics is apparent. Future work includes fine-tuning optimization techniques for other applications, such as clustering and regression. And proving the hypothesis that better weighted discernability metric values than general discernability metric values mean better performance measures for the target applications. Further, an expansive evaluation for the resultant anonymized data needs to be devised, that compares the different implementations of the optimization techniques viz. the naïve adaptive implementation (as implemented currently), the probabilistic adaptive implementation (as defined in the classification/clustering/querying optimization techniques), the strict adaptive implementation (as the attribute weight heuristic) and the non-adaptive implementation (does not incorporate attribute weights).

7. 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] Eytan Adar. User 4xxxxx9: Anonymizing query logs. In *Query Log Analysis Workshop at WWW Conference*, 2007.

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

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

[5] Gagan Aggarwal, Tom´as Feder, Krishnaram Kenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas, and An Zhu. Achieving anonymity via clustering. In *PODS*, pages 153–162, 2006.

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

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

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

[9] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proc. of the ACM SIGMOD*

Conference on Management of Data, pages 439–450. ACM Press, May 2000.
[10] R. Mahaadevan B. A. Beckwith, U. J. Balis, and F. Kuo. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making*, 6(12), 2006.
[11] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
[12] Roberto J. Bayardo and Rakesh 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.
[13] E. Bertino, B.C. Ooi, Y. Yang, and R. H. Deng. Privacy and ownership preserving of outsourced medical data. In *ICDE*, 2005.
[14] Sourav S. Bhowmick, Le Gruenwald,Mizuho Iwaihara, and Somchai Chatvichienchai. Private-iye: A framework
for privacy preserving data integration. In *ICDE Workshops*, page 91, 2006.
[15] Shaofeng Bu, Laks V. S. Lakshmanan, Raymond T. Ng, and Ganesh Ramesh. Preservation of patterns and input-output privacy. In *ICDE*, pages 696–705, 2007.
[16] 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.
[17] Alexandre V. Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy
preserving data mining. In *PODS*, pages 211–222, 2003.
[18] Alexandre V. Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. *Inf. Syst.*, 29(4):343–364, 2004.
[19] 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.
[20] GAO. Efforts to improve information sharing need to be strenghthened, 2003. Homeland Security Highlights.
[21] D. Gupta, M. Saul, and J. Gilbertson. Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research. *American Journal of Clinical Pathology*, 2004.
[22] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition, 2006.
[23] Zhengli Huang,Wenliang Du, and Biao Chen. Deriving private information from randomized data. In *SIGMOD*
Conference, pages 37–48, 2005.
[24] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *SIGKDD*, 2002.
[25] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.
[26] S. Jajodia and R. Sandhu. Toward a multilevel secure relational data model. In *ACM SIGMOD*, 1991.
[27] Elisa Bertino Ji-Won Byun and Ninghui Li. Purpose based access control of complex data for privacy protection.
In *ACM Symposium on Access Control Models and Technologies (SACMAT)*, 2005.
[28] 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.
[29] M. Kantarcioglu and C. Clifton. Privacy preserving k-nn classifier. In *ICDE*, 2005.
[30] Murat Kantarcoglu and J. Vaidya. Privacy preserving naive bayes classifier for horizontally partitioned data. In *IEEE ICDM Workshop on Privacy Preserving Data Mining*, 2003.
[31] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties
of random data perturbation techniques. In *ICDM*, pages 99–106, 2003.
[32] Daniel Kifer and Johannes Gehrke. Injecting utility into anonymized datasets. In *SIGMOD Conference*, pages 217–228, 2006.
[33] Ravi Kumar, Jasmine Novak, Bo Pang, and Andrew Tomkins. On anonymizing query logs via token-based hashing. In *WWW Conference*, 2007.
[34] 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.
[35] K. LeFevre, D. Dewitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *ACM SIGMOD*
International Conference on Management of Data, 2005.
[36] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *IEEE ICDE*, 2006.
[37] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *SIGKDD*, 2006.
[38] Ninghui Li and Tiancheng Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *To appear in International*
Conference on Data Engineering (ICDE), 2007.
[39] Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of Cryptology*, 15(3), 2002.

[40] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. ldiversity:
Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, page 24, 2006.
[41] David J. Martin, Daniel Kifer, Ashwin Machanavajjhala, Johannes Gehrke, and Joseph Y. Halpern. Worst-case
background knowledge for privacy-preserving data publishing. In *ICDE*, pages 126–135, 2007.
[42] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *PODS*, pages 223–228,
2004.
[43] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD Conference*, pages 665–676, 2007.
[44] Mehmet Ercan Nergiz and Chris Clifton. Thoughts on k-anonymization. In *ICDE Workshops*, page 96, 2006.
[45] M. T. Ozsu and P. Valduriez. *Principles of distributed database systems*. prentice hall, 2nd edition, 1999.
[46] Shariq Rizvi and Jayant R. Haritsa. Maintaining data privacy in association rule mining. In *VLDB*, pages 682–693, 2002.
[47] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, fuzziness and knowledge-based systems*, 10(5), 2002.
[48] Latanya Sweeney. Replacing personally-identifying information in medical records, the scrub system. *Journal of the American Informatics Association*, pages 333–337, 1996.
[49] Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*,
10(5):557–570, 2002.
[50] R. K. Taira, A. A. Bui, and H. Kangarloo. Identification of patient name references within medical documents
using semantic selectional restrictions. pages 757–761, 2002.
[51] Zhouxuan Teng and Wenliang Du. Comparisons of k-anonymization and randomization schemes under linking attacks. In *ICDM*, pages 1091–1096, 2006.
[52] S. M. Thomas, B. Mamlin, and G. Schado adn C. McDonald. A successful technique for removing names in pathology reports. pages 777–781, 2002.
[53] TraianMarius Truta and Bindu Vinay. Privacy protection: p-sensitive k-anonymity property. In *ICDEWorkshops*,
page 94, 2006.
[54] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.
[55] J. vaidya and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *SIGKDD*, 2003.
[56] Jaideep Vaidya and Chris Clifton. Privacy preserving nave bayes classifier for vertically partitioned data. In ,
The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.
[57] Jaideep Vaidya and Chris Clifton. Privacy-preserving top-k queries. In *ICDE*, 2005.
[58] 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.
[59] K.Wang and B. C. M. Fung. Anonymizing sequential releases. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 414–423, Philadelphia, PA, August 2006.
[60] 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.
[61] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150, 2006.
[62] Xiaokui Xiao and Yufei Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *SIGMOD Conference*, pages 689–700, 2007.
[63] L. Xiong, S. Chitti, and L. Liu. Topk queries across multiple private databases. In *25th International Conference on Distributed Computing Systems (ICDCS 2005)*, 2005.
[64] L. Xiong, S. Chitti, and L. Liu. K nearest neighbor classification across multiple private databases. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, Arlington, VA, November 2006.
[65] L. Xiong, S. Chitti, and L. Liu. Preserving data privacy for outsourcing data aggregation services. Technical
Report TR-2007-013-A, Emory University Department ofMathematics and Computer Science, 2007. To Appear in ACM Transactions on Internet Technology (TOIT), Special Issue on Internet and Outsourcing.
[66] Li Xiong and Eugene Agichtein. Towards privacy preserving query log publishing. In *Query Log Analysis Workshop at WWW Conference*, 2007.
[67] Li Xiong, Subramanyam Chitti, and Ling Liu. Mining multiple private databases using a knn classifier. In *ACM Symposium of Applied Computing (SAC)*, pages 435–440, 2007.

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

[69] Qing Zhang, Nick Koudas, Divesh Srivastava, and Ting Yu. Aggregate query answering on anonymized tables. In *ICDE*, pages 116–125, 2007.

[70] Sheng Zhong, Zhiqiang Yang, and Rebecca N. Wright. Privacy-enhancing k-anonymization of customer data.
In *PODS '05: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 139–147, New York, NY, USA, 2005. ACM Press.