Database Privacy- Issues and Solutions

Jyoti Devi
School of engineering and sciences
Bhagat phool singh mahila vishwavidyalaya, khanpur
Sonepat, India
kadiayn.niharika@gmail.com

Kavita
School of engineering and sciences
Bhagat phool singh mahila vishwavidyalaya, khanpur
Sonepat, India
yadavkavita42@gmail.com

VikasMalik
Astt Proff (School of engineering and sciences)
Bhagat phool singh mahila vishwavidyalaya, khanpur
Sonepat, India
vikassmalik@gmail.com

Abstract— Data Mining, fourth and analytical step of Knowledge discovery in database process is a process of discovering new and interesting patterns in the large datasets. For example, data miner can derive different patterns based on age, countries or continents on HIV-AIDS if they get suitable dataset to be mined. However, data publishing publish some sensitive information which can lead to chaos, e.g. if data is published as such then HIV effected people loose their privacy .The challenge for data publisher is to publish the data in a form that is most suitable in terms of utility and anonymity. Motivation of this paper comes from the challenges that data publisher face and benefits that data miner can get from data publishing. The aim of this pa-per is to make data publishers way simple enough so that Data Miner or Adversary cannot extract sensitive information out of the dataset. Minimizing the tradeoff between anonymizing and utility of dataset is the primary objective. The method used to accomplish the stated objective is anonymization of raw data using various Privacy Models.

Keywords: Database privacy issue and solution

1.Introduction

As data storage hardware going cheaper day by day, the amount of data we have increases exponentially. There is a high probability to draw important conclusions by analyz-ing these dataset scientifically. So, data publishing which is a dataset transaction between two parties, one, data owner and other one data analyzer has become a very important topic in recent time. Data publishing helps data analyzers while it creates two big challenges for the owners. First, how he change the raw data T in a form T* which he want to publish so that it contains no sensitive information and at the same time it is not anonymized more than optimal which would lead to bad patterns, second how he make sure that T* don't have any sensitive information contained in it. Two challenges are a bit different in that first ask for the methods for converting T to T* while second ask for the methods to derive that T* does not contain any sensitive information. Example 1.1 Group Insurance Commission (GIC), Mas-sachusetts collected medical data of for approximately 135,000 state employees and their families. Assuming the data to be anonymous they gave one copy to researchers and sold another to the industries. In 2002, [3] L. Sweeney could uniquely identify individuals by linking ZIP, date of birth and sex attributes present in of Medical record to that of Voter List. These kinds of linking attacks can be avoid if the

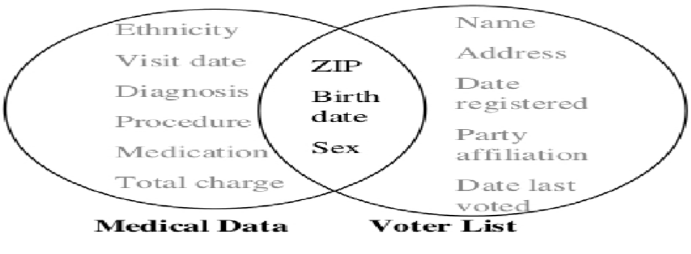


Figure 1: Linking to re-identify data. Both data have three attributes in common which can be used to link individuals in Voter List to their diseases in Medical table

raw data is refined keeping all the ways in consideration by which a linking attack can happen. Solution can be divided in two parts each handing two challenges of data owner. Grouping and Breaking and Perturbation are two types of Anonymization which tackle the first challenge while different Privacy Models tackle with the second challenge of owner. This paper describes different privacy models, their limitations and their power which include k-anonymity, l-diversity, (,)- anonymity. It also describes how to quantify the information loss which occur when the raw data, T is modified to published data, T*. It tackles the problem of tradeoff between utility and anonymity of data.

2. ANONYMIZATION

The process of generalizing the raw data T in a form T* which is published to the data miner is called Anonymization e.g. Table 1[4][1] is the raw table of data owner which he modifies to Table 2 in which Name attribute has been removed. Before dealing with the core of this topic, lets de ne some terms that are used extensively in rest of the paper.

Table 1: Original Medical Table

Name	Gender	Nationality	Age	Disease
Peter	Male	Japanese	26	HIV
John	Male	Malaysian	30	u
Mary	Female	American	36	HIV
Sally	Female	Canadian	40	HIV
Eason	Male	American	40	u
Louis	Male	Chinese	36	u

Table 2: Anonymization, where the attribute Name has been removed from original table

Gender	Nationality	Age	Disease
Male	Japanese	26	HIV
Male	Malaysian	30	U
Female	American	36	HIV
Female	Canadian	40	HIV
Male	American	40	U
Male	Chinese	36	U

2.1 Definitions

2.1.1 Attribute

The columns in a table are the attributes, the things that tell us about the instance in the row. Name, Nationality, Gender, Age and Disease are attributes of Table 1.

2.1.2 Quasi-identifier Attributes

Those attributes which can serve as identifier for instances of data sets. These are written as QI-Attributes. Name, Nationality, Gender, Age are Quasi-identifier attributes of Table 1.

2.1.3 Sensitive Attributes

Those attributes which are prone to contain some sensitive information that if fall in wrong hands can create chaos or we can say the attributes which contain individuals private information. Disease is the sensitive attribute in Table 1.

2.1.4 QI-Group

A collection of all data instances which have identical QI-Attributes values. QI-Group is also known as equivalence class

2.1.5 Micro Data

The original form of data or the unmodified data that the data owner has is called Micro Data. It is denoted as T in the whole paper while modified data is denoted by T*. Table 1 is an example of micro data.

2.1.6 Utility

The degree of information that the modified data T* contains or quality of patterns that can be derived from data T* is called utility of the data T*.

2.1.7 Background Knowledge

The information or knowledge that the data miner has related to the raw date T which can help him to extract sensitive information is called background knowledge. In Example 2.1, the knowledge contained in the Voter List is the background knowledge.

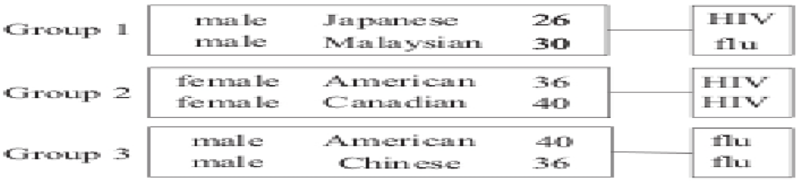
2.2 Methods of Anonymization

2.2.1 Grouping and Breaking

The reason that Data Miner like Sweeney L., was able to extract sensitive information like the diseases one suffer from in Example 1 is that he could link quasi identifier with the corresponding sensitive information with the help of background knowledge of voter list. To remove these kinds of what we call Linking Attacks, the relation between QI-attributes must be broken before publishing the data. The main idea behind grouping is to make tuples or each instance of dataset indistinguishable from others. After grouping the instances, the exact linkage between QI-values is broken, so called Breaking step.

Table 3: Table 2 is grouped in three groups based on Gender attribute

Male Male	Japanese Malaysian	26 30	HIV u
Female Female	American Canadian	36 40	HIV HIV
Male Male	American Chinese	40 36	u u



Group 1	male male	Japanese Malaysian	26 30	HIV flu
Group 2	female female	American Canadian	36 40	HIV HIV
Group 3	male male	American Chinese	40 36	flu flu

Figure 2: Breaking of Table 3.

Before understanding the exact methods for Grouping and Breaking, we need to understand what is called Generalization Taxonomy. Taxonomy is a tree like structure whose different levels has different degree of anonymized values of data attributes. The total values of each level in the tree makes one **Generalization Domain**. When data owner anonymize the table he can generalize a value with some other more generalized and less informative values. He uses a Taxonomy tree to make his decision. Following is an ex-ample of Taxonomy Tree, based on Nationality and Age.

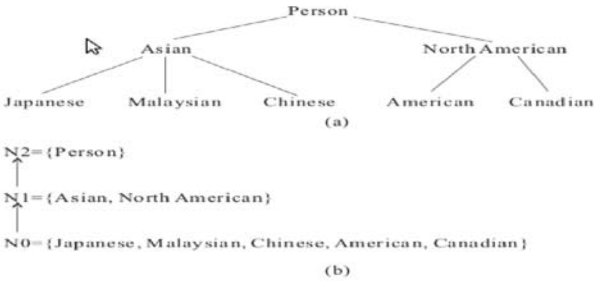


Figure 3: Taxonomy based on Nationality

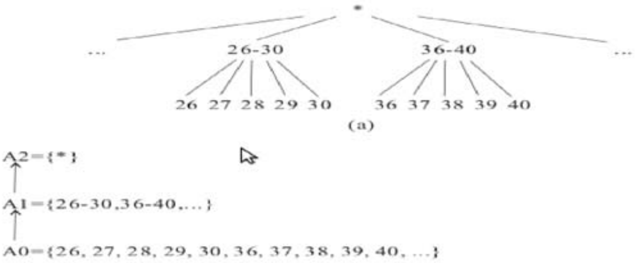


Figure 4: Taxonomy based on Age

Here as one goes from leaf to root the values become more generalized while utility is decreased. The total values of each level in the tree makes one Generalization Do-main.N0,N1,N2 are generalization domains of taxonomy tree in figure 3.Three methods to achieve grouping and breaking are

Suppression

Generalization

Bucketization

2.2.2 Suppression

It is a process where some of attributes are generalized to the root of taxonomy tree making more than one instances of data look exactly same. The linkage between QI-values and sensitive values is automatically broken as one cannot identify individual's sensitive information from the suppressed data. The main problem with this method is that it reduces the utility of data to an extent where analyzer cannot derive interesting patterns.

Person and Age is anonymized with root of age taxonomy tree, *. All males can be group in one group and female in another. Now adversary cannot differentiate between first two rows which correspond to Peter and John and he can-not identify that Peter has HIV. But there are two problems, one is that adversary can easily identify that Sally and Merry both have HIV, as they are in one group where both the sensitive values are HIV and second is that sup-pressed table is over-generalized in that one cannot derive HIV related patterns based on countries or age ranges. First problems is solved by Generalization while second is solved by Perturbation.

Table 4: Table 2 suppressed on Nationality and Age

Gender	Nationality	Age	Disease
Male	Person	*	HIV
Male	Person	*	u
Female	Person	*	HIV
Female	Person	*	HIV
Male	Person	*	u
Male	Person	*	u

2.2.3 Generalization

It solves much of the problem that suppression created. Here, the attributes are encoded to some level of taxonomy tree, not necessarily root of the tree. Two main kind of encoding are

1. Global Encoding: Here all values of an attribute are generalized from one generalization domain. Though it has more degree of utility than the data produced by suppression but still it suffers from over generalization. Table 5 shows the global encoding of table 2, where Nationality is generalized from generalization domain N1 and age is generalized from A2 (see figure 3 & 4). Here data miner can draw patterns based on continents but he cannot make any decisions based on age ranges.

Table 5: Global encoding of Table 2. Attribute Nationality is generalized to first level of taxonomy tree.

Gender	Nationality	Age	Disease
Male	Asian	*	HIV
Male	Asian	*	u
Female	North American	*	HIV
Female	North American	*	HIV
Male	North American	*	u
Male	Asian	*	u

2. Local encoding: Here values of an attribute can be generalized from more than one generalization domains. The modified dataset has more utility and resulting patterns would be more accurate but it makes the dataset less readable or hard to detect the patterns, as the attributes don't follow one pattern. Table 6 is local encoding of table 2. Now data miner can de-rive some HIV patterns based on age ranges, countries and continents. Table looks like the original table and doesn't suffer from over-generalization.

Table 6: Local encoding of Table 2.

Gender	Nationality	Age	Disease
Male	Japanese	26	HIV
Male	Malaysian	26-30	u
Female	North American	36-40	HIV
Female	North American	40	HIV
Male	American	*	U
Male	Asian	*	U

2.2.4 Bucketization

First the table is divided into many partitions and then each partition is given an ID called GID. Here the attribute’s values are not changed instead the dataset is divided in more two subsets. One is called QI-Table and another one is called Sensitive-Table, both having an extra attribute, GID. QI-attributes and Sensitive-attributes are grouped on the bases of GIDs. This lead to a data with high degree of utility as no data instance is changed to any generalized value.

Table 7: Bucketization of Table 1, QI-Table

Gender	Nationality	Age	GID
Male	Japanese	26	1
Male	Malaysian	30	1
Female	American	36	2
Female	Canadian	40	2
Male	American	40	3
Male	Chinese	36	3

Table 8: Bucketization of Table 1, Sensitive-Table

GID	Disease
1	HIV
1	U
2	HIV
2	HIV
3	U
3	U

2.2.5 Perturbation

Here the attributes values are changed to some arbitrary values making the dataset more noisy. Two main methods for Perturbation are

1. Adding Noise:

   This method is generally applicable to numeric at-tributes. The numeric values X_i are modified to X_i + X_i. The noise is added in a way that maintains some statistical values like Mean and vary standard deviations of the numeric data. The problem with this method is that values after adding noise may not exist in real world. The noise is added using standard functions like Normal Distribution, Gaussian Function, - function etc. which have well defined mean and variance. For small dataset Noise can be added by adding value in some values and subtracting from others. Table 9 shows adding noise to Age attribute of the Table 1 where mean value of Age remained unchanged while standard deviation is changed.

Table 9: Adding noise to Age in Table 1

Gender	Nationality	Age	Disease
Male	Japanese	26+1	HIV
Male	Malaysian	30+1	u
Female	American	36-2	HIV
Female	Canadian	40+2	HIV
Male	American	40-4	u
Male	Chinese	36+4	u

2. Value Swapping: Here the values of an attribute are swapped between two tuples. It don’t suffer with the problem of adding noise as the swapped value are the value that existed in original table or in other words they exist in real world. In Table 10 Peter’s Nationality, Japanese is swapped with the Eason ’Nationality, American and his age is swapped with Louis age. Now if adversary has cannot deduce from voter list that Peter has HIV as his Nationality is no more Japanese.

Table 10: Value Swapping

Male	American	36	HIV
Male	Malaysian	30	u
Female	American	36	HIV
Female	Canadian	40	HIV
Male	Japanese	40	u
Male	Chinese	26	u

2.3 Quantifying information loss

When micro data T is changed to T*, how much information is lost? One can quantify the information loss by using taxonomy tree^{0}s height values. Exact procedure of quantifying information loss can be understood in following terms

2.3.1 Distortion

It is the degree of deviation that dataset T* deviate from original dataset T. Distortion of a value V_i in T* is denoted by d_i.

$$d_i = \frac{\text{Height of } G_i : V_i \; 2 \; G_i}{\text{Height of Taxonomy Tree}} : \qquad (1)$$

Distortion of whole dataset, dT8*_ =

P

i di. Distortion of

Second attribute value of _first row in Table 5 is 1/2.

2.3.2 Fully Generalized Dataset, T_f

The dataset where each value is generalized to root of taxonomy tree of corresponding value. It is maximum possible anonymized dataset that could be published. Table 11 shows T_f of Table 2.

Table 11: Fully Generalized Table

Gender	Nationality	Age	Disease
*	Person	*	HIV
*	Person	*	U
*	Person	*	HIV
*	Person	*	HIV
*	Person	*	U
*	Person	*	U

2.3.3 Distortion Ratio of Dataset T , Dr

$$DrT = \frac{dT}{dTf} \qquad (2)$$

2.3.4 Information Loss Metric

The measure of distortion ratio of different anonymized datasets T* of same dataset T is stored in a table called in-formation loss metric. Based on these ratios, the best fitting T* is selected which has least tradeoff between utility and anonymization. The decision needs some extra knowledge of privacy models which are described in next section.

For example, as stated in earlier section that global encoding suffers from more generalization than the local en-coding. It can be seen by calculating distortion ratio of both encoding. Distortion of global encoding Table 5 is 1+1+1+1+1+1+ 0.5+0.5+0.5+0.5+0.5+0.5 =9 while that of local encoding Table 6 is 0.5+0.5+ 0.5+0.5+1+1= 4. Distortion of fully generalized Table 11 is 6+6+6=18.

$$D_r^{Table5} = \frac{9}{18} = 50\% \qquad (3)$$

$$D_r^{Table6} = \frac{4}{18} = 22:22\% \qquad (4)$$

So, now we have two values in our Information Loss Metric with two different values of distortion ratios. Data owner will select Table 6.

3. PRIVACY MODELS

Having described what anonymization means, what are different methods to achieve it and how to quantify it, now there is a need to tackle the second challenge of data publisher i.e. how he ensure that data that he is publishing doesn't contain any sensitive information. This is solved by making various privacy models which are very helpful in quantifying individual privacy. Following are some famous privacy models

3.1 k-Anonymity

3.1.1 Definition

Let T (A_1, A1, AN) be a dataset, with fQIG$_i$g as its quasi-

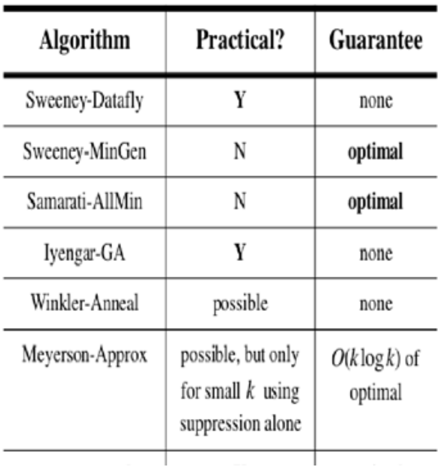
Identifier groups. A QIG$_i$ is k-anonymous if it has at least k instances of dataset T. T(A_1,A_1,...,A_N) is said to satisfy k-anonymity if every QIG$_i$ is k-anonymous.

3.1.2 Significance

The significance of k-anonymity is that it ensures each data instance in a group indistinguishable from at least other k-1 data instances. It means that adversary cannot distinguish between k instances of a group even if he has some back-ground knowledge. This tackles the second challenge of data publisher. Now he can be sure that sensitive information is secure with 1/k probability. Table 12 satisfy 2-anonymity. It has three QIG groups each having 2 instances of dataset. Now having said that the table is 2-anonymous, adversary cannot decide that Peter has HIV as there are two instances in his QI group and it might be other one who has HIV. It ensures that Peter's privacy is 50% safe. Even data owner was able to secure Peter's privacy; still he is disposing Merry and Sally's privacy by 100%. Merry and Sally are in the same

group but they both have HIV. Data miner cannot point which instance corresponds to which lady but he is sure that both have HIV. This problem is solved by l-diversity described in next subsection. The question is that can we k-anonymize a T (A_1,A_1,...,A_N) with minimum information loss? Sadly, proving this is a NP-Hard Problem. However people has contributed a lot in this area. Some algorithms with their

Optimality and running time is displayed in Figure 5[1][2]



Algorithm	Practical?	Guarantee
Sweeney-Datafly	Y	none
Sweeney-MinGen	N	**optimal**
Samarati-AllMin	N	**optimal**
Iyengar-GA	Y	none
Winkler-Anneal	possible	none
Meyerson-Approx	possible, but only for small k using suppression alone	$O(k\log k)$ of optimal

Figure 5: A breakdown of known approaches to k-anonymity

Table 12: 2-anonymous table

Gender	Nationality	Age	Disease
Male	Asian	26-30	HIV
Male	Asian	26-30	U
Female	North American	36-40	HIV
Female	North American	36-40	HIV
Male	Person	36-40	U
Male	Person	36-40	U

3.1.3 Attacks against k-anonymity

As describe above k-anonymity leaves a room for leakage of sensitive information. There are some other attacks[3] that can extract sensitive information out of k-anonymous table.

1. Unsorted matching attack against k-anonymity:

   This attack is due to the order in which tuples occur in the table. Table 13 and Table 14 are two tables which satisfy 2-anonymity. Now Table 13 is released first, knowing that we have anonymized age to the *, we are safe. Then we released another version of Table 13 i.e. Table 14. Now again we were sure about the privacy preserving because Nationality is anonymized to Person. But adversary can link each tuple of two tables and form a new table with Nationality from Table 13 and Age from Table 14. In this way he can match this new table from his background knowledge, Table 14 and found that only Asian whose age is 26 is Peter. This kind of attacks can be resolved by disordering tuples in a random order before publishing.

Table 13: 1st Version

Gender	Nationality	Age	Disease
Male	Asian	*	HIV
Male	Asian	*	u
Female	North American	*	HIV
Female	North American	*	HIV
Male	American	*	u
Male	Asian	*	u

Table 14: Second Version

Gender	Nationality	Age	Disease
Male	Person	26	HIV
Male	Person	30	U
Female	Person	36	HIV
Female	Person	40	HIV
Male	Person	40	U
Male	Person	36	U

Table 15: Background Knowledge of Adversary

Name	Nationality	Age
Peter	Person	26
Danny	Person	30
John	Person	36
Persia	Person	40
Arya	Person	40
Louis	Person	36

2. Temporal Attacks against k-anonymity:

   Suppose Table T_1 is released at time t_1 which satisfied k-anonymity and Table T_2 which is a modified version of T_2 (we added or deleted some tuples) is released at time t_2. Then link these two tables can leak some privacy. These attacks are called Temporal Attacks.

3.2 l-diversity

3.2.1 Definition

Let $T(A_1,A_1,...,A_N)$ be a dataset, with fQIG$_i$g as its quasi-

Identifier groups. A QIG$_i$ is l-diverse if the probability that a data instance of this group is linked to a sensitive value is at most 1/l. T $(A_1,A_1,...,A_N)$ is said to satisfy l-diversity if every QIG$_i$ is l-diverse.

3.2.2 Significance

The significance of l-diversity over l-anonymity is that it links QI-attributes to the sensitive values. K-anonymity didn't have any relation

between QI-attributes and sensitive values which lead to privacy leakage. Table 16 satisfies 2-diversity as it have 2 QIGs and the probability that a tuple is linked to HIV is exactly 0.5 in both the groups. In this table adversary cannot identify Merry and Sally^{0}s HIV disease. It solves the problem that was in k-anonymity.

Table 16: 2-diverse

Gender	Nationality	Age	Disease
Male	Person	26-30	HIV
Male	Person	30-30	u
*	Person	36-40	HIV
*	Person	36-40	HIV
*	Person	36-40	u
*	Person	36-40	u

3.3 (,)-anonymity

3.3.1 Definition

A table $T(A_1,A_1,...,A_N)$ is said to be (,)-anonymous) if it satisfy -anonymity and l-diversity, where l=1/ . Here 2 [0,1].

3.3.2 Significance

It consider both QI-attributes and sensitive attributes. k-anonymity is special cases of (,)-anonymity, when = k and l=1. Table 16 satisfy (0.5,2)-anonymity as it is 2-anonymous and 2-diverse.

3.4 Other Privacy models

There are many other privacy models present in the DBMS security market. (k,e)-anonymity and (,m)-anonymity are some other famous privacy models

4. SOME OTHER POSSIBLE SOLUTIONS

There are some other solutions[2][4] to the problem of privacy leakage in data publishing. We provide just overview of the methods without dealing with how they are accomplished.

4.1 Limiting Access

The approach is usually taken by secure DBMS community. Controlling the access for data can certainly combat the problem of privacy leakage. However, it need to quantify how much control data owner want to give to the data miner. It also restrict limited quality pattern extractions.

4.2 Fuzz the Data

This is another term for anonymization. If we alter some of the values, data miner will confuse himself if he starts extracting sensitive information as the data is not totally correct. It suffers from the problems that we discussed in Anonymization Section.

4.3 Eliminate Unnecessary Grouping

Usually the published data has various contiguous sequences of data. If data Miner and information about one member form the group then he can extract information about other members of the group. For example, consider the Voter ID number provided to the people, which are sequential in a particular locality. Also, the phone numbers are also allocated sequentially based on their Voter ID number. Then if data miner found the telephone number of one of the member in locality, he can extract whole set of phone numbers of that locality and can sell to the company. To prevent such kind of attacks data should never be grouped sequentially while the groups should be broken before publishing the data.

4.4 Augment the Data

If data is augmented with some instances of data then the data become more safe in that if adversary will make a query for some data there is a high probability that he will end up with more number of results other than actual result.

4.5 Audit

Though not very feasible but it is also another solution to the stated problem. If data owner audit the mined data time to time then he can get to know if their if something illegal going on and can hand over the data miner in custody. It require a bond between the two parties before setting up a deal.

5. CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LATEX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

ACKNOWLEDGEMENT

We would like to give our sincere gratitude to our guide **Mr. Vikas Malik** who guides us to pursue this topic and help us for completion of this topic

6. REFERENCES

[1] Roberto J. Bayardo and Rakesh Agarwal. Security and privacy implications of data mining. In Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), page 2. IEEE-COmputer Society, 2005.

[2] Chris Clifton and Don Marks. Security and privacy implications of data mining. In Proceedings of the 1996 ACM SIGMOD Workshop on Data Mining and Knowledge Discovery, pages 2{3. ACM SIGMOD, 1996.

[3] L. Sweeney. k-anonymity: A model for protecting privacy. International Journal on Uncertainty,Fuzziness and Knowledge-based Systems, pages 557{570, May 2002.

[4] Raymond Chi-Wing Wong and Ada Wai-Chee Fu. Privacy-Preserving Data Publishing: An Overview.