An Approach to Overcome Inference Channels on *k*-anonymous Data

Ousseynou Sané[1], Fodé Camara[1], Samba Ndiaye[1] and Yahya Slimani[2]

[1]*Department of Mathematics, Cheikh Anta Diop University*
Dakar, Senegal
{ousseynou.sane, fode.camara, samba.ndiaye}@ucad.edu.sn

[2]*University Department of Computer Science, Faculty of Sciences of Tunis*
1060 Tunis, Tunisia
yahya.slimani@fst.rnu.tn

Abstract

The concept of k-anonymity protection model has been proposed as an effective way to protect the identities of subjects in a disclosed database. However, from a k-anonymous dataset it may be possible to directly infer private data. This direct disclosure is called attribute linkage. k-anonymity also suffer to another form of attack based on data mining results. In fact, data mining models and patterns pose a privacy threat even if the k-anonymity is satisfied. In this paper, we discuss how the privacy requirements characterized by k-anonymity can be violated by data mining results and introduce an approach to limit privacy breaches. We experiment it by using the adult dataset from the UCI KDD archive. We report the experimental results which show its effectiveness.

Keywords: *Data anonymization, Data mining, Knowledge hiding, Privacy preserving*

1. Introduction

Information sharing has become part of the activities of many individuals, companies, organizations, and government agencies. If it is today a need that is unceasingly growing, because it can bring much advantages in collaboration between various organizations. Whereas, it raises many questions relating to individual privacy. As a scenario to illustrate this problem, we can consider the example of a hospital that collects large volumes of sensitive data of individuals that are valuable for research and decision making. Sharing or publishing data about individuals is however prone to privacy attacks, breaches, and disclosures.

The concept of *k*-anonymity [1] has been proposed as an effective way to protect the identities of subjects in a disclosed database. Because of its conceptual simplicity, it has been widely discussed as a viable definition of privacy in data publishing, and due to algorithmic advances in creating *k*-anonymous versions of a dataset [1], *k*-anonymity has grown in popularity. However, *k*-anonymity is vulnerable to attribute linkage attacks (e.g. homogeneity attacks, background knowledge attacks). As shown in [2], *k*-anonymity also suffer to another form of attack based on knowledge discovery results. In fact, data mining models and patterns pose a privacy threat even if the *k*-anonymity is satisfied.

1.1. How Knowledge Discovery Results Create Inference Channels on *k*-Anonymous Data

The application of data mining task to a collection of anonymous data can result in disclosure of sensitive knowledge. To illustrate this threat, we first recall the C4.5 decision tree algorithm [3], secondly we highlight the inference channels created by the decision rules.

1.1.1. Overview of Decision Tree Construction: We specially studied C4.5 decision tree classifier [3] which is one of the best data mining algorithms. The C4.5 algorithm is an extension of the ID3 algorithm [3], and has been proposed by Quinlan in [3]. Some of the improvements to ID3 are: (i) handling both numeric and categorical attributes; (ii) handling training data with missing attribute values; (iii) pruning trees after creation, C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes. At each node of the tree, C4.5 chooses one attribute of the dataset that most effectively splits its set of samples into subsets.

One of the more attracting aspects of decision trees resides in their interpretation based on the decision rules which they generate [4]. The rules can very simply be built from a decision tree while crossing all the ways of the root towards any leaf. This complete set of decision rules generated by a decision tree is equivalent (in terms of decision-making) to the decision tree itself.

A decision rule is an implication of the form <*if antecedent, then consequent*>. It can also be represented as follows: *antecedent* → *consequent*. The consequent is formed by a value of attribute class (i.e. a leaf node of the decision tree). The support of a decision rule relates to the proportion of records of the training set which belong to the leaf node (i.e. predicted attribute). The confidence of a decision rule is the proportion of records of the leaf node for which the rule is true. If confidence is equal to 100% (=1), the leaf node is *pure* and the decision rule is *perfect*.

1.1.2. Inference channels created by decision rules on *k*-anonymous data: After the anonymization process, it is also possible to infer sensitive data. To illustrate this, we give the example in Table 1 which represents a *4*-anonyme version of a medical data. After building the C4.5 decision tree, we have certain sensitive decision rules as the following rule: 3*⇒ Cancer, confidence=1. This kind of rules is often useful for medical research, but can allow inferring the disease of certain individuals because it expresses clearly that all the patients in the initial database, who are old between 30 and 39 are cancerous. Through this example, we can see that data mining results can violate privacy even when a *k*-anonymity definition is satisfied.

Table 1. A k-anonymous Dataset

	ZipCode	Age	Nationality	Disease
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Heart Disease
7	1485*	≥40	*	Viral Infection
8	1485*	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

1.2. Contribution and Paper Outline

In this paper, we highlight that data mining results can cause privacy breaches. More specifically, we show that C4.5 decision rules create inference channels which an adversary can use to find private data of an individual. Our contribution is twofold:

- We define the concept of sensitive decision rules which potentially threat anonymity of data source.

- We develop an effective algorithm which aims to eliminate the threats to anonymity by reducing the confidence of the sensitive decision rules under a threshold chosen by the miner.

The remainder of this paper is organized as follows. We begin by reviewing related prior research on attack techniques on Privacy. Our proposed is presented in Section 3. Section 4 presents our experimental studies and Section 5 provides and discusses the obtained results. Finally, Section 6 concludes with a discussion of the contributions of our proposal and our current research plans.

2. Related Research and Insights

A privacy threat occurs either when an identity is linked to a record or when an identity is linked to a value on some sensitive attribute. These threats are respectively called *record linkage* and *attribute linkage*.

2.1. Record Linkage

The record linkage occurs when some values q of quasi-identifiers Q identifies a smaller number of records in the released dataset T. In this case, the record holder having the value q is vulnerable to being linked to a small number of records in T.

The notion of k-anonymity [1] was proposed to combat record linkage. k-anonymity is a stronger model of privacy protection. It limits disclosure risk to an acceptable level. The guarantee obtained with k-anonymity is that no information can be linked to groups of less than k individuals. Therefore, the degree of uncertainty of sensitive attribute is at least $1/k$. However the main drawback in k-anonymity is its vulnerability to attribute linkage.

2.2. Attribute Linkage

If some sensitive values are predominate in a group, an attacker has no difficulty to infer such sensitive values for a record holder belonging to this group. Such attacks are called *attribute linkage*. More specifically, *k*-anonymity suffers from two types of attribute linkage:

- *Homogeneity attacks. k*-anonymity protection model can create groups that leak information due to lack of diversity in the sensitive attribute. In fact, *k*-anonymization process is based on generalizing the quasi-identifiers but does not address the sensitive attributes which can reveal information to an attacker.

- *Background knowledge attacks*. Beside to homogeneity attacks, the background knowledge attacks can compromise privacy in *k*-anonymous database. In fact, an adversary can have knowledge that a priori enables him to guess sensitive data with high confidence. This kind of attacks depends on other information available to an attacker.

Given these two weaknesses, several models are introduced to combat attribute linkage. Among this models, we can cite *l*-diversity, (α, *k*)-Anonymity and (*X*, *Y*)-privacy [1]. However, the latter are may be difficult to achieve and generally compromise the usefulness and significance of the mining results [5]. In fact, finding equilibrium between the amount of privacy and the utility loss resulting from the anonymization process is an important issue.

3. Proposed Approach

3.1. Problem Definition

Before introducing our anonymity preservation problem, we need to define the sensitive decision rules, which are potentially threat to privacy.

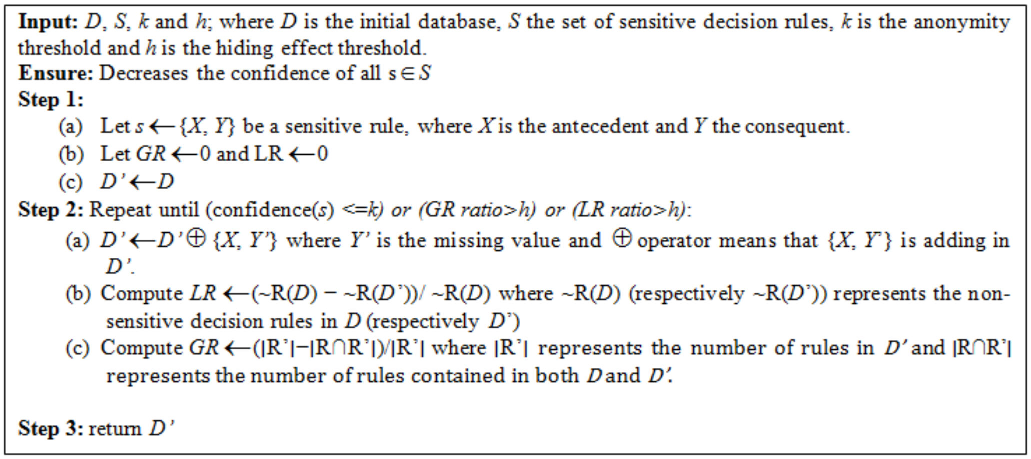
Definition 1. Let *T* be a database and $A= \{a_1, a_2, ..., a_m\}$ be a set of attributes. A decision rule is an implication of the form $X \Rightarrow Y$, where $X \subset A$, $Y \subset A$, and $X \cap Y = \phi$. The decision rule $X \Rightarrow Y$ holds in *T* with confidence *c* equals to 1 is habitually named *perfect decision rule*. In our anonymity preservation problem, *perfect decision rules* are denoted *sensitive decision rules*.

Informally, we call inference channel any subset of the collection of sensitive decision rules, from which it is possible to infer private data of an individual.

Definition 2. Given *S* a collection of sensitive decision rules mining from a database *T* and an anonymity threshold *k*, our problem consists in reducing the confidence of each rule $s \in S$: $0<\text{conf}(s)<k$.

3.2. Algorithm

When hiding inference channels, one always needs to find a good equilibrium between the amount of privacy and the utility loss resulting from this hiding process. In order to control the utility loss, we use the following metrics: the *Lost Rules Ratio* (*LR*) and the *Ghost Rules Ratio* (*GR*) [6]. The first refers to the percentage of the non-sensitive rules in the sanitized dataset to the total non-sensitive rules in the initial dataset. And the second refers to the percentage of the ghost rules in the sanitized *D’* to total rules in *D’*. If either *LR* or *GR* is higher than a *hiding effect threshold* denoted by *h* (chosen by the miner), the reducing confidence process is stopped, and then, the corresponding sanitized database *D*' is returned.



```
Input: $D$, $S$, $k$ and $h$; where $D$ is the initial database, $S$ the set of sensitive decision rules, $k$ is the anonymity threshold and $h$ is the hiding effect threshold.
Ensure: Decreases the confidence of all $s \in S$
Step 1:
    (a) Let $s \leftarrow \{X, Y\}$ be a sensitive rule, where $X$ is the antecedent and $Y$ the consequent.
    (b) Let $GR \leftarrow 0$ and $LR \leftarrow 0$
    (c) $D' \leftarrow D$
Step 2: Repeat until (confidence($s$) <=$k$) or (GR ratio>$h$) or (LR ratio>$h$):
    (a) $D' \leftarrow D' \oplus \{X, Y'\}$ where $Y'$ is the missing value and $\oplus$ operator means that $\{X, Y'\}$ is adding in $D'$.
    (b) Compute $LR \leftarrow (\sim R(D) - \sim R(D'))/ \sim R(D)$ where $\sim R(D)$ (respectively $\sim R(D')$) represents the non-sensitive decision rules in $D$ (respectively $D'$)
    (c) Compute $GR \leftarrow (|R'| - |R \cap R'|)/|R'|$ where $|R'|$ represents the number of rules in $D'$ and $|R \cap R'|$ represents the number of rules contained in both $D$ and $D'$.

Step 3: return $D'$
```

Algorithm 1. Blocking Inference Channels Algorithm

4. Experimental Validation

We ran experiments on the adult dataset from the UCI machine learning repository [7]. It is question there of predicting whether income exceeds $50K/yr based on census data. We used the following ten original attributes: *education*, *race*, *sex*, *work-class*, *marital-status*, *age*, *relationship*, *native-country*, *occupation* and *salary*. In order to keep the usefulness of the data for data mining task, we only consider the *age* and *sex* attributes to compose the quasi-identifiers. The attribute *salary* was considered as sensitive attribute. The records with missing values were removed and the resulting dataset contained 45,222 records. Experiments can be broken into three steps:

- **Step 1: *k*-anonymization process.** To anonymize this dataset, we used the UT Dallas Anonymization Toolbox [8] which currently contains 6 different anonymization methods (e.g. Datafly, Incognito with *k*-anonymity, Incognito with *l*-diversity, Incognito with *t*-closeness, Mondrian and Anatomy[1]) over 3 different privacy definitions (e.g. *k*-anonymity, *l*-diversity and *t*-closeness [5]). We specially used the Datafly method which is the first algorithm to satisfy *k*-anonymity privacy definition [9]. The algorithm uses full-domain generalization until every combination of quasi-identifier values appears at least *k* times. Figure 1 gives a representation of the value generalization hierarchies (VGHs) of "*age*" and "*sex*" attributes. We chose *k*=10. Therefore, the degree of uncertainty of the sensitive attribute in the generated *10*-anonymous is at least *1/k*.

- **Step 2: Detecting Inference Channels.** For building a decision tree on the 10-anonymous dataset obtained in the first step, we ran the C4.5 decision tree algorithm using '*J48 Decision tree classifier*' of Weka[10], a widely used data mining toolkit. We detect all possible inference channels as described in *Definition 1*.

- **Step 3: Hiding Inference channels.** To hide the inference channels created by the sensitive decision rules generated in the previous step, we used the *Algorithm 1*. To assess the information loss from this hiding strategy, we ran '*J48 Decision tree* *classifier*' on the three datasets: the initial dataset, the *10*-anonymous dataset obtained in the first step, and that resulting from the integrated privacy preserving process. In each one of them, we used 70% as training set and 30% as test set. Table 2 displays the obtained results.

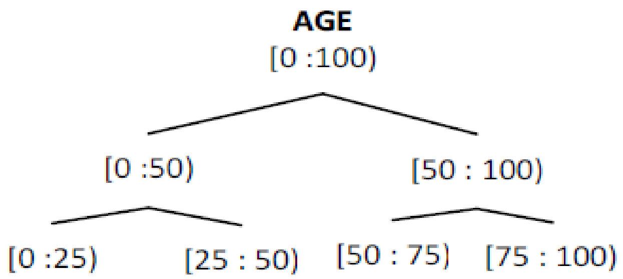


Figure 1. Graphical Representation of VGHs of "*age*" and "*sex*" Attributes

5. Results and Discussion

Tables 2, 3 highlight the differences between the traditional *k*-anonymity and the Blocking Inference Channels in *k*-anonymity (BIN *k*-anonymity) which we proposed under two metrics: the data quality and the privacy preserving.

Table 2 shows that the data quality resulting from our proposed approach is acceptable because it is slightly decreased. As analyzed in [11], privacy preserving process through insertion false information usually causes the decrease of the data quality. It is obvious that the more the changes are made to the database, the less the database reflects the domain of interest. If data quality is too degraded, the released database is useless for the purpose of knowledge extraction. For this reason, in each addition, Algorithm 1 tests if the data quality is not degraded too much by computing the *Lost Rules Ratio* (*LR*) and the *Ghost Rules Ratio*. Note that the classifier accuracy is closely related to the *information loss* resulting from the hiding strategy: the less is the information loss; the better is the data quality.

Table 2. Data Quality Comparison

Dataset	**Classification Accuracy**
Initial Adult Dataset	83,23%
10-anonymous Adult Dataset	82,54%
BIC 10-anonymous Adult Dataset	82,09%

Although the blocking inference channels in *k*-anonymity decrease the data quality, it improves the privacy protection level. Table 3 gives a comparison between the traditional *k*-anonymity and BIC *k*-anonymity (Blocking inference channels in *k*-anonymity) under the three families of attacks we discussed in Section 2. As opposite to *k*-anonymity, our BIC *k*-anonymity controls the sensitive rules-based inferences.

Table 3. Privacy Protection Comparison

Privacy Model	Record Linkage	Attribute Linkage	Sensitive Rules-based Inferences
k-anonymity	✓		
BIC *k*-anonymity	✓		✓

We can summarize that BIC *k*-anonymity improves the *k*-anonymity definition because it provides better privacy protection with degrading slightly the data utility (data quality). As shown in the literature, maximizing both the privacy protection and the data utility is not possible: better privacy protection is, more utility loss is important, and vis-versa [11].

6. Conclusion

In this paper, we studied the data anonymization problem which has two conflicting goals: privacy protection and the data utility preserving. In order to find a good trade-off between the level of privacy and the data quality, we introduce a novel approach which improves the *k*-anonymity protection while keeping the usefulness of the data. In the future, we plan to extend this work by overcoming the attribute linkage attacks and the probabilistic attacks while also maintaining the data utility.

References

[1] V. Ciriani, S. De Capitani di Vimercati, S. Foresti and P. Samarati, "k-Anonymous Data Mining: A Survey", in Privacy-Preserving Data Mining: Models and Algorithms, Charu C. Aggarwal and Philip S. Yu (eds), Springer-Verlag, **(2008)**.

[2] M. Atzori, F. Bonchi, F. Giannotti and Dino Pedreschi, "k-anonymous patterns", 9th European Conference On Principles And Practice Of Knowledge Discovery In Databases **(2005)**.

[3] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, **(1993)**.

[4] D. T. Larose, "Discovering Knowledge in Data", An Introduction to Data Mining, John Wiley & Sons, Inc., **(2005)**.

[5] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, **(2007)** April 15-20, pp.106-115.

[6] T. Johnsten and V. V. Raghavan, "A methodology for hiding knowledge in databases", Proceedings of the IEEE International Conference on Privacy, Security and Data Mining, Maebashi City, Japan, vol. 14, pp. 9-17.

[7] U. C. Irvine, "Machine Learning Repository", http://www.ics.uci.edu/mlearn/mlrepository.html.

[8] UTD Anonymization Toolbox, http://cs.utdallas.edu/dspl/cgi-bin/toolbox/.

[9] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", Int. J. Uncertain. Fuzziness Knowl.-Based System, vol. 10, no. 5, **(2002)**, pp. 571-588.

[10] H.W. Ian and F. Eibe, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann, **(1999)** October.

[11] E. Bertino, I. Fovino and I. Provenza, "A Framework for Evaluating Privacy-Preserving Data Mining Algorithms", Data Mining and Knowledge Discovery Journal, vol. 11, no. 2, **(2005)**.