# Privacy Leakage in Multi-relational Databases via Pattern based Semi-supervised Learning

Hui Xiong
MSIS Department
Rutgers University
hui@rbs.rutgers.edu

Michael Steinbach
Computer Science
University of Minnesota
steinbac@cs.umn.edu

Vipin Kumar
Computer Science
University of Minnesota
kumar@cs.umn.edu

## ABSTRACT

In multi-relational databases, a view, which is a context- and content-dependent subset of one or more tables (or other views), is often used to preserve privacy by hiding sensitive information. However, recent developments in data mining present a new challenge for database security even when traditional database security techniques, such as database access control, are employed. This paper presents a data mining framework using semi-supervised learning that demonstrates the potential for privacy leakage in multi-relational databases. Many different types of semi-supervised learning techniques, such as the K-nearest neighbor (KNN) method, can be used to demonstrate privacy leakage. However, we also introduce a new approach to semi-supervised learning, hyperclique pattern based semi-supervised learning (HPSL), which differs from traditional semi-supervised learning approaches in that it considers the similarity among groups of objects instead of only pairs of objects. Our experimental results show that both the KNN and HPSL methods have the ability to compromise database security, although HPSL is better at this privacy violation than the KNN method.

**Categories and Subject Descriptors:** H.1 [Information Systems-Models and Principles]:Miscellaneous

**General Terms:** Security, Algorithms.

**Keywords:** Semi-supervised Learning, Database Security, Hyperclique Patterns, Privacy Preserving Data Mining.

## 1. INTRODUCTION

This paper investigates privacy leakage in database views via semi-supervised learning.[1] Figure 1 illustrates this problem. In the figure, the view contains $m$ attributes and $n$ tuples (objects); all this information is known by the user. There are also $p$ attributes that are in base tables but not in the view; the information in these $p$ attributes is unknown to a user of the database view. However, if, for some objects, these $p$ attributes are known to a user of the database view, then such a user may use semi-supervised learning techniques to predict these $p$ attributes for other objects. This is the problem addressed in this paper.[2]

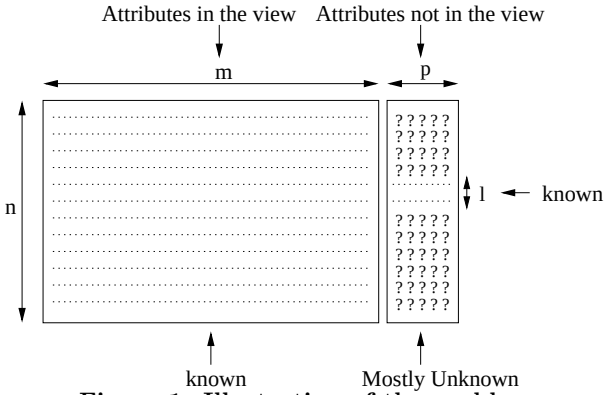


**Figure 1: Illustration of the problem.**

This problem is challenging, since the number of training objects is much smaller than the number of objects that are to be predicted. This is typically referred to as the small training sample size problem [3][4] in machine learning. Indeed, it was demonstrated that supervised classification techniques cannot obtain reliable results if only a very small set of samples are available [1]. To this end, semi-supervised learning techniques [5], which make use of both unlabeled and labeled data, have recently been proposed to cope with the above challenge.

## 2. PROPOSED APPROACHES

We describe three proposed approaches for privacy violation using semi-supervised learning.

**The KNNS method.** For an object with a class label, the KNNS method labels the $k$ nearest (unlabeled) neighbors with the same class label as the object. If a predicted object is found to be one of $k$ nearest neighbors of more than one given object, then the KNNS method assigns the label of the given object with the highest similarity. The KNNS method only considers pairs of similar objects when labeling the data objects. However, in real world data sets, it is possible that two objects are often nearest neighbors without

[1]This work was partially supported by NASA grant #NCC 2 1231, NSF grant #ACI-0325949 and by Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014. The content of this work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by the AHPCRC and the Minnesota Supercomputing Institute.


*CIKM'05,* October 31–November 5, 2005, Bremen, Germany.


[2]Full details of this research are given in [6]

belonging to the same class. In addition, the KNNS method predicts an equal number of objects for each object with a class label; that is, this method gives each labeled object equal weight as a predictor. This may not be appropriate in real-world data sets where, typically, objects from a high-density cluster predict more objects with a higher accuracy. In the worst case, a labeled object can be noise or an outlier that is completely unsuitable for prediction.

**The TOP-K NNS method.** For $n$ given objects with class labels, the TOP-K NNS method finds the $k$ objects with the highest level of similarity from the neighborhood of these $n$ objects. The TOP-K NNS method assigns different predictive power to different labeled objects. Thus, unlike the KNNS method, the TOP-K NNS method can avoid prediction errors that result when some labeled objects are noise or outliers. However, like KNNS, the prediction mechanism of the TOP-K NNS method is also solely based on pairwise similarity. In real world data sets, it is possible that two objects can be nearest neighbors without belonging to the same class. Therefore, Top-K NNS can also generate many prediction errors.

**Hyperclique pattern based semi-supervised learning (HPSL) method.** Recently, we have defined a new pattern for association analysis—the *hyperclique pattern* [7]—that demonstrates a particularly strong connection between the overall similarity of a set of attributes (or objects) and the itemset (local pattern) in which they are involved. The hyperclique pattern possesses the *strong affinity property*, i.e.; the attributes (objects) in a hyperclique pattern have a guaranteed level of global pairwise similarity to one another as measured by the cosine similarity measure. Intuitively, a hyperclique pattern includes objects which tend to be from the same class category. Based on this observation, we propose a new semi-supervised learning approach, the hyperclique pattern based semi-supervised learning (HPSL) method. By considering the similarity among all objects in a hyperclique instead of the similarity between only pairs of objects, we can improve semi-supervised learning results over those based on KNN approaches.

More specifically, for an object with a class label, we find a maximal hyperclique pattern that contains this object and then label all other objects in the pattern with the label of this object. If the hyperclique pattern contains objects with different class labels, then our algorithm assigns an unlabeled object the class label of the labeled object that has the highest similarity to the unlabeled object. A similar strategy can be applied when an unlabeled object is located in two different maximal hyperclique patterns.

There are several benefits of the HPSL method. First, this method, like KNNS and TOP-K NNS, only predicts class labels for objects strongly connected to objects with known class labels. Second, unlike these two approaches, HPSL considers the similarity among groups of objects instead of just pairs of objects. Third, hyperclique patterns represent unique concepts that may potentially help guide better information inference in databases. Finally, the application of the HPSL method for attacking database security reveals an interesting direction for multi-relational data mining [2].

# 3. EXPERIMENTAL EVALUATION

In this section, we discuss some experimental results to (1) show the information leakage in databases via the HPSL method with experiments on several real-world data sets, and (2) compare the relative performance of HPSL, KNNS, and TOP-K NNS. For our experiments, we used several real life data sets. Results shown in this paper used the WAP data set from the WebACE project, which has 1560 documents, 8460 terms, and 20 classes.

Consider Figure 2, which shows the prediction accuracy of the HPSL method and the nearest neighbor based approaches, KNNS and TOP-K NNS, as the number of objects with known class labels is increased. In this experiment, we specified the total number of predicted objects to be approximately five times more than the number of objects with class labels. We also performed random sampling to select objects with class labels. Finally, in order to reduce the random effect, we conducted 10 trials for each experiment.

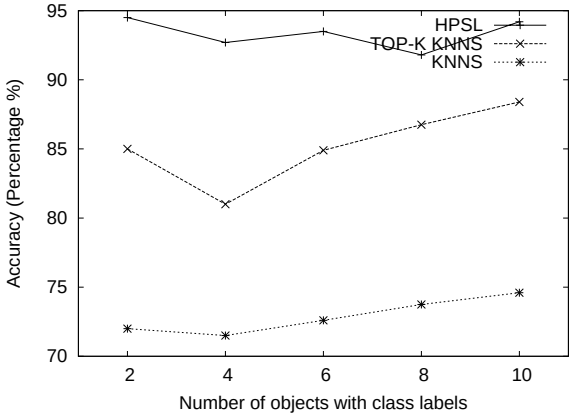


**Figure 2: Relative classification performance of KNNS, TOP-K NNS, and HPSL on WAP.**

First note that the classification accuracy of all three techniques is above 70% and ranges up to 95%, indicating a high degree of information leakage. Similar results were obtained on other data sets [6]. As can also be seen from this figure, for most observed numbers of objects with known class labels, the achieved accuracy of the HPSL method is significantly and systematically better than that of the KNNS and TOP-K NNS methods. This is due to the fact that the HPSL method has the power to eliminate the isolated data objects that often result in prediction errors in nearest neighbor approaches, such as KNNS and TOP-K NNS. Another observation is that the TOP-K NNS method performs much better than KNNS in terms of accuracy.

# 4. REFERENCES


[1] C. Clifton. Using sample size to limit exposure to data mining. *J. Comput. Secur.*, 8(4):281–307, 2000.

[2] P. Domingos. Prospects and challenges for multi- relational data mining. *SIGKDD Explorations*, 2003.

[3] R. Duin. Classifiers in almost empty spaces. In *Proc. 15th Int'l Conference on Pattern Recognition*, 2000.

[4] S. Raudys and A. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE TPAMI*, 13(3):252–264, 1991.

[5] M. Seeger. Learning with labeled and unlabeled data. In *Technical Report, University of Edinburgh*, 2001.

[6] H. Xiong, M. Steinbach, and V. Kumar. Privacy leakage in multi-relational databases via pattern based semi-supervised learning. Technical Report 04-023, University of Minnesota, 2004.

[7] H. Xiong, P. Tan, and V. Kumar. Mining strong affinity association patterns in data sets with skewed support distribution. In *Proc. of ICDM*, 2003.