Privacy Preserving in Association Rule mining

Nilesh R. Radadiya[1], Nilesh B. Prajapati[2], Krupali H. Shah[3]

[1]*Research scholar student, B.V.M., V.V.Nagar, GTU,INDIA.*

[2]*Information Technology Department, B.V.M.,V.V.Nagar, GTU, INDIA.*

[3]*Information Technology Department, B.V.M., V.V.Nagar, GTU, INDIA.*

[1]radadiya_nilesh@ymail.com,[2]nilesh.prajapati@bvmengeering.ac.in,
[3]krupali.shah@bvmengeering.ac.in

***Abstract*— The recent advances of data mining techniques are widely used for many applications in today's information world. However the misuse of these techniques may disclose the private or sensitive information which database owners do not want reveal to others. Therefore many of the researchers in knowledge hiding field have recently made efforts to preserve privacy for sensitive knowledge. In this Paper, we propose a heuristic based algorithm named ADSRRC (Advance Decrease Support of R.H.S. item of Rule Clusters) to hide the sensitive association rules with multiple items in consequent (R.H.S) and antecedent (L.H.S). This algorithm overcomes the limitation of existing rule hiding algorithm DSRRC. Proposed algorithm ADSRRC selects the items and transactions based on certain criteria which modify transactions to hide the sensitive information. Example illustrating the proposed approach is also given. Then performance of the algorithm is evaluated to show various effects of it.**

***Keywords*— Data mining, Frequent Itemset Hiding, Association Rule Hiding, Sensitivity.**

I. Introduction

In recent years, data mining or knowledge discovery techniques such as association rule mining, classification, clustering, sequence mining etc. have been most widely used in information world. Successful application of data mining have been demonstrated in marketing, medical analysis, business, bioinformatics, product control and some other areas that benefit commercial, social and human activities. Along with the importance and successfulness of these techniques, they also pose a threat to data privacy, in a way that one (e.g. adversary or malicious user) is intentionally able to infer sensitive information or knowledge including customer's behaviour, business policies, marketing strategies etc. by using these techniques. So, before releasing database sensitive information or knowledge must be protected from unauthorized access. Therefore to solve privacy problem privacy preserving data mining (PPDM) has become a hotspot in data mining and database security field.

Association rule mining technique is widely used in data mining to find relationship between item sets. Many organizations disclose their information or database for mutual benefit to find some useful information for some decision making purpose and improve their business schemes. But this database may contain some private data and which the organization does not want to disclose. The issue of privacy plays important role when several organizations share their data for mutual benefit but no one want to disclose their private data. Therefore before disclosing the database, sensitive patterns must be hidden and to solve this issue PPDM techniques are helpful to enhance the security of database.

Motivation example shows importance of sensitive patterns in business applications. Let a clothes store that purchase jeans from two companies, ABC and CDE, and both can access customers' database of the store. Now ABC applies data mining techniques and mines association rules related to CDE's products. ABC had found that most of the customer who buy jeans of the CDE also buy belt. Now ABC offers some discount on belt if customer purchases ABC's jeans. As result the business of CDE goes down. So releasing the database with sensitive information cause the problem. This scenario gives the direction to research on sensitive rules (or knowledge) hiding in database.

The proposed algorithm is the improved version of DSRRC [3]. DSRRC could not hide association rules with multiple items in antecedent (L.H.S) and consequent (R.H.S.) and automatic sensitive rule generation. To overcome this limitation we have proposed an algorithm ADSRRC which uses count of items in consequent of the sensitive rules. It modifies the minimum number of transactions to hide maximum sensitive rules and maintain data quality.

Problem Descriptions

Association rule hiding problem can be defined as : convert the original database into sanitized database so that data mining techniques will not be able to mine sensitive rules from the database while all non sensitive rules remain visible.

A general definition of problem can be given as: Given transnational database D, Minimum confidence, Minimum support, and generated set of association rules R from D, a subset SR of R as sensitive rules, which database owner want to hide. Problem is to find the sanitized database D' such that when mining technique is applied on the D', all sensitive rules in set SR will be hidden while all non sensitive rules can be mined.

The aim of association rule hiding is to satisfy the following conditions

1) Sanitized database must not reveal any sensitive rules.

2) Sanitized database must facilitate mining of all non sensitive rules.

3) Sanitized database must not introduce any new rule, not present in D.

The problem of finding an optimized sanitized database, which satisfies all these conditions has been proved as NP-hard in [1].

The structure of remaining paper is as follow: Section II describes a literature review of existing approaches and some theoretical background. Section III describes proposed ADSRRC algorithm in detail. Section IV presents an example of ADSRRC algorithm. Section V elaborates the performance results and compared it with existing algorithm DSRRC [3]. At last section VI concludes our work and gives direction for future work.

II. LITERATURE REVIEW AND THEORETICAL BACKGROUND

This section provides overall idea for association rule mining and goal for sensitive association rule hiding. After that we will discuss and classify existing approaches for hiding sensitive knowledge. This section is written according to literature survey done by us while hiding sensitive knowledge application.

Association rule generation is done by the Apriori algorithm as in [3].Association rule using support and confidence can be defined as follows. Let I= {i1,…,im}be a set of items. Database D={T1,….,Tn}is a set of transactions, where Ti⊆I(1⩽i⩽m). Each transaction T is an itemset such that T⊆I. A transaction T supports X, a set of items in I, if X⊆I. The association rule is an implication formula like X⇒Y, where X⊂I, Y⊂I and X∩Y=Ø. The rule with support s and confidence c is called, if |X∪Y|/|D| ⩾ s and |X∪Y|/|X| ⩾ c. Because of interestingness, we consider user specified thresholds for support and confidence, MST (minimum support threshold) and MCT (minimum confidence threshold). A detailed overview of association rule mining algorithms are presented in [3].

Many approaches have been proposed to preserve privacy for sensitive knowledge or sensitive association rules in database. They can be classified in to following classes: heuristic based approaches, border based approaches, exact approaches, reconstruction based approaches, and cryptography based approaches. In following, a detailed overview of these approaches is given.

A. Heuristic Based Approaches

These approaches can be further divided in to two groups based on data modification techniques: data distortion techniques and data blocking techniques.

Data distortion techniques try to hide association rules by decreasing or increasing support (or confidence). To increase or decrease support (or confidence), they replace 0's by 1's or vice versa in selected transactions. So they can be used to address the complexity issue. But they produce undesirable side effects in the new database, which lead them to suboptimal solution. M.Attallah et al. [1] we the first proposed heuristic algorithms. The proof of NP-hardness of optimal sanitization is also given in [1]. Verykios et al. [11] proposed five assumptions which are used to hide sensitive knowledge in database by reducing support or confidence of sensitive rules. Y-H Wu et al. [14] proposed method to reduce the side effects in sanitized database, which are produced by other approaches [11].

Data blocking techniques replace the 0's and 1's by unknowns ("?") in selected transaction instead of inserting or deleting items. So it is difficult for an adversary to know the value behind "?". Y.Saygin et al. [13] were the first to introduce blocking based technique for sensitive rule hiding. The safety margin is also introduced in [13] to show how much below the minimum threshold new support and confidence of a sensitive rule should.

B. Border Based Approaches[12]

Border based approaches use the notion of borders presented in [12]. These approaches preprocess the sensitive rules so that minimum numbers of rules are given as input to hiding process. So, they maintain database quality while minimizing side effects. Sun and Yu [12] were the first to propose the border revision process. Hiding process in [10] greedily selects those modifications that lead to minimal side effects. The authors in [14] presented more efficient algorithms than other similar approaches presented in [12].

C. Exact Approaches

Exact approaches formulate hiding problem to constraint satisfaction problem (CSP) and solve it by using binary integer programming (BIP). They provide an exact (optimal) solution that satisfies all the constraints. However if there is no exact solution exists in database, some of the constraint are relaxed. These approaches provide better solution than other approaches. But they suffer from high time complexity to CSP. Gkoulalas and Verykios [5] proposed an approach to find optimal solution for rule hiding problem. The authors in [15] proposed a partitioning approach for the scalability of the algorithm.

D. Reconstruction Based Approaches[10]

Reconstruction based approaches generate privacy aware database by extracting sensitive characteristics from the original database. These approaches generate lesser side effects in database than heuristic approaches.

E. Cryptography Based Approaches[11]

Cryptography based approaches used in multiparty computation. If the database of one organization is distributed among several sites, then secure computation is needed between them. These approaches encrypt original database instead of distorting it for sharing. So they provide input privacy. Vaidya and Clifton [11] proposed a secure approach for sharing association rules when data are vertically partitioned.

We proposed a more efficient heuristic algorithm than other heuristic approaches presented in this section.

DSRRC algorithm is presented in [3]. The disadvantage of the algorithm presented in [3] is as under

I. It can only hide rules with single item in the antecedent and consequent.

II. User has to select which rule is sensitive.

To overcome this limitation of the DSRRC algorithm we presented the Advance DSRRC algorithm in next section.

III. Proposed ADSRRC Algorithm

In order to hide the sensitive rule like X→Y , we can decrease either confidence or support of the rule below the user specified minimum threshold. To decrease the confidence of the rule, we can choose two methods like (1) increase the support of X(L.H.S. of the sensitive rule) but not support of X ∪ Y ,or (2) decrease the support of X ∪ Y by decreasing support of Y(R.H.S of the sensitive rule) because it decrease the confidence of the rule faster than simply decreasing the support of X ∪ Y. Proposed algorithm hides rules with multiple items in L.H.S and multiple items in R.H.S. So the rule is like aX→bY where a,b∈ I and X,Y⊂ I. Here b is an item selected by proposed algorithm to decrease the support of the R.H.S. and decrease the confidence of the rule below MCT. We replace '1' to '0' in some transaction to decrease the support of selected items.

Some important definitions of terms are use in the proposed algorithm is as follow:

- Sensitivity of item: - is the quantity of data item exists in sensitive association rule set or the number of the sensitive association rule that are associated with this data item.
- Sensitivity of rule:- the sum of the sensitivities of each item of association rule is the sensitivities of the association rule.
- Sensitive transaction:- Transaction that contains sensitive items is called sensitive transaction.
- IS={*is0*, is1...isk} k≤n, Set of items present in R.H.S of sensitive rules with decreasing order of their frequency in R.H.S of sensitive rules.
- *is0* = Item with highest count in R.H.S of sensitive rules.

The proposed algorithm starts with mining the association rule from the original database D using association rule mining algorithm e.g. Apriori algorithm [4]. Then user specifies some items as sensitive and based on that sensitive rules (SR) are generated by the association rule mining algorithm. Then algorithm counts occurrences of each item in R.H.S of sensitive rules. Now algorithm finds IS={is0; is1:::isk} k≤n, by arranging those items in decreasing order of their counts. After that sensitivity of each item is calculated then sensitivity of each transaction is calculated. Then transactions which support is0 are sorted in descending order of their sensitivities.

Now rule hiding process starts by selecting first transaction from the sorted transactions with higher sensitivity, delete item is0 from that transaction. Then update support and confidence of all sensitive rules and if any rule have support and confidence below MST and MCT respectively then delete it from SR. Then update sensitivity of each item ,transaction and IS. Again select transaction with higher sensitivity and delete is0 from it. This process continue until all sensitive rules are hidden.

As a result, modified transactions are updated in the original database and new database is generated which is called sanitized database D', which preserves the privacy of sensitive information and maintains database quality.

Proposed algorithm MDSRRC is shown below, which is used to hide the sensitive rules from database. Given a database D, MCT (minimum confidence threshold) and MST(minimum support threshold) algorithm generates sanitized database D'. Sanitized database hides all sensitive rules and maintains data quality.

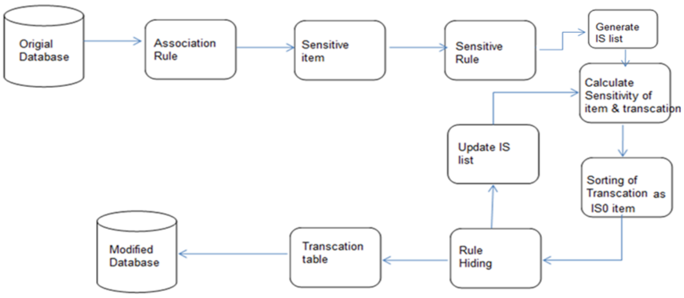


Figure 1. Framework of ADSRRC algorithm.

ADSRRC Algorithm.

INPUT:

MCT (Minimum Confidence Threshold), Original database D, MST(Minimum support threshold).

OUTPUT:

Sanitized Database D' with all sensitive rules hidden.

```
1.  Apply Apriori algorithm [3] on given database D. Generate all possible association rules R.
2.  Select the Sensitive item from the list.
3.  Generate set of rules SR⊂R as sensitive rules.
4.  Calculate sensitivity of each item j ∈ D.
5.  Calculate sensitivity of each Transaction.
6.  Count occurrences of each item in R.H.S of sensitive rules, find IS={is0,is1⋯isk} k≤n, by arranging those items in descending order of their count. If two items have same count then sort them in descending order of their actual support counts.
7.  Select the transactions which supports is0, then sort them in descending order of their sensitivity. If two transactions have same sensitivity then sort them in increasing order of their lengths.
8.  While(SR is not empty)
9.  {
10.     Start with first transaction from sorted transactions,
11.     Delete item is0 from that transaction.
12.         For each rule r ∈ SR
13.         {
14.             Update support and confidence of the rule r.
15.     If(support of r < MST or confidence of r < MCT)
16.     {
17.             Delete Rule r from SR.
18.             Update sensitivity of each item.
19.             Update IS (This may change is0).
20.             Update the sensitivity of each transaction.
21.             Select the transactions which supports is0.
22.             Sort those transactions in descending order
                of their sensitivities.
23.     }
24.     Else
25.     {
26.             Take next transaction from sorted
transactions, go to step 11.
27.     }
28.     }
29.     }
30.     End
```

ADSRRC select best items so that deleting those items hide maximum rules from database to maintain data quality.

IV. Example

To understand the ADSRRC algorithm following example is illustrated. In Table I transnational database D is shown. With 3 as MST, 40% as MCT and ‘a’,’d’,’e’ as the sensitive item the possible generated association rule as shown in table IV. Then we calculate the sensitivity of each item which is present in sensitive rule based on that we calculate the transactions sensitivity as shown in table V. then we create the IS list and sort it. Based on the *IS0* item we select the transactions from which we can delete the item as shown in fig-1.

Table - I
Sample Database

TID	Items
1	a b c d e
2	a c d
3	a b d f g
4	b c d e
5	a b d
6	c d e f h
7	a b c g
8	a c d e
9	a c d h

Table - II
Itemsets with support count ≥3

Item sets with Support count 3 or more
a:7,b:5,c:7,d:8,e:4, abd:3, acd:4, cde:4, ab:4,ac:5,bc:3, ad:6, bd:4, cd:6. ce:4, de:4,abd:3, acd:4, cde:4

Table - III
Selected Sensitive Item

Selected Sensitive Item
a d e

Table - IV
Generated Sensitive rue with Confidance ≥ 40%

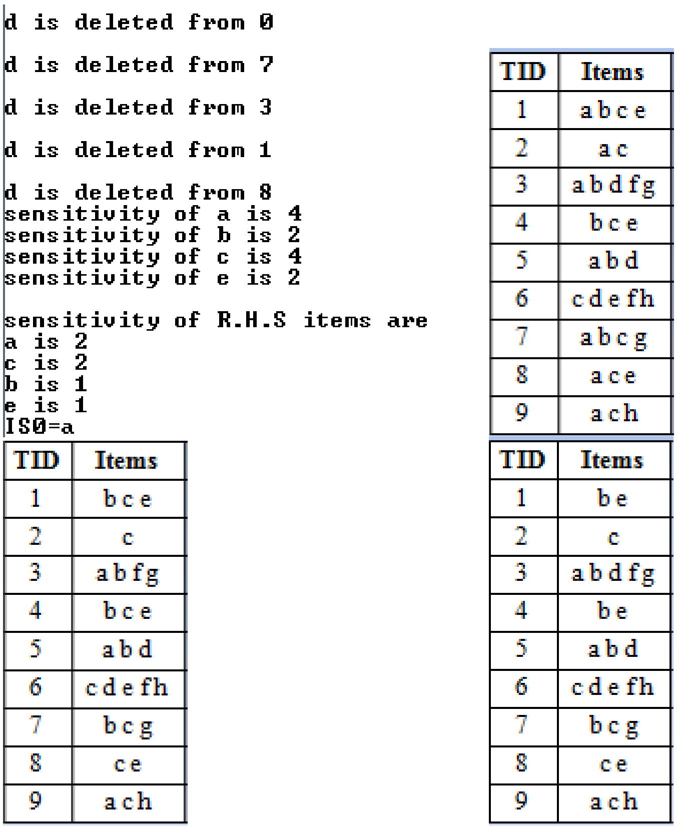
Sensitive	Confidence(%)	Sensitive	Confidence(%)
Rule		Rule	
a – b	57.14	a – bd	42.85
b – a	80	b – ad	60
a – c	71.42	ab –d	75
c – a	71.42	ad – b	50
a – d	85.71	bd – a	75
d – a	75	a – cd	57.14
b – d	80	c – ad	57.14
d – b	50	ac – d	80
c – d	85.71	d - ac	50
d - c	75	ad - c	66.66
c - e	57.14	cd - a	66.66
e - c	100	c - de	57.14
d - e	50	d - ce	50
e - d	100	cd - e	66.66
ce – d ,de - c	100 , 100	e – cd	100

Table - V
Sensitivity of Item

Sensitive item	Sensitivity of Item
a	17
b	9
c	18
d	25
e	10

Table - VI
Sorted IS list

Sensitive RHS item	Sensitivity of RHS Item
d	13
c	9
a	8
e	5
b	4



```
d is deleted from 0

d is deleted from 7

d is deleted from 3

d is deleted from 1

d is deleted from 8
sensitivity of a is 4
sensitivity of b is 2
sensitivity of c is 4
sensitivity of e is 2

sensitivity of R.H.S items are
a is 2
c is 2
b is 1
e is 1
IS0=a
```

TID	Items
1	abce
2	ac
3	abdfg
4	bce
5	abd
6	cdefh
7	abcg
8	ace
9	ach

TID	Items
1	bce
2	c
3	abfg
4	bce
5	abd
6	cdefh
7	bcg
8	ce
9	ach

TID	Items
1	be
2	c
3	abdfg
4	be
5	abd
6	cdefh
7	bcg
8	ce
9	ach

Fig 1 : *IS0=d* deleted form the transaction updated main transaction and updated IS list with new *IS0=a*. Similarly update the transaction and IS all the Sensitive rule is hidden and the final Sanitized database D' is obtained which is as show in the last table

The algorithm works as follows first of all it ask for the MCT and MST and then ask for the Sensitive item then it calculate the sensitivity of transactions and items and generate the IS list then sort the transaction based on *IS0* item and delete the *IS0* items from the selected transactions until all the rule having item *IS0* is hidden then update the IS list again and do the same procedure again until all sensitive rule not hidden.

As result ADSRRC generate sanitized database which hides all sensitive rules and maintain data quality.

V. EXPERIMENTAL RESULT AND ANALYSIS OF PROPOSED ALGORITHM

As we cannot compare the result of the DSRRC and ADSRRC because in ADSRRC has capability to generate the sensitive rule automatically and therefore it have more rule to hide so database modification also be very large but for comparison purpose we select the sensitive rule manually and then hide it through DSRRC and ADSRRC and compared the result.

We used algorithm DSRRC and our proposed algorithm (ADSRRC) to hide the three sensitive rules on sample database, as shown in Table I. After applying Apriori algorithm with 3 as MST and 40% as MCT, we select 3 sensitive rules(a-bd, a-cd, d-ac) from the generated rules. After applying both algorithms on sample database we have done evaluation by considering the performance parameters which are given in [16] viz. (a) HF (hiding failure), (b) MC (misses cost), (c) AP (artificial patterns), (d) DISS (dissimilarity) and (e) SEF (side effect factor). Experimental results show that ADSRRC increase efficiency and reduce modification of transactions in database.

Table - VII
Perfomance Results

Parameter	DSRRC[3]	ADSRRC(ProposedAlgo)
HF	0%	0%
MC	36%	26.5%
AP	0%	0%
DISS(D,D')	6.4%	5.3%
SEF	36.5%	26.5%

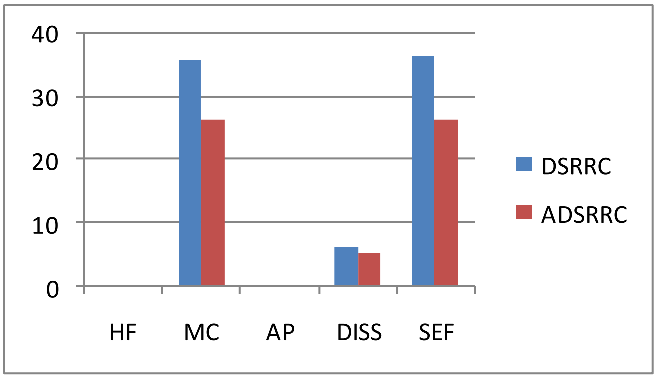


Fig 2 : Performance Comparisons of DSRRC and ADSRRC

VI. CONCLUSION AND FUTURE SCOPE

We proposed an algorithm named ADSRRC which hides sensitive association rules with fewer modifications on database to maintain data quality and to reduce the side effect on sanitized database. Functionality of proposed algorithm is shown using sample database with three sensitive rules. Experimental results show that proposed algorithm works better than DSRRC.

In future, ADSRRC algorithm can be extended to increase the efficiency and reduce the side effects by minimizing the modifications on database.

REFERENCES

[1] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios "Disclosure limitation of sensitive rules," In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pp. 45–52, 1999.

[2] C. N. Modi, U. P. Rao, and D. R. Patel, "Maintaining privacy and data quality in privacy preserving association rule mining," 2010 Second International conference on Computing, Communication and Networking Technologies, pp. 1–6, Jul. 2010.

[3] Han Jiawei and Kamber, Micheline. "Data Mining: Concepts and Techniques". 2006. Morgon Kaufmann. Sanfransico, CA.

[4] X. Sun and P.S. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM '05), pp. 426–433, Nov. 2005.

[5] A. Gkoulalas-Divanis and V.S. Verykios, “An Integer Programming Approach for Frequent Itemset Hiding,” In Proc. ACM Conf. Information and Knowledge Management (CIKM ’06), Nov. 2006.

[6] Charu C. Aggarwal, Philip S. Yu, Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company Incorporated, 2008, pp. 267-286.

[7] Yi-Hung Wu, Chia-Ming Chiang and Chen A.L.P., “Hiding Sensitive Association Rules with Limited Side Effects,” IEEE Transactions on Knowledge and Data Engineering, Vol.19, no. 1, Jan. 2007.

[8] Vassilios S. Verykios, Elisa Bertino, and Igor Nai Fovino, “State-of-the-art in Privacy Preserving Data Mining,” ACM SIGMOD, Vol. 33, no. 1, March 2004.

[9] A. Gkoulalas-Divanis and V.S. Verykios, “Exact Knowledge Hiding through Database Extension,” IEEE Transactions on Knowledge and Data Engineering, vol. 21(5), pp. 699–713, May 2009.

[10] Y. Guo, “Reconstruction-Based Association Rule Hiding,” In Proc. Of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 2007.

[11] J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data,” In proc. Int’l Conf. Knowledge Discovery and Data Mining, pp. 639–644, July 2002.

[12] X. Sun and P.S. Yu, “A Border-Based Approach for Hiding Sensitive Frequent Itemsets,” In Proc. Fifth IEEE Int’l Conf. Data Mining (ICDM ’05), pp. 426–433, Nov. 2005.

[13] Y.Saygin, V. S. Verykios, and C. Clifton, “Using Unknowns to Prevent Discovery of Association Rules,” ACM SIGMOD, vol.30(4), pp. 45–54, Dec. 2001.

[14] Moustakides and V.S. Verykios, “A Max-Min Approach for Hiding Frequent Itemsets,” In Proc. Sixth IEEE Int’l Conf. Data Mining (IC

[15] DM ’06), pp. 502–506, April 2006.

[16] A. Gkoulalas-Divanis and V.S. Verykios, “Exact Knowledge Hiding through Database Extension,” IEEE Transactions on Knowledge and Data Engineering, vol. 21(5), pp. 699–713, May 2009.

[17] V. Verykios and A. Gkoulalas-Divanis, A Survey of Association Rule Hiding Methods for Privacy, ser. Advances in Database Systems. Springer US, 2008, vol. 34.