# Privacy Preserving Big Data Mining: Association Rule Hiding

Golnar Assadat Afzali*
Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
g.afzali@gmail.com
Shahriar Mohammadi
Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
mohammadi@kntu.ac.ir

Received: 05/Apr/2016 Revised: 16/May/2016 Accepted: 06/Jun/2016

**Abstract**

Data repositories contain sensitive information which must be protected from unauthorized access. Existing data mining techniques can be considered as a privacy threat to sensitive data. Association rule mining is one of the utmost data mining techniques which tries to cover relationships between seemingly unrelated data in a data base.. Association rule hiding is a research area in privacy preserving data mining (PPDM) which addresses a solution for hiding sensitive rules within the data problem. Many researches have be done in this area, but most of them focus on reducing undesired side effect of deleting sensitive association rules in static databases. However, in the age of big data, we confront with dynamic data bases with new data entrance at any time. So, most of existing techniques would not be practical and must be updated in order to be appropriate for these huge volume data bases. In this paper, data anonymization technique is used for association rule hiding, while parallelization and scalability features are also embedded in the proposed model, in order to speed up big data mining process. In this way, instead of removing some instances of an existing important association rule, generalization is used to anonymize items in appropriate level. So, if necessary, we can update important association rules based on the new data entrances. We have conducted some experiments using three datasets in order to evaluate performance of the proposed model in comparison with Max-Min2 and HSCRIL. Experimental results show that the information loss of the proposed model is less than existing researches in this area and this model can be executed in a parallel manner for less execution time

**Keywords:** Big Data; Association Rule; Privacy Preserving; Anonymization; Data Mining.

## 1. Introduction

Data mining is the process of extracting hidden but useful knowledge from large data bases [1]. Nowadays different sources are creating data with high speed [2]. Distributed infrastructures such as cloud computing present the opportunity to store large volume data bases for further analysis and knowledge discovery.

Big data mining is the capability of extracting desired information from large data bases or data streams [3]. Association rule mining is one of the most important data mining techniques. However, misuse of this technique may lead to disclosure of sensitive information about users [4,5]. Many algorithms have been proposed in the literature for rule hiding [6,7,8,9,10]. Most of them are based on the idea of modifying main data base to decrease the support or confidence value of sensitive association rules. The main drawback of existing works is the undesired side effect of removing some item-set on non-sensitive association rules.

In this paper, we use anonymization techniques as an alternative for removing some repeated instance of frequent item-sets. The main idea of the proposed model is that removing frequent item-sets (which is used in existing related works) has undesired side effect on new entrance data. But, by using data anonymization any necessary change can be applied to existing anonymization level to support this new data. In other word, it is possible to change (increase or decrease) the anonymization level by new data entrance. As in big data mining, we deal with dynamic datasets, with any new data entrance, association rules can change. So, the proposed model we replace removing some instance of association rules with rule anonymity.

The remainder of this paper is organized as follow: next section reviews the related works. In section III, the proposed approach for big data association rule hiding is described. Experimental results of comparing the performance of our approach with previous works are described in section IV. At last, section V concludes this paper.

### 1.1 Related Work

#### A. Big Data

In term of definition, big data refers to high volume of structured, semi-structured and unstructured data with high velocity which can be mined for information [3]. Big data mining refers to the capability of extracting information from massive datasets that due to specific features cannot be done using existing data mining techniques [1].

In many situations, it is infeasible to store this huge amount of data, so the knowledge extraction should be

* Corresponding Author

done real-time. For processing big data, a cluster of computers with high computing performance is needed and this framework would be practical with paralleling tools such as MapReduce [11].

### B. Anonymity

Information dissemination is usually with the risk of sensitive information disclosure [12]. Data usually contain sensitive information and this proves the importance of employing anonymity approaches [12,13]. Generally, there are three techniques for anonymization which include generalization, suppression and randomization. Different approaches for anonymization such as k-anonymity, l-diversity, t-closeness and etc. use these techniques.

In generalization, values of attributes are replaced with a more general one [14]. For example, if the value of attribute "age" is equal to 16, it can be replaced with appropriate range such as (10-20).

Suppression refers to stop releasing the real value of an attribute. In this way, occurrence of the value is replaced with a notation such as "*", this means that any value can be replaced instead [15]. For example, if the related value of an attribute is equal to 56497, it can be replaced with 5649*.

Randomization refers to substitution of real value with a random value. In this technique, noise is added to data so that real value of attributes is masked [16].

In this paper proposed model generalization technique is used for anonymity, while suppression technique is not suitable for quantitative data because in quantitative data we cannot substitute some parts of the data with "*". A secure randomization technique needs a defined and not reversible function. Therefore for each data, related assigned noise should be saved to make it possible to retrieve the real value, if necessary. Therefore, this technique imposes significant overhead to systems.

### C. Association Rule Hiding

Association rule mining is an interesting approach to find out unknown relations between variables in large databases [6]. However, misuse of these techniques may cause disclosure of sensitive information [17]. So, many researchers worked on hiding sensitive association rules. The main purpose of association rule hiding approaches is to hide sensitive rules, without any side effect on non-sensitive rules. Le et al. [8] proposed HSCRIL model as a heuristic approach to hide a set of association rules from relational databases in retail industry. The main steps of their proposed algorithm are: identification of victim items that their modifications have least impact on other frequent item-sets, determination of minimum number of transactions which should be modified, and removing victim items from specified transactions. In this research, generation set of frequent item-sets is maintained. This generation set causes least impact on non-sensitive item-sets during sensitive rule hiding. The main result of this model is an acceptable information loss. But, this model is based on determined generation sets; however, in big data mining, with new data entrance, generation sets will change. So, this idea cannot be applicable for big data.

Max-Min2 model of Moustakides and Verykios [18], used Max-Min theorem in association rule hiding. The main idea of this theorem is to maximize the minimum gain. In fact, they are trying to maximize sensitive rule hiding while at the same time minimize the side effect on non-sensitive rules. This model hides sensitive association rules by decreasing the support of sensitive item-sets. Results of this research show that the information loss of this model is less than existing related works. This model tried to hide sensitive association rules by removing some instances of them. However, in dynamic data sets, sensitive association rules will change with new data entrances and removing them cannot be a good idea.

Wang et al. in [19] proposed a model in which two algorithms are used to hide sensitive association rules. They used ISL (Increase support of left hand side) and DSR (decrease support of right hand side) to achieve their purpose. Removing sensitive association rules by these two algorithms causes mentioned problems of Max-Min2 algorithm.

In the research of [20] by Wang et al., all existence transactions are represented in the form of a binary matrix. In this matrix, if item i participates in transaction j, Dij will be 1, otherwise it is equal to 0. Then, based on the defined threshold of support value in this system, matrix S would be determined so that D'= S *D. In this definition, D is the matrix related to the main database, S is the hiding matrix, and D' is the matrix related to hidden database. Based on the "volume" feature of big data, defining related matrix is time consuming and needs high storage capacity.

Dasseni et al. [21] considered the hiding of both sensitive association rules and frequent item-sets. They develop three strategies for this purpose: Increasing the support of left hand side (LHS), decreasing the support of right hand side (RHS), and decreasing the support of right and left hand sides, simultaneously. In this paper, three strategies cause less undesired effect on non-sensitive association rules. However, main disadvantage of this model is similar to Max-Min2.

Jung et al. [22] use Hadoop for association rule hiding in large scale datasets. Privacy threats which are considered in this paper are related to the flow of data to untrusted cloud service providers. So, at the first step, association rules are determined. Then, some noises are added to the item-sets to prevent frequent item-set disclosure of them. This model can prevent exposure of sensitive data without data utility degradation, but adding noise to data causes endures computing cost to systems and is not suitable for big data mining and real time processing.

Xu et al [23] concentrate on information security on big data analysis. They identify four types of users involved in data mining application. Namely, data provider, data collector, data miner and decision maker. For each group, security threats are defined and appropriate solutions considered, too.

In this paper's proposed model, anonymization technique is used to hide sensitive information. So, at first,

two criteria are defined to select best item-set(s) for anonymization. Then, based on these two criteria, in any sensitive association rule, the item-set with the least undesired side effect is selected in order to be hidden.

For selected item-set(s), quasi-identifier attributes are anonymized in appropriate level. In other word, in this proposed model, none of repeated item-sets would be removed from database and only sensitive values would be hidden. So, if with new data entrance, each association rule changes between "sensitive" and "non-sensitive" mode, only by changing anonymity level, main purpose which is retrieving related information or hiding information, would be acquired.

Table 1 summarizes these related works.

Table 1. summary of related work

| Author(s) | Method | Information hiding | Association rule hiding |
|---|---|---|---|
| [8] Le et al. | - Identification of informative items<br>- Specification of generation set related to frequent item-sets.<br>- Removing determined item-set form database | No | Yes |
| [18] Moustakides & Verykios | - Using Max-Min theorem to maximum information hiding with minimum side effect.<br>- Reducing support of frequent item-sets to less than defined threshold. | No | Yes |
| [19] Wang et al. | - Decreasing support value of frequent item-sets to less than defined threshold.<br>- Decreasing confidence value of frequent item-sets to less than defined threshold.<br>- Utilizing ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) functions to achieve mentioned purposes. | No | Yes |
| [20] Wang et al. | - Binary indicator matrix of items in transactions, named as D.<br>- Hiding matrix S is determined based on the defined threshold for support.<br>- Matrix D' related to hidden data set, is determined based on S and D | No | Yes |
| [21] Dasseni et al. | - Sensitive information hiding besides association rule hiding.<br>- Increasing the support value of LHS (Left Hand Side).<br>- Decreasing the support value of RHS (Right Hand Side).<br>- Decreasing the support of RHS (Right Hand Side) and LHS (Left Hand Side), simultaneously. | Yes | Yes |
| [22] Jung et al. | - Determine sensitive association rules.<br>- Adding noise to sensitive association rule to hide them from undefined user, without significant information loss. | No | Yes |
| Proposed model | - Sensitive information hiding besides association rule hiding.<br>- Using anonymity approach for hiding sensitive association rules. | Yes | Yes |

As mentioned below, as in the proposed model, anonymity technique is used instead of removing instances of association rules, if with any new data entrance, each association rule changes from sensitive to non-sensitive (or vice versa), we can update dataset easily. This feature makes the proposed model appropriate for dynamic datasets. While in all of existing models, authors only has concentrated on static datasets.

## 2. Proposed Model for Big Data Association Rule Hiding

As mentioned, association rules should not be disclosed since they may be used to infer sensitive information. Many researches have done in association rule hiding which most of them have significant drawbacks:

- Undesired side effect of hiding sensitive association rules on non-sensitive rules.
- The impossibility of using in big data analysis.

To solve these mentioned problems, anonymity techniques could be used for rule hiding as an alternative for deleting some of the most repeated items. In this paper two criteria are defined in order to support new data entrance in our big data base which are represented below. It is notable that features such as parallelization and scalability which considered in this model make it suitable for big data analysis.

The proposed model consists of three main steps which are described in follow:

**Step 1: Association Rule Mining**

There are many association rule mining algorithms such as Apriori [24] or FP-growth [25]. Let $\alpha(H)$ be the support value of item-set H, this item-set is called frequent item-set if $\alpha(H)>\sigma$ , which $\sigma$ is the defined support threshold. An association rule A->B is considered as a sensitive association rule if $\alpha(A->B) >= \sigma$ and

$\beta(A->B)>=\delta$, which $\beta(X)$ refers to the confidence value of this rule and $\delta$ is the defined confidence threshold.

**Step 2: Best Item-set Selection**

Because of the velocity feature of big data, selection of the best item(s) for anonymization should be done based on the two criteria:

- Undesired side effect of anonymization on other existing non-sensitive association rules.
- Undesired side effect of anonymization on probable new entrance data.

The Best approach is to decrease these values as much as possible.

Suppose that we want to hide a rule such as A->B. The main problem is to determine the best item-set for anonymization. For this, anonymization effect of each right or left hand side item should be evaluated based on the two mentioned criteria and then, the item with the least side effect is selected.

At first, Association rules are sorted based on their confidence value. Then, these factors are used for the best item selection.

The First criterion has a static view on data set (without new data entrance). So, information loss which is caused by this anonymization could be computed with formula presented in (1).

$$\text{InfoLoss} = \frac{Ni}{Ni+Nj} \quad (1)$$

In formula (1), Ni is the number of non-sensitive association rules which A is involved in, while Nj is the number of sensitive association rules which A is involved in.

In the second criterion, we have dynamic view on our data set. In this manner, the best item is one which has greater chance to convert related non-sensitive association rules to sensitive association rule. This can cause lower information loss. So, the difference between defined confidence threshold and confidence value of existing non-sensitive association rules can be considered as the second criterion for the best item selection. This measure can be evaluated based on the formula presented in (2).

$$\text{DoC} = \sqrt{\sum_{i=1}^{n}(Ci - CLj)^2} \quad (2)$$

In (2), Ci is the confidence value of i'th non-sensitive association rule which A is involved in, while CLj is the defined confidence threshold.

Finally, the best item selection could be done by combining InfoLoss and DoC values, but with appropriate effective weight, as (3);

$$BI = \alpha1 * \text{InfoLoss} + \alpha2 * \text{DoC} \quad (3)$$

The item with less BI value could be selected as the best item for anonymization. In (3), $\alpha1$ and $\alpha2$ are effective weights and their values can be changed based on the importance ratio of related criterions in each specific context. By default, $\alpha1$ and $\alpha2$ have same value and are equal to 0.5.

**Step 3: Data Anonymization**

As mentioned, generalization technique is used as the proposed anonymization technique. Attributes of each item-set can be classified in three categories: identifier attributes are attributes containing identifying information such as Social Security Number (SSN); sensitive attributes are set of attributes that contain personal privacy information and should be protected; quasi-identifier (QI) attributes are attributes that do not contain identifying attributes, but can be linked to other information to cause identification disclosure [8].

So, in this model, after selecting the best item-set for anonymization, exact value of sensitive and identifier attributes would be removed and then quasi-identifier attributes of this item-set would be generalized to an acceptable level.

The pseudo code of the proposed model is shown in figure 1.

```
Initialize list of sensitive association rules
While sensitive association rules lists is not empty
{
Sort sensitive association rules in decreasing order of confidence value
Select the association rule with the maximum confidence value
While selected association rule is not anonymized yet
  {
  Calculate InfoLoss and DoC for each item-set of association rule
  Calculate the BI for each item-set
  Choose the item-set with maximum BI value
  Anonymize selected item-set
  Remove this association rule from sensitive association rules
  }
}
```

## 3. Parallelization of the Proposed Method for Big Data Mining

As said before, in order to facilitate the implementation of the proposed model for big data processing, features such as parallelism should be considered in this model. Distributed computing infrastructures, such as cloud computing, can provide the required infrastructure for this purpose. Now, it is required that besides considering tree structure for our database, as shown in figure 2, basic operations such as association rule mining to be done in a distributed and parallel manner.

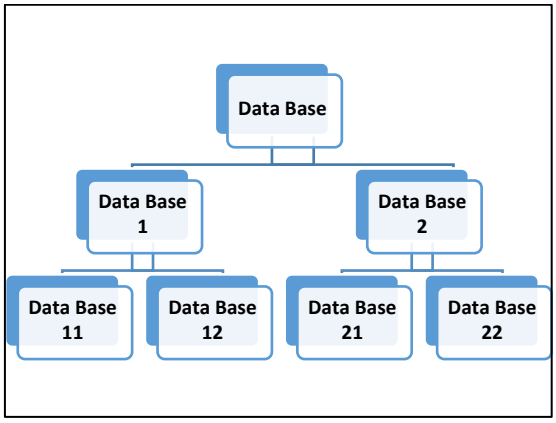



Fig. 2. proposed hierarchical structure for big data base

In addition to defining a tree structured data set, proper changes should be considered in the defined threshold of support and confidence values. Ideally, data set generator of each association rule is divided equally to slave nodes (nodes which are responsible for data storing and running the computations). In this manner, defined threshold of support and confidence values in each node is changed based on formula presented in (4).

$$\text{Threshold}_{new} = \text{Threshold}_{old} * \frac{1}{n} \quad (4)$$

In (4), thresholdold is the defined threshold for support and confidence parameters, n is the number of slave nodes (leaf nodes in hierarchical tree structure), and thresholdnew is the new defined threshold in each data node.

Normally, it is possible that the distribution of the main data set on data nodes would not be according to the mentioned ideal form. In this manner, if in any slave node, the computed support and confidence values of each association rule is higher than the new threshold, this rule may be a sensitive association rule. So, existence of this rule in other slave nodes should be checked and according to this information, this rule would be defined as a sensitive or non-sensitive association rule.

Appropriate tree structure for data set (as shown in figure 2) can facilitate scalability feature, too. In this structure, every node can extend the number of its children. Therefore, the number of slave nodes can be extended until required computing power reached.

## 4. Evaluation

In order to evaluate the proposed model performance, some experiments have been done and the results are compared with Max-Min2. Max-Min2 algorithm has gained better results in minimizing undesired side effect compared with other existing association rule hiding approaches.

### A. Dataset Description

Experiments have been done using three datasets. First dataset named Brijs dataset, contains market basket data from a Belgian retail supermarket store. Dataset contains 88162 transactions and 16469 product IDs.

Other two datasets are BMS-WebView-1 and BMS-WebView-2. These datasets are well-known datasets in association rule mining and contain click-straem data which are collected from two e-commerce web sites. The main goal of these two data sets are to determine the association between products which are viewed by visitors.

In order to increase the volume of datasets and make the dataset suitable for big data analysis, some instances of transactions are sampled randomly and repeated. Each database is divided into six partitions. Size of the first partition is equal to 500K and other partitions are added in next phases to this database in order to simulate the data stream feature of big data.

### B. Experiment Process

As mentioned above, the proposed model is compared with Max-Min2 and HSCRIL algorithms. Three metrics which are used for this comparison are: percentage of lost rules, ghost rules, and false rules, where

Lost rule: a non-sensitive association rule which are lost during association rule hiding process and are not in the released database [8].

Ghost rule: a non-sensitive association rule which cannot be mined form main database but can be mined from released database [8].

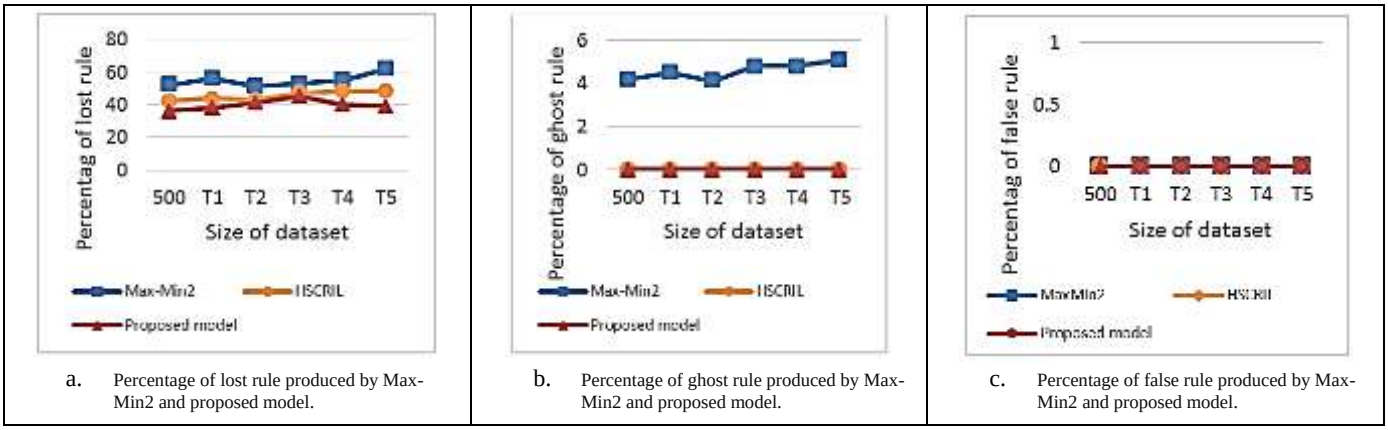
False rule: a sensitive association rule which cannot be hidden using the proposed association rule hiding process [8]. Figures 3,4 and 5 compare the performance of the Max-Min2, HSCRIL and the proposed model.

In these figures, part a, shows the lost rules of these models in each dataset, part b, shows the ghost rules of the these models and part c, is related to the false rules which are produced by them. As shown in figure 3.a, at first, number of lost rules in the proposed model is higher than Max-Min2 and HSCRIL models; but it starts to work better as new data arrives. The main reason is that these models have static view on database. For example, consider at time t1, rule A->B is considered as a sensitive association rule and the appropriate item-set is removed from some transactions in database. Now, if with the entrance of new data, the confidence value of this rule in the main database is decreased to be less than defined confidence threshold, this rule is a non-sensitive rule and should not be hidden. However, there is not the chance to retrieve this removed rule.

It should be noticed that another approach is to check the main database (which is not hidden) in order to retrieve such non-sensitive hidden rule. It is clear that because of the huge volume of data in big data mining, this approach is very time consuming and would be an impractical way.

Number of ghost rules produced by the proposed model is less than MaxMin2, in all of datasets.

Any of these models would not produce false rules. So, the percentage of false rule for all of them is equal to zero.

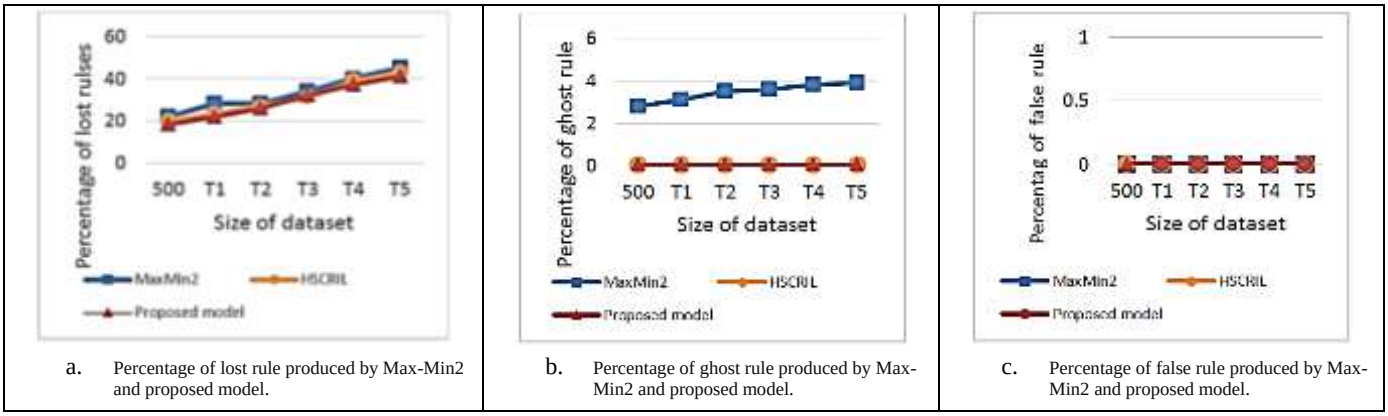



a. Percentage of lost rule produced by Max-Min2 and proposed model.

b. Percentage of ghost rule produced by Max-Min2 and proposed model.

c. Percentage of false rule produced by Max-Min2 and proposed model.

Fig. 3. comparison of the proposed model, HSCRIL and MaxMin2, Brijs dataset.

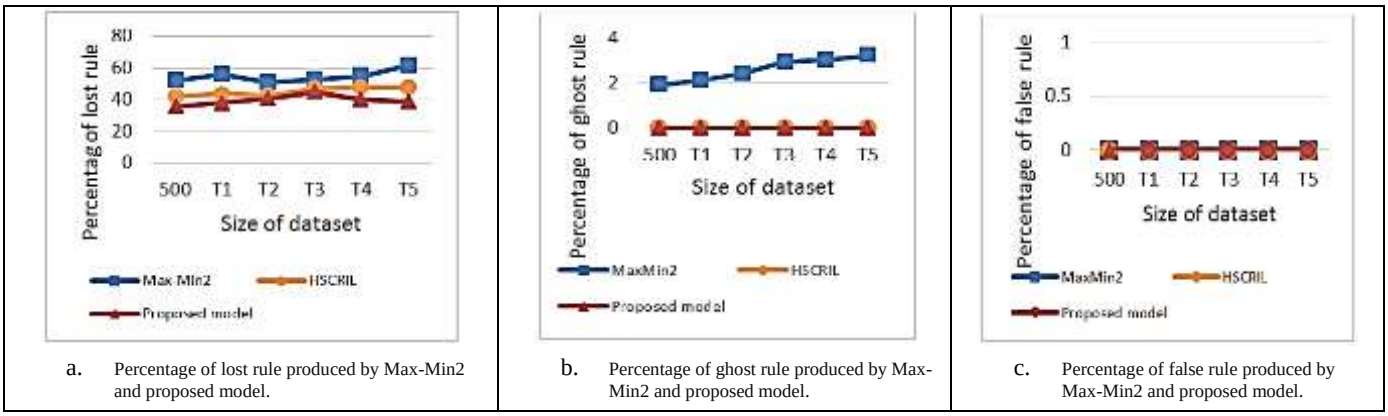



a. Percentage of lost rule produced by Max-Min2 and proposed model.

b. Percentage of ghost rule produced by Max-Min2 and proposed model.

c. Percentage of false rule produced by Max-Min2 and proposed model.

Fig. 4. comparison of the proposed model, HSCRIL and MaxMin2, BMS-WebView-1 dataset




a. Percentage of lost rule produced by Max-Min2 and proposed model.

b. Percentage of ghost rule produced by Max-Min2 and proposed model.

c. Percentage of false rule produced by Max-Min2 and proposed model.

Fig. 5. comparison of the proposed model, HSCRIL and MaxMin2, BMS-WebView-2 dataset.

Percentage of released lost rules, ghost rules and false rules in Brijs, BMS-WebView-1 and BMS-WebView-2 datasets are mentioned in table 2,3 and 4.

In order to evaluate the scalability and parallel processing capability features of the proposed model, multi thread processing is used to simulate the distributed computing infrastructure. Number of thread has been changed from 4 to 10. At each manner, defined threshold of the support and confidence values changed based on the number of threads and formula 4. Execution time of the proposed model is shown in figure 6. As shown in figure 6, as the number of the threads increases, execution time of the proposed model will decrease.

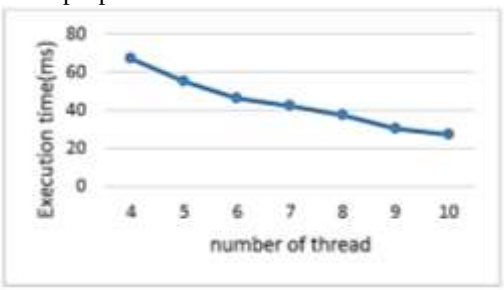



Fig. 6. execution time vs number of threads.

Table 2. Percentage of lost rule

| | Brijs dataset | | | | | | BMS-WebView-1 dataset | | | | | | BMS-WebView-2 dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Max-Min2 | 40 | 39 | 42 | 45 | 40 | 43 | 22 | 28 | 28 | 34 | 40 | 45 | 52 | 56 | 51 | 53 | 55 | 62 |
| HSCRIL | 33 | 34 | 37 | 31 | 29 | 31 | 20 | 23 | 27 | 39 | 39 | 43 | 42 | 44 | 43 | 47 | 48 | 48 |
| Proposed model | 30 | 26 | 28 | 25 | 22 | 29 | 18 | 22 | 26 | 32 | 37 | 41 | 36 | 38 | 41 | 45 | 40 | 39 |

Table 3. Percentage of ghost rule

| | Brijs dataset | | | | | | BMS-WebView-1 dataset | | | | | | BMS-WebView-2 dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Max-Min2 | 4.2 | 4.5 | 4.1 | 4.8 | 4.8 | 5.1 | 2.8 | 3.1 | 3.5 | 3.6 | 3.8 | 3.9 | 1.9 | 2.1 | 2.4 | 2.9 | 3 | 3.2 |
| HSCRIL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Proposed model | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 4. Percentage of false rule

| | Brijs dataset | | | | | | BMS-WebView-1 dataset | | | | | | BMS-WebView-2 dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Max-Min2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HSCRIL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Proposed model | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## 5. Conclusion

Association rule mining is a data mining technique which besides its benefits in discovering unclear relationships between data, will result privacy violation. Association rule hiding can help to protect sensitive association rules to be discovered. Many different techniques have been considered to hide sensitive association rules but most of them try to select item-sets and remove them, in order to decrease the confidence value of the related rule(s) to be less that the defined threshold. In this model, instead of removing some instances of the frequent item-sets, item-sets are assigned to appropriate anonymity level. None of existing approaches can be executed in a parallel and scalable manner, to be appropriate for big data mining. Besides, removing item-sets from the database can cause serious information loss as new data stream arrives. In this research, new big data association rule hiding technique is presented which tries to decrease undesired side effect of sensitive rule hiding on non- sensitive rules in data streams. Features such as parallelism and scalability are embedded in the proposed model to provide the facility of implementing this model for huge volume of data. Empirical evaluations show that the proposed model have less number of lost rules and ghost rules in data stream. Therefore, the performance of this model is better than other existing researches and embedded features such as parallelism and scalability can make it suitable for big data mining. So, it can be concluded that the proposed model is more effective in big data mining than existing rule hiding approaches.

As future work, we will try to decrease undesired side effect of the proposed model to gain less information loss.

## References


[1] C.L.P.Chen and Ch.Zhang. “Data Intensive Applications, Challenges, Techniques, and Technologies: A Survey on Big Data”. Information Science, Vol.275, pp.314-347, 2014.

[2] O.Kwon. N.Lee. and B.Shin. “Data Quality Management, Data Usage Experience and Acquisition Intention of Big Data Analytics”, International Journal of Information Management, Vol.34, No.3,pp 387-394, 2014.

[3] A.Cuzzocrea. C.K.S. Leung and R.K.Mackinnon. “Mining Constrained Frequent Item-Sets from Distributed Uncertain Data”, Future Generation Computer Systems, Vol.37, pp 117-126. 2014.

[4] X.Zhang. Ch.Liu. S.Nepal. Ch.Yang. W.Dou. and J.Chen. “A Hybrid Approach for Scalable Sub-Tree Anonymization over Big Data using MapReduce on Cloud”, Journal of Computer and System Science, Vol.80, No.5, pp 1008-1020, 2014.

[5] Y.Li. M.Chen. Q.Li. and W.Zhen. “Enabling Multilevel Trust in Privacy Preserving Data Mining”, IEEE Transaction on Knowledge and Data Engineering, Vol.24, No.9, pp 1589-1612, 2012.

[6] Y.H.Wu. C.Chiang and A.L.P.Chen. “Hiding Sensitive Association Rules with Limited Side Effects”, IEEE Transaction on Knowledge and Data Engineering, Vol.19, No.1, pp 29-42, 2007.

[7] A.Gkoulalas-Divanis and V.S.Verykios. “Exact Knowledge Hiding through Database Extension”, IEEE Transaction on Knowledge and Data Engineering, Vol.21, No.5, pp 699-713, 2009.

[8] H.Q.Le. S.Arch-int. H.X.Nguyen and N.Arch-int. “Association Rule Hiding in Risk Management for Retail Supply Chain Collaboration”, Computer in Industry, Vol.64, No.4, pp776-784, 2013.

[9] Y.Ch.Li. J.S.Yeh. and Ch.Chang. "MCIF: An Effective Sanitization Algorithm for Hiding Sensitive Patterns on Data Mining", Advanced Engineering Informatics, Vol.21, No.3, pp 269-280, 2007.
[10] B.N.Keshavamurthy. D.Toshniwal. and B.K.Eshwar. "Hiding Co-Occurring Prioritized Sensitive Patterns over Distributed Progressive Sequential Data Streams", Journal of Network and Computer Applications, Vol.35, No.3, pp1116-1129, 2012.
[11] X.Wu. X.Zhu. G.Wu. and W.Ding. "Data Mining with Big Data", IEEE Transaction on Knowledge and Data Engineering, Vol.26, No.1, pp 97-107, 2013.
[12] M.E.Nergiz. and M.Z.Gok. "Hybrid K-Anonymity", Computers & Security, Vol.44, pp 51-63, 2014.
[13] B.Li. E.Erdin. M.H.Gunes. G.Bebis. T.Shipley. "An Overview of Anonymity Technology Usage", Computer Communication, Vol.36, No.12, pp 1269-1283, 2013.
[14] A.Monreale. G.Andrienko. N.Andrienko. F.Giannotti. D.Pedreschi. S.Rinzivillo. and S.Wrobel. "Movement Data Anonymity through Generalization", Transactions on Data Privacy, Vol.3, No.2, 2010.
[15] S.Kisilevich. L.Rokach. Y.Elovici. and B.Shapira. "Efficient Multidimensional Suppression for K-Anonymity", IEEE Transaction on Knowledge and Data Engineering, Vol.22, No.3, pp 334-347, 2010.
[16] G.Zhang. Y.Yang. X.Liu. and J.Chen. "A Time-Series Pattern Based Noise Generation Strategy for Privacy Protection in Cloud Computing, International Symposium on Cluster", Cloud and Grid Computing (CCGrid), pp 458-465, 2010.
[17] H.Wang. "Quality Measurement for Association Rule Hiding", AASRI Procedia, Vol.5, pp 228-234, 2013.
[18] G.V.Moustakides. and V.S.Verykios. "A MaxMin Approach for Hiding Frequent Item Sets", Data & Knowledge Engineering, Vol.65, No.1, pp 75-89, 2008.
[19] S.Wang. B.Parikh. and A.Jafari. "Hiding Informative Association Rule Sets", Expert Systems and Applications, Vol.33, No.2, pp 316-323, 2007.
[20] Ch.Wang. S.Tseng. and T.Hongm. "Flexible Online Association Rule Mining Based on Multidimensional Pattern Relations", Information Science, Vol.167, No.12, pp 1752-1780, 2006.
[21] E.Dasseni. V.S.Verykios. A.K.Elmagarmid. and E.Bertino. "Hiding Association Rules by Using Confidence and Support", Information Hiding Lecture Notes in Computer Science, Vol.2137, pp 369-383, 2001.
[22] K.Jung. S.Park. S.Cho. and S.Park. "A Novel Privacy Preserving Association Rule Mining using Hadoop", The Third International Conference on Data Analytics, 2014, pp 131-137.
[23] L.Xu. C.Jiang. J.Wang. J.Yuan. and Y.Ren. "Information Security in Big Data: Privacy and Data Mining". IEEE Access, vol.2, pp. 1149-1176, 2014.
[24] Ch.Borgelt. and R.Kruse. "Introduction of Association Rules: Apriori Implementation", Compstat, Physica-Verlog Heidelberg, pp 395-400.
[25] Ch.Borgelt. "An Implementation of the FP-Growth Algorith", Proceeding of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations, 2005, pp 1-5.

**Golnar Assadat Afzali** is a M.Sc graduate of Information Technology Engineering at K.N.Toosi University of Technology. She received her B.Sc degree from Isfahan University of Technology (IUT). Her research interests include Network Security, Trust and Privacy and Big Data mining.

**Shahriar Mohammadi** is a former senior lecturer at the University of Derby, UK. He received his Ph.D in Computer Science (Network Security) from University of Salford, Manchester UK in 1993. Then, while he used to be a Network consultant, he worked in UK universities of Salford and Derby of the UK for more than fifteen years as a lecturer and senior lecturer. He currently is a lecturer in the Industrial Eng. Department of the University Of K.N.Toosi, of Iran. His main research interests and lectures are in the fields of Networking, Data Security, Network Security, e-commerce and e-commerce Security. He has published more than hundred and twenty papers in various journals and conferences as well as seven books.