A Survival Study on Privacy Preservation of Data Sharing with Optimal Side Effects

P. Tamil Selvan,
Research Scholar,
Department of Computer Science,
Karpagam University
Coimbatore, India

Dr. S. Veni,
Research Supervisior,
Department of Computer Science,
Karpagam University
Coimbatore, India

Abstract - **Privacy preserving is a significant task in the success delivery of the data to the users using data mining techniques. Privacy preserving data mining (PPDM) protects the individual sensitive data while sharing to the public users. PPDM was used to reduce privacy threats by hiding sensitive information while allowing required information to be mined from databases. Many existing PPDM techniques like data sanitization do not hide the sensitive information. The authenticity of original database using sensitive item hiding technique also alters the originality of the database in data sanitization techniques. In this work, association rule mining technique is used to send data to all the users with optimal side effects. Our research work helps to maintain the individual privacy for these sensitive attributes**

Keywords: Privacy Preserving Data Mining, Data Sanitization, Optimal Side Effects Sensitive Attributes.

I.INTRODUCTION

Privacy preserving data mining (PPDM) is a primary issue in determining the results of privatizing user's data. PPDM techniques adopting sensitive item hiding alters the innovation of the record. The existing method of privacy-preservation results in the failure of information for data mining functions. The loss of information is taken as a loss of efficiency in privacy preserving data mining functions. Traditional data mining techniques examine database to identify potential relations between items. Several applications require protection beside the disclosure of private, confidential, or secure data. PPDM technique is used to minimize the privacy threats by hiding sensitive information when permitting the necessary information extracted from databases.

In PPDM, data sanitization is used to hide sensitive information with the minimum side effects for preserving the original database as reliable. The spontaneous method of data sanitization to hide sensitive information is straightly to remove sensitive information from amount of data. The key goal in many distributed methods for privacy-preserving data mining is to allow the computation of useful aggregate statistics over the whole data set without compromising the privacy of the individual data sets for various participants.

Privacy preserving data mining is an innovative analysis in data mining and also in statistical databases. In PPDM, data mining algorithms are examined for side effects attaining the data privacy. There are two fold concerns in privacy preserving data mining techniques. Initially, it is sensitive raw data that is secured from unauthorized access like identifiers, names, addresses adapted from original database for receiver of data which fails to compromise another person's privacy. Next, sensitive knowledge is eliminated from a database using data mining algorithms.

The privacy preservation data mining using association rule mining aim is to attain quality privacy preservation for distributed data mining with optimal side effects on the original database. This also improves the efficiency of privacy preserving association rule mining with constraint minimization and the privacy preserving mechanism with efficient data utility.

This paper is organized as follows: Section II discusses Privacy Preserving Data Mining, Section III shows the study and analysis of the existing privacy preserving techniques on Data mining, Section IV identifies the possible comparison between them and Section V discusses the limitation of privacy preserving data mining techniques.

II. LITERATURE SURVEY

Data mining is the combination of various fields counting machine learning, database systems, data visualization, statistics and information theory. From [1], Hiding-Missing-Artificial Utility (HMAU) algorithm is designed to hide sensitive information through item set removal. Though, HMAU algorithm fails to control the highest frequency in sensitive rules regarding current sensitive transaction. In [4], Fast Distributed Mining Algorithm presented secured mining of association rules in parallel distributed databases with its secure multi-party protocol for calculating combination of private subsets with two secure multi-party algorithms.

Perturbation-based PPDM to Multilevel Trust (MLT-PPDM) [2] facilitated flexibility and created perturbed copies of data for different trust levels, though, data owner is unable to forecast all possible trust level that requires prior requirement. As described in [3], Exact Knowledge Hiding through Database Extension arrives in optimal solution for hiding sensitive frequent item sets. It protects sensitive knowledge with minimum effect on sensitive item and tests on real-world data sets with minimum threshold. However, exact border-based are typically of lower quality and an increment observed in number of constraints are produced.

Privacy preserving mining of association rules from outsourced transaction databases [5] calculated encrypt/decrypt (E/D) module to alter client data sooner than it is distributed to server and regain true patterns with exact support. Though E/D module supposes the attacker which fails to contain knowledge on hiding aspect and relaxation break encryption scheme and takes privacy vulnerabilities. Logical framework [6] is planned to reveal secret information and databases that enclose nulls. Query answers implicitly informative lack of exposing original content. But, query rewriting method does not include combination of nulls and negation.

In [7], Anonymous Publication of Sensitive Transactional Data changes the binary data into a band matrix by executing the variation of rows and columns in the unusual table. Numerous anonymization methods like generalization and bucketization are intended for privacy preserving, however methods fails to offer an efficient data utility. Slicing divides the data both horizontally and vertically and protects better data utility than generalization [8] and tradeoff happens in controlling the constant attributes by minimizing the dimensionality.

III. Privacy Preservation Data Mining Using Association Rule Mining

In privacy preserving data mining techniques, association rules are used for examining and forecasting customer behavior. Increase in the demand for privacy, secure data mining is the development of techniques that includes the privacy and security with effective data circulating. Complexity of running data mining algorithms on private data is carried out in privacy-preserving data mining (PPDM) techniques. The key goal in many distributed methods for privacy-preserving data mining is to allow the computation of useful aggregate statistics over the whole data set without compromising the privacy of the individual data sets for various participants.

A. *Minimization of Side Effects on Hiding Sensitive Itemsets in Privacy Preserving Data Mining*

Data mining is accepted to regain and examine knowledge from large amount of data. Privacy preserving data mining (PPDM) is used to reduce privacy threats by hiding sensitive information while permitting the necessary information mined from databases. Privacy information contains some personal or confidential information in business like social security numbers, home address, credit card numbers, credit ratings, purchasing behavior, and best-selling commodity. In PPDM, data sanitization is employed to hide sensitive information with the minimal side effects for keeping the original database .

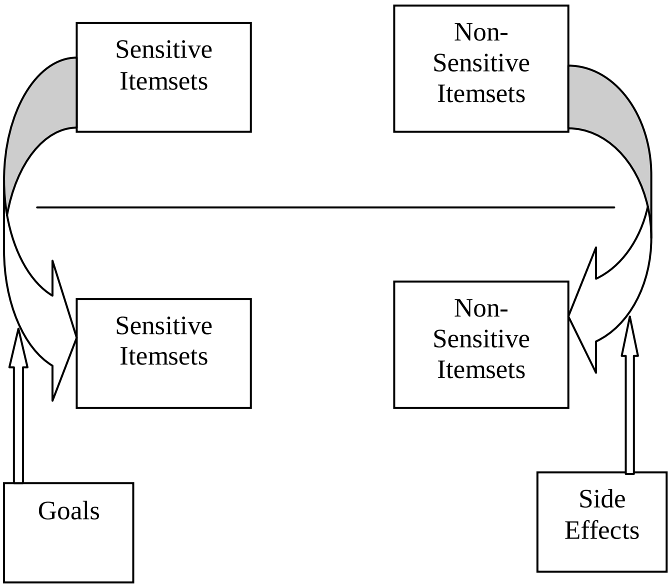


Figure. 1 Goals and Side Effects of Privacy Preserving

From figure.1, the main goal is to transfer the data from one sensitive itemsets to another sensitive item sets. When the non sensitive information is transferred, the side effects are created. Privacy preserving data mining (PPDM) resulted in major issue for hiding private, confidential, or secure information. The original database is sanitized for hiding sensitive information. The sensitive technique of data sanitization is employed to hide sensitive information which used to delete the sensitive information directly from amount of data. Data sanitization process has three side effects namely hiding failure, missing cost, and artificial cost.

Hiding-missing-artificial utility (HMAU) algorithm is introduced for calculating the operations required to be removed for hiding sensitive itemsets by taking three dimensions as hiding failure dimension (HFD), missing item set dimension (MID), and artificial itemset dimension (AID). The transactions with sensitive item set are to be identified with the minimum HMAU values among transactions that eliminated from the database. Data sanitization is the general technique for sensitive knowledge from disclosure in PPDM.

B. *Privacy-Preserving Mining of Association Rules*

An organization subcontracts its mining requirements like resources or capability to a third party service provider. The issues of association rule mining task is studied in privacy preserving framework. The main function of privacy preserving is carried out on privacy-preserving data mining (PPDM) techniques with grouping of frameworks. Feature of the patterns mined from the data is planned to share with parties than the data owner

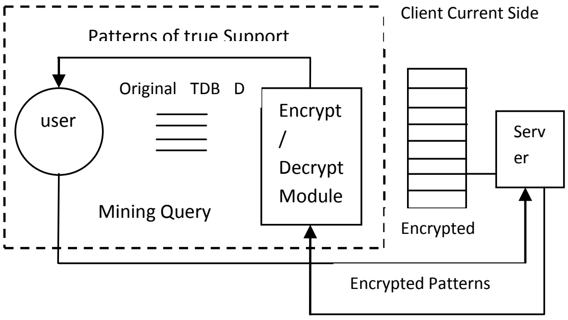


Figure.2 Architecture of Mining-as-a-Service

Figure. 2 describes the mining as-a service. The client/owner encrypts its data using encrypt/decrypt (ED) module is taken as “black box” from its viewpoint. This module is dependable for altering the input data into an encrypted database. The server performs data mining and sends the (encrypted) patterns to the owner. Encryption scheme has the property which returned supports are not true supports. The ED module recovers the true identity of the returned patterns as true supports.

Slicing manage high-dimensional data without a clear separation of QIs and SAs. Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH) transform data into band matrix by performing permutations of rows and columns. Efficient linear-time heuristic creates anonymized groups based on data organization. Initially, an attack model is designed for the adversary and creates the environment knowledge for an exact solution. Idea of privacy requires for each ciphertext item where at least k−1 distinct cipher items are identical from the item about their supports. Next, an encryption scheme called RobFrugal is the E/D module facilitates to change client data before it is shipped to the server.

Third, E/D module allows for enhancing the true patterns and exact support, it creates and preserves a compact structure called synopsis. Privacy preserving mining provide the E/D module with an effective approach for preserving the synopsis in the form of attachments. Next, a formal analysis is designed on attack model for privacy preserving mining and verifies the probability an individual item, a transaction, or patterns that broken by the server are controlled by the owner by locating the anonymity threshold.

C.Anonymous Publication of Sensitive Transactional Data

Privacy Preserving Data Mining is the search of data mining side-effects on privacy that obtains a rising attention from the research community. Anonymity is the condition of having one’s name or identity unknown or concealed. Anonymous provides valuable social ideas and allows individuals against institutions by bordering examination, however it is used by incorrect achievers to cover the events or avoid the ability to permit anonymous contact to services that avoid tracking of user's personal information and user behavior like user location, frequency of a service usage.

Two reorganization methods for sensitive transactional data are designed. The initial method changes the data into a band matrix by executing the variations of rows and columns in the innovative table. The changes obtain the benefits of data sparseness and find the nonzero openings near the main diagonal. The benefit is the neighboring rows containing high correlation. The second data transformation method depends on arranging regarding Gray encoding: the QID items in each transaction t are recognized as the Gray code of t. The transaction set is sorted consistent with the rank in the Gray sequence. The result of two methods is fed to an efficient linear-time heuristic which groups mutually the transactions resulting in the minimization of the search space of the solution. Since both data transformations capture correlation well, groups contain transactions with similar QID leading to increased data utility.

IV. COMPARISON OF PRIVACY PRESERVATION DATA MINING WITH OPTIMAL SIDE EFFECTS AND SUGGESTIONS

In order to compare the privacy perseverance in data mining using association rule mining, number of files is taken to execute the experiment. Various parameters are used to measure the privacy preserving of the data mining techniques.

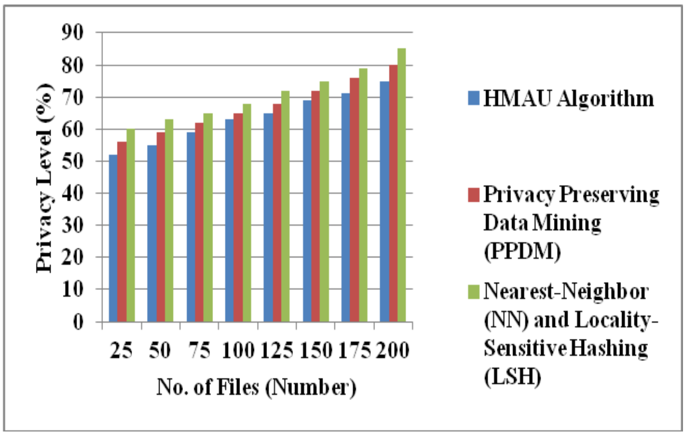
A. Privacy Level

Privacy level is defined as the level at which the data is privately sent to the required user without showing to the public users. Privacy level increases the information delivery to the private users. It is measured in terms of percentage (%).

Table 4.1Tabulation for Privacy Level on Privacy Preserving Data Mining with Optimal Side Effects

No. of Files (Number)	Privacy Level (%)		
	HMAU Algorithm	*Privacy Preserving Data Mining (PPDM)*	*Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH)*
25	52	56	60
50	55	59	63
75	59	62	65
100	63	65	68
125	65	68	72
150	69	72	75
175	71	76	79
200	75	80	85

The privacy level comparison takes place on existing Hiding-Missing-Artificial Utility (HMAU) Algorithm, Privacy Preserving Data Mining (PPDM) and Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH).



.Figure. 4.1 Privacy Level on Privacy Preserving Data Mining with Optimal Side Effects

Figure 4.1describes the privacy level on Privacy preserving Data Mining. As the number of files increases, privacy level gets increased automatically. The experiment shows that Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH) greatly lift up the privacy level when compared with Hiding-Missing-Artificial Utility (HMAU) Algorithm and Privacy Preserving Data Mining (PPDM). Research in Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH) is 10 – 15 % higher private when compared to Hiding-Missing-Artificial Utility (HMAU) Algorithm and 4-7 % higher private when compared with Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH).

B.Throughput

Throughput is defined as the rate of successful data delivery over a communication channel. Throughput increases the overall efficiency of the system. Throughput values provide the information about the delivery of the data while transferring the data to the various network channels. Throughput level is measured in terms of percentage (%).

Table 4.2 Tabulation for Throughput on Privacy Preserving Data Mining with Optimal Side Effects

No. of Files (Number)	Throughput (%)		
	HMAU Algorithm	*Privacy Preserving Data Mining (PPDM)*	*Nearest-Neighbor (NN) and Locality-Sensitive*
25	65	55	60
50	68	59	62
75	70	62	65
100	73	66	69
125	75	69	71
150	79	73	75
175	82	75	78
200	85	78	83

The throughput comparison takes place on existing Hiding-Missing-Artificial Utility (HMAU) Algorithm**,** Privacy Preserving Data Mining (PPDM) and Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH).

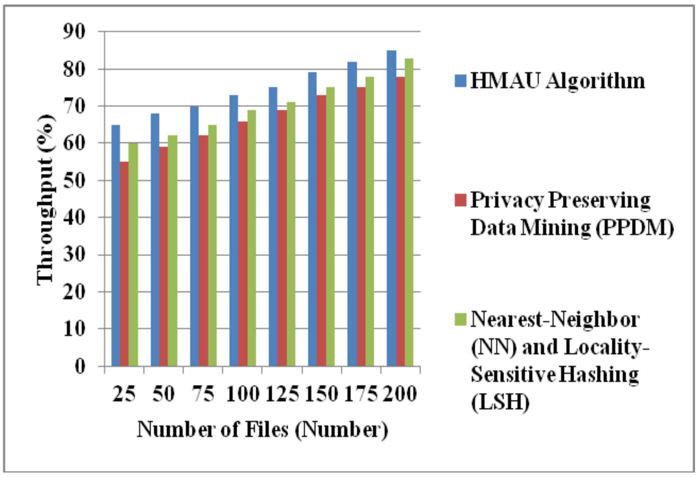


Figure.4.2 Throughput on Privacy Preserving Data Mining with Optimal Side Effects

Figure 4.2 shows that throughput level of Privacy Preserving Data Mining (PPDM). Research in Hiding-Missing-Artificial Utility (HMAU) Algorithm has higher throughput than Privacy Preserving Data Mining (PPDM) and Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH). Throughput of HMAU Algorithm is 8-15% higher than the Privacy Preserving Data Mining (PPDM) and 2-8 % higher than Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH).

C. Privacy Preserving Efficiency

Privacy preserving efficiency is defined as the rate at which the data is effectively transferred to correct user with high privacy. Privacy preserving efficiency plays an important part in delivering the data to the private user without showing the data to the public users. Privacy preserving efficiency is measured in terms percentage (%).

Table 4.3 Tabulation for Efficiency on Privacy Preserving Data Mining with Optimal Side Effects

No. of Files (Number)	Privacy Preserving Efficiency (%)		
	HMAU Algorithm	*Privacy Preserving Data Mining (PPDM)*	*Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH)*
25	48	62	54
50	50	65	57
75	55	68	60
100	59	71	63
125	62	75	65
150	65	78	67
175	67	81	71
200	70	85	75

Privacy preserving efficiency comparison takes place on existing Hiding-Missing-Artificial Utility (HMAU) Algorithm**,** Privacy Preserving Data Mining (PPDM) and Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH).

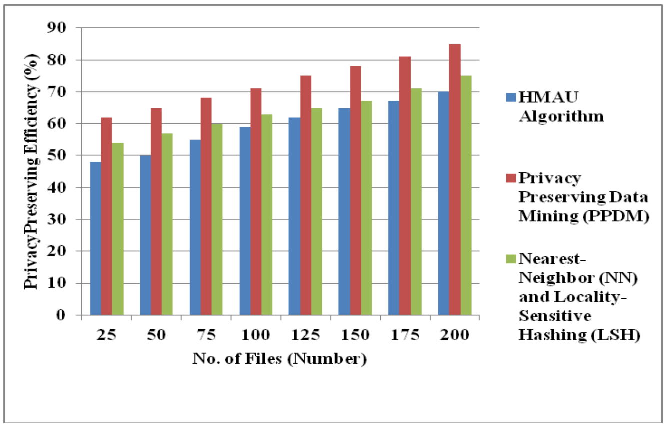


Figure 4.3 Efficiency on Privacy Preserving Data Mining with Optimal Side Effects

Figure 4.3 shows the performance of efficiency of existing methods. Privacy Preserving Efficiency of Privacy Preserving Data Mining (PPDM) is comparatively higher than that of HMAU Algorithm and Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH). Research in Privacy Preserving Data Mining (PPDM) has 17-22% higher efficient than HMAU Algorithm and 10-13% higher efficient than Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH) technique.

V.DISCUSSION ON LIMITATION OF PRIVACY PRESERVING DATA MINING TECHNIQUES WITH OPTIMAL SIDE EFFECTS

In Hiding-Missing-Artificial Utility (HMAU) algorithm, PPDM has significant issue for hiding private, confidential, or secure information. Highest frequency in sensitive rules is linked to current sensitive transaction. Noise addition and data modification are significant to hide sensitive information in PPDM is not implemented. Secure multi-party protocol use homomorphism encryption when computational costs are significantly higher. Logarithmic communication overhead happens only when size of intersection of two sets is cleared by a constant. Server obtains no information regarding other pairs in server's database. Privacy-preserving data mining (PPDM) framework supposes the attacker which fails to hold such knowledge. It is not appropriate for commercial privacy where the analytical properties are revealed.

In anti-discrimination techniques, indirect discrimination comprises rules or procedures which are not clearly revealing discriminatory attributes. Post processing approach fails to permit the data set which is to be distributed. The technique has huge effects on of changing minimum support and minimum confidence. In SQL standard, query rewriting methodology does not contain mixture of nulls and negation. SQL nulls in context are more significant queries.

(k, p)- Anonymity requires severe restriction on PR equality and results in serious pattern loss. K-anonymity model fails to deal with issues as it experiences severe pattern loss. Micro aggregation suffers pattern loss in uncontrolled manner. Requirement is not flexible and sometimes impossible to meet. Slicing associations between column values of a bucket and probable to lose the data utility. Column generalization also leads to the information loss. Connections among attributes in different columns are lost in marginal publication. Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH) Dimensionality reduction techniques are more efficient anonymization. Dimensionality mapping is efficient only for low-dimensional QIDs and not suited for the transactional data.

A. Future Directions

The future direction of using the privacy preserving data mining techniques are attaining the quality privacy preservation for distributed data mining with optimal side effects on the original database, improving the efficiency of privacy preserving association rule mining with constraint minimization and improving the privacy preserving mechanism with efficient data utility

VI.CONCLUSION

Examination about the existing privacy preserving data mining techniques such as Hiding-Missing-Artificial Utility (HMAU) Algorithm, Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH), Privacy Preserving Data Mining (PPDM). Novel Hiding-Missing-Artificial Utility (HMAU) algorithm hides sensitive itemsets through transaction deletion and transaction with maximal ratio of sensitive to non-sensitive is chosen to whole deletion. Though, the noise addition and data modification are important for hiding the sensitive information in PPDM is not employed. Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH) transform data into band matrix by performing permutations of rows and columns. However, mapping of dimensionality is efficient only for low-dimensional QIDs and does not suit for the transactional data.

In PPDM, encrypt/decrypt (E/D) module employs to transform client data before it shipped to server. E/D module recovers true patterns and their correct support. However, framework which containing the attacker does not contains such knowledge. Relaxations break the encryption scheme and provide the privacy vulnerabilities. PPDM is not suited for corporate privacy where the analytical properties are revealed. Observation was increasing the privacy preserving data mining efficiency using association rule mining techniques. The wide range of experiments on existing techniques calculates the relative performance of the various privacy preserving techniques and its limitations. The result shows that the research work can be done in the privacy preserving data mining techniques with minimal side effects which increases the privacy preserving efficiency.

REFERENCES

[1] Chun-Wei Lin, Tzung-Pei Hong, and Hung-Chuan Hsu, "Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", Hindawi Publishing Corporation, e Scientific World Journal Volume 2014.

[2] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, "Enabling Multilevel Trust in Privacy Preserving Data Mining", IEEE Transaction on Knowledge and Data Engineering, SEPTEMBER 2012

[3] Aris Gkoulalas-Divanis., and Vassilios S. Verykios., "Exact Knowledge Hiding through Database Extension", IEEE Transaction on Knowledge and Data Engineering, MAY 2009

[4] Tamir Tassa., "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 4, APRIL 2014

[5] Fosca Giannotti., Laks V. S. Lakshmanan., Anna Monreale., Dino Pedreschi., and Hui (Wendy) Wang., "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases" IEEE Systems Journal, VOL. 7, NO. 3, September 2013

[6] Leopoldo Bertossi and Lechen Li, "Achieving Data Privacy through Secrecy Views and Null-Based Virtual Updates" IEEE Transactions on Knowledge and Data Engineering, MAY 2013

[7] Tiancheng Li., Ninghui Li., Jian Zhang., Ian Molloy., "Slicing: A New Approach to Privacy Preserving Data Publishing," IEEE Transactions on Knowledge and Data Engineering, Volume: 24, Issue: 3, 2012

[8] Gabriel Ghinita., Panos Kalnis., and Yufei Tao., "Anonymous Publication of Sensitive Transactional Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO: 2, FEBRUARY 2011

Tamil Selvan P. completed his M.Phil in Computer Science from Karpagam University in 2009. He is working as Assistant Professor in Department of Computer Science, Karpagam University, Coimbatore. His experience is 7 yrs. He has presented a paper in International Conference. His research interests are Data mining and warehousing.

Dr. S. Veni completed her Ph.D in Computer Science from Bharathiar University in 2014. she is working as Associate Professor in Department of Computer Science, Karpagam University, Coimbatore. Her experience is 12 yrs. she has presented various papers in National and International Conference. Her research interests are Computer Networks.