Trilateral Spearman Katz Centrality Based Least Angle Regression for Influential Node Tracing in Social Network

VIMAL KUMAR P. (mevimal2016@gmail.com)
PSR Engineering College https://orcid.org/0000-0003-4707-9978

Balasubramanian C.
PSR Engineering College

Research Article

Keywords: Social Network, Node Cover Preprocessor, Trilateral, Spearman Correlated, Katz Centrality, Least Angle Regression

Posted Date: March 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-184850/v1

TRILATERAL SPEARMAN KATZ CENTRALITY BASED LEAST ANGLE REGRESSION FOR INFLUENTIAL NODE TRACING IN SOCIAL NETWORK

P. Vimal kumar[*1], C. Balasubramanian[2]

[*1]Assistant Professor, Department of CSE, PSR Engineering College, Sivakasi, India.

[2]Professor, Department of CSE, PSR Engineering College, Sivakasi, India.

Corresponding Author's Mail ID: mevimal2016@gmail.com

Abstract—

With the epidemic growth of online social networks (OSNs), a large scale research on information dissemination in OSNs has been made an appearance in contemporary years. One of the essential researches is influence maximization (IM). Most research adopts community structure, greedy stage, and centrality measures, to identify the influence node set. However, the time consumed in analyzing the influence node set for edge server placement, service migration and service recommendation is ignored in terms of propagation delay. Considering the above analysis, we concentrate on the issue of time-sensitive influence maximization and maximize the targeted influence spread. To solve the problem, we propose a method called, Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) for influential node tracing in social network is proposed. Besides, two algorithms are used in our work to find the influential node in social network with maximum influence spread and minimal time, namely Trilateral Statistical Node Extraction algorithm and Katz Centrality Least Angle Influence Node Tracing algorithm, respectively. Extensive experiments on The Telecom dataset demonstrate the efficiency and influence performance of the proposed algorithms on evaluation metrics, namely, sensitivity, specificity, accuracy, time and influence spread

Index Terms—Social Network, Node Cover Preprocessor, Trilateral, Spearman Correlated, Katz Centrality, Least Angle Regression.

I. INTRODUCTION

With the swift evolution of network technology and familiarization of the Internet, social applications, to name a few, like WeChat, Weibo and Snapchat, moderately produce an enormous amount of network data. Several practical applications required to mine profitable information from the profitable network data. Since then, the investigation of Influence Maximization (IM) has become an intense topic of concern in social network, specifically in the recent few years due to its crucial part in an extensive span of fields, like, network monitoring, product/brand recommendation and so on.

Community Based Influence Maximization Algorithm (CBIMA) was proposed in [1] to solve the issues concerning influence maximization in social networks. With this objective, an influence maximization method was designed that in turn obtained influential nodes by means of community structure followed by which the influence distribution difference was presented. Here, to start with, network embedding-based community detection model was designed with which the entire social network was split into distinct high-quality communities.

Next, the solution for influence maximization was divided into candidate stage and the greedy stage. Here, in candidate stage, the candidate nodes were selected via a heuristic algorithm, and as far as the greedy stage was concerned seed nodes were selected via modular property-based Greedy algorithm. With this, better performance was said to be arrived at with minimum running time.

Community Finding Influential Node (CFIN) was designed in [2] was designed with the purpose of selecting optimal users on the basis of the community structure, which in turn maximized the influence spread in the networks. The CFIN was split into two stages. They were, selection of seed and community spreading in a localized manner.

In the first stage, seed nodes were extracted via community detection algorithm. Here, with the objective of reducing the computational complexity involved in selecting seed node, meaningful communities were obtained in a significant manner. Next, influence spread inside communities was determined with which the final seed nodes responsible for distribution was analyzed. With this both the coverage ratio and running time were said to be improved with better coverage ratios.

Despite improvements observed in the coverage ratio with minimum running time and higher coverage ratio for obtaining influential nodes, the accuracy and the sensitivity rate with higher influence spread was less focused. To address these issues in our work, Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) for influential node tracing in social network is proposed.

A. Objectives

The main contributions of this paper are summarized as follows:

- Inspect the Influential Node Tracking (INT) issue as an augmentation to the conventional Influence Maximization problem to maximize the accuracy and influence spread under a dynamic social network.
- To propose Trilateral Statistical Node Extraction algorithm with reduces the tracing time of influence maximization node by first eliminating the ineffective nodes via Node Cover Preprocessor and

then extract optimal nodes by employing Trilateral Statistical functions.

- To design Katz Centrality Least Angle Influence Node Tracing algorithm that with robust node selected via Spearman Correlated and Normalization functions achieves higher rate of sensitivity and specificity.
- We evaluate the performance on large-scale telecom dataset. The experiment results confirm our theoretical findings and show that our proposed Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) method achieve better performance of both influence coverage and running time along with sensitivity, specificity and accuracy.

B. Organization of Paper

The manuscript continues as follows. Section 2 presents a considerable reference list for related work in the field. The results of this exploration account for the novelty of the multi-objective proposal of this work. Section 3 describes the Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) method. Section 4 provides the experimental setup, while Section 5 presents and discusses the main results obtained. Finally, conclusions are included in Section 6.

II. RELATED WORKS

Identifying the most influential nodes in social networks has extensive range of applications. Different types of methods have been in the recent years designed with the purpose of estimating the influential nodes on the basis of their structural location in social network. An algorithmic approach based on Map Reduce with centrality as a measure was proposed in [3] with the purpose of improving the computational efficiency. However, with natural life evolution and influence made by peer groups have resulted in temporal patterns. To address this issue, a temporal preference model was designed in [4] to detect network structure changes on the basis of the node centrality. With this, strong correlations were said to be established. Yet another influence ranking model based on Susceptible Infected Recovered was proposed in [5] to find influential nodes.

From the information transmission aspect, different types of social networks have different information transmission modes due to the function diversity and user structures. Due to this, social network has become a new path for viral marketing that is utilized in promoting products, innovations and opinions. An enhanced multi attribute k shell method was designed in [6] utilizing the iterative information in the decomposition process for ranking influential nodes.

Yet another deep influential evaluation model was proposed in [7] by considering complicated graphic structure. This in turn evaluated influence node in computationally efficient manner. However, with the inception of complex networks, this type of influence node identification remains a major issue to be attended. A machine learning framework was designed in [8] to evaluate the complicated relationship between neighboring and non-neighboring nodes.

In the analysis of influence maximization in social networks, the speed with which the information is spread reduces with two different factors, i.e., elevated time and distance. Therefore, the examination of the transmission of information is of great importance to the supervision and administration of public point of view.

An Influence Maximization algorithm was designed in [9] on the basis of the rate attenuation propagation model. With this, both the accuracy and time were said to be improved. Two new effective algorithms based on centrality measure and local measure were proposed in [10] under the Independence Cascade (IC) and Linear Threshold (LT) models. With this, the time complexity involved in influential node analysis was found to be reduced.

Conventional types of methods, like, centrality based and machine learning, only took into considerations either the structures involved in the network or the features involved in the design to measure the node significance. However, the influential significance has to be determined by both considering the network structures and node features. To solve this issue, a deep learning model was presented in [11] to identify the most influential nodes in a complex network. Graph classification method was applied in [12] for identifying significant nodes.

In recent few years studies conducted on the influential community have identified communities that involve large types of influential members. Several types of metrics are hence involved in influencing and existing techniques however search for influential communities by taking into account only one influence type. In [13], an efficient influential community search method that identified the influential communities based on multiple influence criteria. Yet another deep reinforcement learning was applied in [14] for influence maximization.

In the meanwhile, investigations made on prior analysis have specifically designed on the basis of the static network topology. However, the user's online/offline status and topic preference made it a cumbersome process for conventional methods to be applied in real scenarios. Based on the above analysis, time-sensitive influence maximization was concentrated on [15] and accordingly influence maximization was made. An interchange greedy approach was applied in [16] to maximize joint influence under single network. With this better performance was said to be achieved in both influence coverage and running time.

To improve both influence spread and time efficiency, a Multipath Asynchronous Threshold (MAT) model was designed in [17] by employing both neighboring and non-neighboring influencers into account. A network representation learning model was proposed in [18] by using diffusion based processes, therefore identifying influential nodes. A new optimization model using independent cascade and linear threshold was designed in [19] to maximize influence spread.

Based on the above research and analysis, the proposed method constructs a four step processes to perform influential tracing in social network. That is to say, considering the normalized spearman correlation and Katz Centrality Least Angle, network structure information for influential node identification is maximized with minimum response time.

III. PROPOSED WORK

This section explains the design and development of the proposed algorithm for Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) for influential node tracing in social network. Here, four different steps are effectively combined to design an algorithm for influential nodes selection covering service migration and service recommendation. Figure 1 given below shows the block diagram of Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) for influential node tracing in social network.

As illustrated in the below figure, to identify the influential nodes in social network, four different processes are followed. They are preprocessing, feature extraction, feature selection and influential node tracing. First, with the Telecom dataset provided as input, ineffective nodes are removed by means of Node Cover Preprocessor model. Second, feature extraction or node extraction is performed by employing Trilateral Statistical Node Extraction model where only the optimal features or nodes are extracted. Next, feature selection of normalized nodes used in identifying the influential nodes are selected by applying Spearman Correlated Normalized Feature Selection model. Finally, Katz Centrality Least Angle Regression is applied to the selected nodes for significant influential node tracing. The elaborate description of the proposed method is given below.

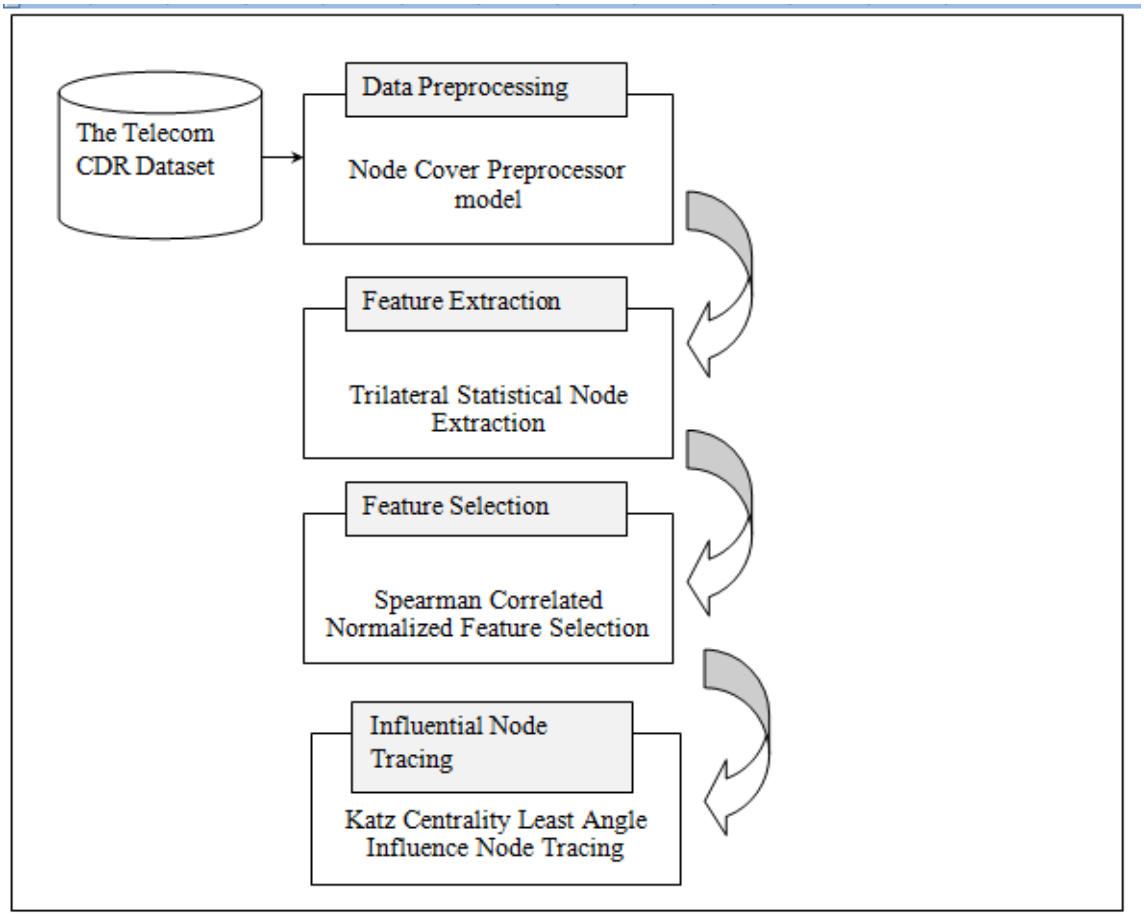


Figure 1 Block diagram of Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR)

3.1 Node Cover Preprocessor model

With the evolution of social networks and extensive utilization between social network users, the magnitude of the social network is surging day by day. So it makes certain issues in identifying the most influential nodes like, memory deficiency, slowdowns of processing and so on. In this work, to address this issue, a Node Cover Preprocessing (NCP) model is designed that provides a pre-processing sample prior to the execution of the indigenous influential node selection algorithm.

This NPC model in turn minimizes the memory and therefore enhancing the process of detecting influential nodes and accelerating overall processes. The proposed method uses the NCP as a pre-processing. Figure 2 given below shows the block diagram of Node Cover Preprocessing (NCP) model.

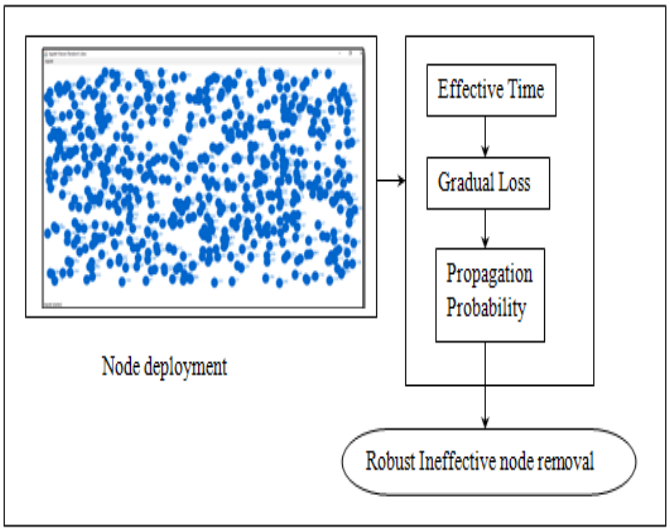


Figure 2 Block diagram of Node Cover Preprocessor model

As shown in the above figure, let us consider the social network modeled as a directed network '$G = (V, E)$', where 'V' represents the individual nodes (i.e., 600 nodes) or social network users while 'E' represents the links between the individual nodes or link between the social network users. In addition, each edge '$(a, b \in E)$' is related with a propagation probability '$Prop^{G}_{a,b}$' specifying the robustness of effect of individual node or social network user 'a' on 'b' with two different conditions, effective or potential '$C_{a,b} \in E \rightarrow Eff/Pot$'.

On one hand the effective edge 'E_{Eff}' refers to the relationship between online users, and on the other hand, the potential edge 'E_{Pot}' refers to that at least one user on both sides of the edge is offline. Due to the nature of users' online status and time constraints of information dissemination, network structure is distinct.

For two information 'In_i' and 'In_j', to be transmitted at different time intervals 'T_i' and 'T_j', then 'T_i' denotes the effective time for information 'In_i' obtained using the deadline of information '$t^{In_i}_{end}$' and start time of information '$t^{In_j}_{start}$' as given below.

$$T_i = t^{In_i}_{end} - t^{In_j}_{start} \quad (1)$$

Next, propagation probability is derived due to the impact of users on neighbors has considerable variability, the propagation probability decreases with increase in distance 'd'. This is mathematically expressed as given below.

$$\mu = \frac{1}{d^2} \quad (2)$$

Finally, the propagation probability 'Prop' for a directed network graph 'G' of two different social network users 'a' and 'b' is mathematically expressed as given below.

$$Prop^{G}_{a,b} = \mu \cup (a, b[T_i]) \quad (3)$$

From the above equation (3), the propagation probability for removing ineffective nodes in the network is obtained by integrating the gradual loss 'μ' and the information to be propagated between two nodes or users '$a, b[T_i]$' respectively. The pseudo code representation of Node Cover Preprocessor model is given below.

Algorithm 1 Node Cover Preprocessor model

Input: Nodes '$N = N_1, N_2, \dots, N_n$'
Output: Effective nodes '$EN = EN_1, EN_2, \dots, EN_n$'
1: **Begin**
2: **For** each Nodes 'N' with a directed network '$G = (V, E)$'
3: **For** information 'In_i' and 'In_j' to be transmitted at different time intervals 'T_i' and 'T_j'
4: Evaluate effective time using (1)
5: Evaluate gradual loss 'μ' using (2)
6: Evaluate propagation probability using (3)
7: **Return** (effective nodes)
8: **End for**
9: **End for**
10: **End**

Algorithm 1 Explanation:

As given in the above Node Cover Preprocessor model, three different measures are evaluated with the objective of eliminating the ineffective nodes. First, the effective time is measured then with the time obtained, the gradual loss based on the distance is evaluated. Finally, propagation probability is measured to obtain the effective nodes and eliminate the ineffective nodes.

3.2 Trilateral Statistical Node Extraction

In this section, with the elimination of ineffective nodes and selection of effective nodes, next node extraction is done by applying trilateral statistical function. Here, trilateral statistical function refers to the application of three different hypotheses and extracting the based on these three hypotheses. For the purpose of influential node tracing in social network, a node in our work is defined as a set of points with three hypotheses that the number of points '$P = P_1, P_2, \dots, P_n$' must be above the threshold 'T' and must be within a predefined transmission range 'TR'. Figure 3 given below shows the block diagram of Trilateral Statistical Node Extraction model.

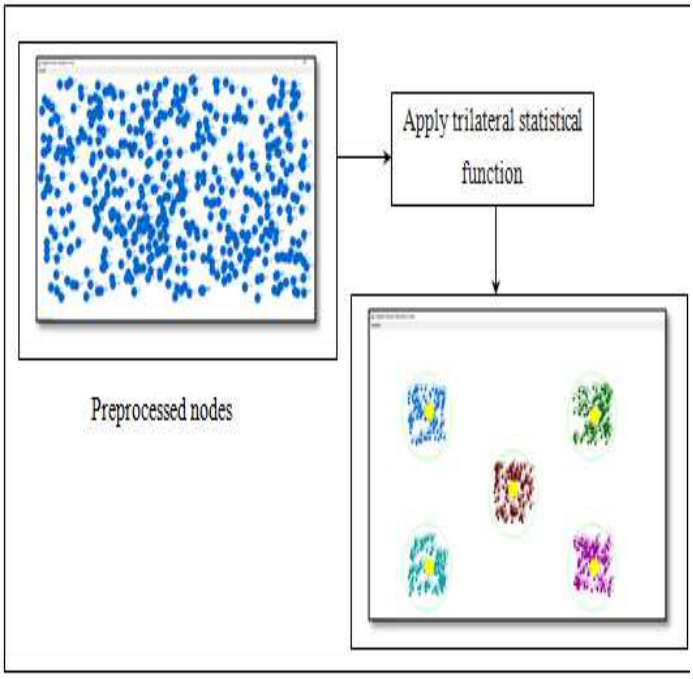


Figure 3 Block diagram of Trilateral Statistical Node Extraction model

As shown in the above figure for the preprocessed nodes (i.e., 500 nodes) provided as input, three hypotheses are applied in our work with the aid of statistical functions using the base station 'BS_i' obtained via its location and the user ID 'UID_i' to extract nodes for identifying influential node in social network. The first hypothesis is expressed as given below.

$$BS_i \rightarrow \sum_{i=1}^{n} \frac{sumof(UID_i)}{N} \quad (4)$$

From the above equation (4), the base station 'BS_i' is recalculated by taking the average of its members (via 'UID') '$\frac{sumof(UID_i)}{N}$'.

The second hypothesis is expressed as given below.

$$BS_i \rightarrow Add(UID_i, P_i) \quad (5)$$

From the above equation (5), the base station 'BS_i' saves the current user 'UID_i'.

$$BS_i \rightarrow Add(Clear(UID_i), P_i) \quad (6)$$

From the above equation (6), the base station 'BS_i' is displaced with the current point 'P_i' by eliminating the current user '$Clear(UID_i)$'. In this way, by employing the location and significance of each node with respect to others, the most vital nodes required for identifying the influential node are extracted. The pseudo code representation of Trilateral Statistical Node Extraction is given below.

Algorithm 2 Trilateral Statistical Node Extraction

Input: Points '$P = P_1, P_2, \dots, P_n$', Base Station '$BS = BS_1, BS_2, \dots, BS_n$'
Output: Optimal Nodes Extracted '$NE = NE_1, NE_2, \dots, NE_n$'
1: **Initialize** transmission range 'TR', threshold 'T', distance threshold 'DisT', users or mobile phones 'N', user threshold 'UIDT'
2: **Begin**
3: **For** each points 'P'
4: Evaluate '$Dis \rightarrow BS_i - P_i$'
5: **If** '$Dis < DisT$' then evaluate first hypothesis using (4)
6: **If** '$Dis < DisT$' && '$UID_i > UIDT$' then evaluate second hypothesis using (5)
7: **If** '$Dis < DisT$' && '$UID_i < UIDT$' then evaluate third hypothesis using (6)
8: **Return** (nodes extracted)
9: **End for**
10: **End**

Algorithm 2 Explanation:

As given in the above Trilateral Statistical Node Extraction algorithm, if the distance 'Dis' is smaller than the threshold distance 'DisT', then this means that the point 'P_i' is detected inside the transmission range 'TR' and hence the point index 'i' is added to prevailing user IDs 'UID_i' and the base station 'BS' with this corresponding location via latitude and longitude is recalculated by taking the mean location of its members (i.e., Users via 'UID'). If the distance 'Dis' is larger than the distance threshold 'DisT' and the number 'N' of its members (i.e., Users via 'UID') is larger than the threshold 'UIDT', then the point 'P_i' is detected outside the transmission range 'TR' and the user points already stored in 'UID_i', therefore the node index 'i' is incremented by 1 to save the current user. Finally, the base station, 'BS_i' is displaced with current point 'P_i' and the 'UID_i' is cleared and then reinitiated with the index 'i'.

3.3 Spearman Correlated Normalized Feature Selection

In this section, with the extracted nodes, features are selected by integrating spearman correlation and a normalization function. Figure 4 shows the block diagram of Spearman Correlated Normalized Feature Selection model

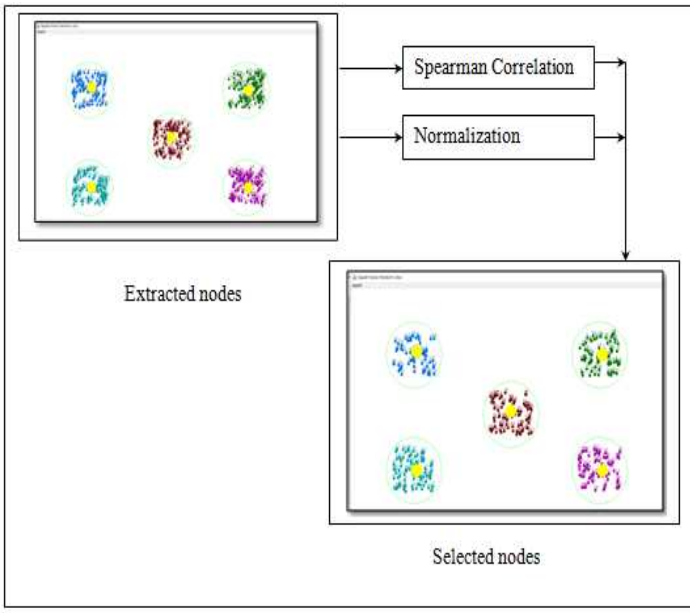


Figure 4 Block diagram of Spearman Correlated Normalized Feature Selection model

As shown in the above figure, the green circle is the transmission range or the boundary of the node (i.e. user ID), the yellow circle is the node centre or the base station location (i.e., the base station) and the small points are locations of the users. For a sample of size 'n', with 'n' representing the extracted features or nodes extracted 'NE', the 'n' raw scores 'NE_i', 'NE_j' are transformed to ranks 'r_{NE_i}', 'r_{NE_j}' and {\displaystyle X_{i},Y_{i}} 'r_s' is evaluated as given below.

$$r_s = \rho r_{NE_i} \rho r_{NE_j} = \frac{\mathrm{cov}\left(r_{NE_i}, r_{NE_j}\right)}{\sigma r_{NE_i}, \sigma r_{NE_j}} \quad (7)$$

From the above equation (7), the rank 'r_s' is measured based on the pearson correlation coefficient 'ρ', covariance of rank of each extracted nodes '$\mathrm{cov}\left(r_{NE_i}, r_{NE_j}\right)$', and the standard deviation of each extracted nodes '$\sigma r_{NE_i}, \sigma r_{NE_j}$' respectively. With distinct integer formulation, the ranks to measure correlation between extracted nodes for selecting normalized nodes are expressed as given below.

$$r_s = 1 - \frac{6\sum[r(NE_i) - r(NE_j)]^2}{n(n^2-1)} \quad (8)$$

From the above equation (8), '$r(NE_i) - r(NE_j)$' refers to the difference between the ranks of two correlated extracted nodes 'NE_i' and 'NE_j' with '' representing the total number of observations.

$$L\left(Prob_{NE_j} | Prob^{BS}\right) = \frac{1 - rs\left(Prob^{BS} * Prob_{NE_j}\right)}{1 - rs\left(Prob^{BS}\right)} * \lambda \left| Prob_{NE_j} \right| \quad (9)$$

This normalization is necessary because of the reason that as we advance through the algorithm, succeeding feature may explain less of the outcome of selected nodes 'NS' and so change in '$1 - rs\left(Prob^{BS} * Prob_{NE_j}\right)$' by splitting '$NE_j$' are smaller. The consequence of this would be a coarser selection of base station, resulting in premature convergence and so the normalization ensures that this is not the case the feature is selected consistently by avoiding premature convergence at each stage of the algorithm. The pseudo code representation of Spearman Correlated Normalized Feature Selection is given below.

Algorithm 3 Spearman Correlated Normalized Feature Selection

Input: Nodes Extracted '$NE = NE_1, NE_2, \ldots, NE_n$'
Output: Normalized node selection '$NS = NS_1, NS_2, \ldots, NS_n$'
1: **Initialize** 'ρ'
2: **Begin**
3: **For** each Nodes Extracted 'NE'
4: Evaluate ranks for each extracted nodes using (7)
5: Evaluate distinct ranks using (8)
6: Perform normalization using (9)
7: **Return** (nodes selected)
8: **End for**
9: **End**

Algorithm 3 Explanation:

As given in the above Spearman Correlated Normalized Feature Selection algorithm, the objective here remains in improving the accuracy involved in tracing the influential nodes in social network. This is achieved by applying two different functions, spearman correlation and normalization. First, by applying the spearman correlation the significance of influential node is shown and allows for further analysis of obtaining the influential node. Next, by employing normalization, premature convergence is avoided and hence has the probability of analyzing all the base stations for obtaining the influential node.

3.4 Katz Centrality Least Angle Influence Node Tracing

Finally, with the selected features or nodes, the actual influence node tracing in social network is performed by applying Katz Centrality to the Least Angle Regression model. Let '$NS = NS_1, NS_2, \dots, NS_p$' be the predictors and 'Y' be the response and let '$NS = [NS_1, NS_2, \dots, NS_p]$' represent the matrix with columns denoting the predictors. Let us also consider the regression coefficient as '$\beta = (\beta_1, \beta_2, \dots, \beta_p)'$' provide us with the estimator response 'Y'' that is expressed as given below.

$$Y' = \sum_{j=1}^{p} NS_j \beta_j \qquad (10)$$

From the above equation (10), the estimator response 'Y'', is obtained by employing the node selected 'NS_j' and the regression coefficient 'β_j' respectively. With this, Katz Centrality is applied to the estimator response for measuring the relative influence and this is expressed as given below.

$$C_{Katz}(Y_i^i) = \alpha \sum_{j=1}^{n} NS_{ji}\, C_{Katz}(Y_j^i) + \beta \qquad (11)$$

From the above equation (11), the Katz Centrality 'C_{Katz} for measuring influence nodes based on neighboring and non-neighboring set is obtained based on the constant '$\alpha = 0.5$'', bias constant 'β'. Moreover, 'NS_{ji}' take value '1' if a predictor node is connected to node 'j' and take value '0' if a predictor node is not connected to node 'j'. Next, with the aid of LAR model, 'β' is selected by minimizing total squared error subject as expressed below.

$$\text{Min}(Y - Y'); \text{Subjectto} \sum_{j=1}^{p} |\beta_j| \le t \qquad (12)$$

From the above equation (12), the estimator response is evaluated based on the minimization of total squared error, therefore considering the neighborhood and non-neighborhood set while identifying the influential nodes in social network. The pseudo code representation of Katz Centrality Least Angle Influence Node Tracing is given below.

Algorithm 4 Katz Centrality Least Angle Influence Node Tracing

Input: Normalized node selected '$NS = NS_1, NS_2, \dots, NS_p$'
Output: Accurate and precise influence node tracing
1: **Initialize** regression coefficient as '$\beta = (\beta_1, \beta_2, \dots, \beta_p)'$',
2: **Begin**
3: **For** each Normalized node selected 'NS'
4: **Evaluate** estimator response 'Y' using (10)
5: Apply Katz Centrality applied to the estimator response using (11) subject to constraint using (12)
6: **Return** (influential node)
7: **End for**
8: **End**

Algorithm 4 Explanation:

As given in the above Katz Centrality Least Angle Influence Node Tracing algorithm, the objective remains in retrieving the influential node in social network by minimizing the total squared error subject to constraint while selecting toe predictors. This is achieved by first considering regression coefficient to accurately derive the relationship between the predictor variable (i.e., nodes selected) and the response (i.e., the influential node). This is obtained by second using a centrality factor considering both the neighboring and non-neighboring node while tracing the influential node.

IV. Experimental Setup and Qualitative Analysis

4.1 Experimental settings

Experimental evaluation of proposed Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) method and existing methods Community Based Influence Maximization Algorithm (CBIMA) [1] and Community based algorithm for Finding Influential Nodes (CFIN) in complex social networks [2] are implemented in Java.

The experimental evaluation is performed with the Telecom Dataset http://sguangwang.com/TelecomDataset.html [20] [21][22] [23]. The dataset, provided by Shanghai Telecom consists of greater than 7.2 million records of effective accessing of the Internet via 3,233 base stations from 9,481 mobile phones acquired for a period of six months. This dataset is very commonly used for influential node selection serving topics such as edge server placement, service migration, service recommendation, etc.For the experimental consideration, six parameters, such as Month, Data, Start Time, End Time, Base Station Location, Mobile Phone ID are taken as input for influential node tracing in social network. The performances of proposed and existing methods are evaluated using different metrics such as sensitivity,

specificity, response time, accuracy and influence spread with a number of nodes.

4.2 Qualitative Analysis

In this section the qualitative analysis with a sample of 10 nodes in social network is analyzed. First, Node Cover Preprocessing (NCP) model is applied to the sample nodes then the information to be transmitted at different time intervals is estimated as given below in table 1 using the deadline of information and start time of information. Followed by which the propagation probability is evaluated based on the distance 'd'.

Table 1 Measure of transmitted information time and propagation probability

S.No	User ID	Using the deadline of information	Start time of information	Transmitted information time	Distance 'd'	Propagation probability
1	T_1	2250	300	1950	2	0.25
2	T_2	2250	350	1900	2	0.25
3	T_3	2250	400	1850	3	0.11
4	T_4	2250	250	2000	2	0.25
5	T_5	2250	500	1750	4	0.06
6	T_6	2250	150	2100	2	0.25
7	T_7	2250	600	1650	4	0.06
8	T_8	2250	200	2050	1	1
9	T_9	2250	800	1450	2	0.25
10	T_{10}	2250	600	1650	2	0.25

Based on the propagation probability from (3), the effective nodes selected are 'T_1', 'T_2', 'T_4', 'T_6', 'T_9' and 'T_{10}' respectively. With the obtained effective nodes, most vital nodes required for identifying the influential node have to be extracted. This is performed by means of distance threshold 'DisT' and user threshold 'UIDT'. Let the distance threshold be assumed to be 'DisT = 4' and user threshold to be assumed to be 'UIDT = 3'.

Table 2 Preprocessed results

User ID	Using the deadline of information	Start time of information	Transmitted information time	Propagation probability
T_1	2250	300	1950	0.25
T_2	2250	350	1900	0.25
T_4	2250	250	2000	0.25
T_6	2250	150	2100	0.25
T_9	2250	800	1450	0.25
T_{10}	2250	600	1650	0.25

Then from the first hypothesis, the base station 'BS_i' is recalculated by taking the average of its members, i.e., given as below.

$$\frac{1950 + 1900 + 2000 + 2100 + 1450 + 1650}{6} = 1841.66$$

Then, the base station 'BS_i' saves the current user 'UID_i' based on the results of the second hypothesis results as 'T_4', 'T_6', 'T_9', 'T_{10}'. The results of the optimal node extraction are given below.

Table 3 Feature extracted results

User ID	Using the deadline of information	Start time of information	Transmitted information time	Propagation probability
T_4	2250	250	2000	0.25
T_6	2250	150	2100	0.25
T_9	2250	800	1450	0.25
T_{10}	2250	600	1650	0.25

Then, extracted features or nodes extracted 'NE' is transformed to rank 'r_{NE_i}', 'r_{NE_j}' and {\displaystyle X_{i},Y_{i}} 'r_s' using (7) and is given below.

Standard deviation of '$\sigma r_{NE_i} = 70.71(T_4, T_6)$'
Standard deviation of '$\sigma r_{NE_j} = 141.42(T_9, T_{10})$'
Covariance of $\mathbf{cov}\left(r_{NE_i}, r_{NE_j}\right) = 5000$

The rank 'r_s' is measured based on the pearson correlation coefficient 'ρ', covariance of rank of each extracted nodes '$\mathrm{cov}\left(r_{NE_i}, r_{NE_j}\right)$', and the standard deviation of each extracted nodes '$\sigma r_{NE_i}, \sigma r_{NE_j}$' as given below

$$r_s = \frac{5000}{70.71 * 141.42} = 0.50$$

Normalization for 'NE_j' is evaluated as given below.

$$\frac{1-0.50(0.25*141.42)}{1-0.50(0.25)}*0.01\ (141.42)$$

$$\frac{1-17.6775}{1-0.125}*1.41=2.70$$

Normalization for 'NE_i' is evaluated as given below.

$$\frac{1-0.50(0.25*70.71)}{1-0.50(0.25)}*0.01\ (70.71)$$

$$\frac{1-8.83875}{1-0.125}*0.77781=6.89$$

So the features selected are from '$NE_i = T_4, T_6$'

Table 4 Feature selected results

User ID	Using the deadline of information	Start time of information	Transmitted information time	Propagation probability
T_4	2250	250	2000	0.25
T_6	2250	150	2100	0.25

Finally, the influential node is estimated based on the Katz Centrality using (11) and (12). Let us assume the regression coefficient 'β_j' to be between '{0.7to 1.0}' and constant '$\alpha = 0.5$', then, the estimator response for node ID 'T_4' is given as below.

$$0.5*1*y+0.7=y*0.5=-0.7$$

$$y=-\frac{0.7}{0.5}=-1.4$$

In a similar manner, the estimator response for node ID 'T_6' is given as below.

$$0.5*1*y+0.8=y*0.5=-0.8$$

$$y=-\frac{0.8}{0.5}=-1.6$$

Therefore, the influential node is 'T_6'.

V. RESULTS & DISCUSSION

5.1 Case 1: Sensitivity

In influence node tracing sensitivity is a measure of how well a test can identify true positives. In other words, the sensitivity is its potentiality to determine the influential node tracing correctly. To estimate it, we should calculate the ratio of true positive in influential node tracing cases. Mathematically, this can be stated as given below.

$$\text{Sensitivity}=\frac{TP}{TP+FN} \quad (13)$$

From the above equation (13), the sensitivity rate 'Sensitivity' is measured based on the true positive 'TP' (i.e., the number of nodes correctly identified as influential node if so) and the false negative 'FN' (i.e., the number of nodes incorrectly identified as normal node) respectively. The sensitivity of a test can assist in showing how well it can classify samples (i.e. nodes) that have the condition. In other words, a high sensitivity value means a test correctly classifies a sample. Table 5 given below shows the sensitivity rate of the proposed TSKC-LAR and existing methods, CBIMA [1] and CFIN [2].

Table 5 Performance measure of sensitivity

Nodes	Sensitivity (%)		
	TSKC-LAR	CBIMA	CFIN
500	80.76	72.72	65.5
1000	80.25	71.35	64.85
1500	78.95	71.15	64.55
2000	78.35	71	64.15
2500	78.15	70.85	64
3000	77.45	70.65	63.95
3500	77.25	70.25	63.75
4000	77	70	63.55
4500	76.55	69.55	62.55
5000	75	69.25	62

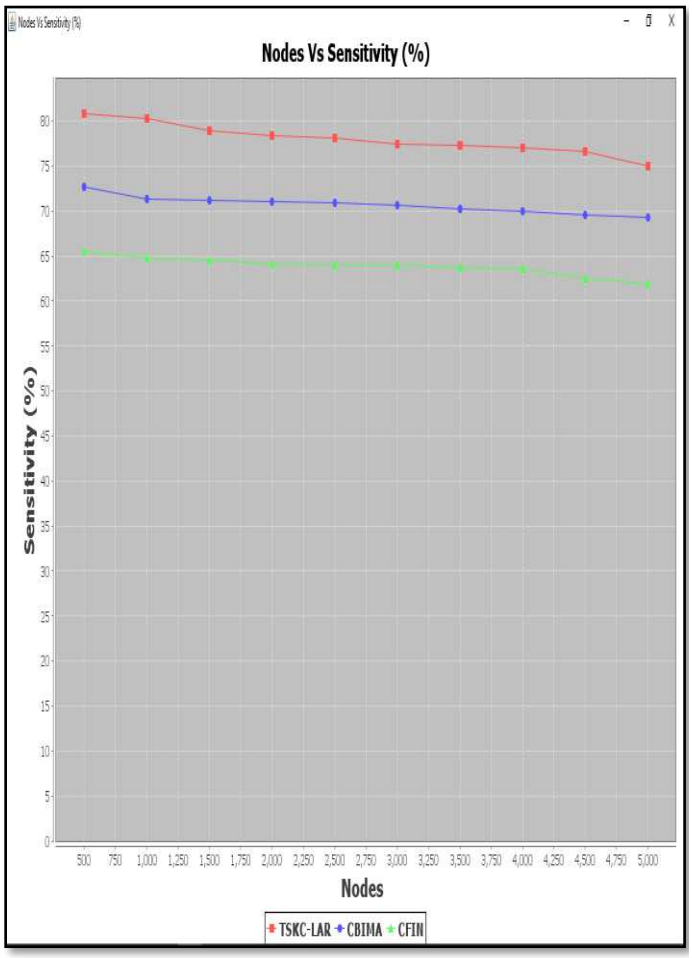


Figure 5 Graphical representation of sensitivity

Figure 5 given above shows the sensitivity measure for tracing the influential node in social network. A simulation of 5000 nodes with unique user ID (i.e., mobile phone) is considered and the sensitivity rate is measured accordingly. With '500' nodes considered for simulation, '42' number of nodes correctly identified as influential node and '10' number of nodes incorrectly identified as normal node using TSKC-LAR, '40' number of nodes correctly identified as influential node and '15' number of nodes incorrectly identified as normal node using [1] and '38' number of nodes correctly identified as influential node and '20' number of nodes incorrectly identified as normal node using [2], the sensitivity rate was observed to be 80.76%, 72.72% and 65.5% respectively. From the simulation results it is inferred that the sensitivity rate though decreases with the increase in the number of nodes, the sensitivity rate using TSKC-LAR is found to be better than [1] and [2]. This improvement is due to the application of Node Cover Preprocessor algorithm. By applying this algorithm, initially the effective time is measured and then followed by which the gradual loss based on the distance is evaluated. Finally, propagation probability is estimated to arrive at the effective nodes and discard the ineffective nodes. With this the sensitivity rate using TSKC-LAR method is found to be improved by 10% compared to [1] and 22% compared to [2].

5.2 Case 2: Specificity

In influential node tracing, specificity is a measure of how well a test can identify true negatives. It refers to the percentage of the true negatives out of all the samples that do not have the condition (true negatives and false positives). Mathematically, this can be stated as given below.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (14)$$

From the above equation (14), the specificity rate 'Specificity' is measured based on the true negative 'TN' (i.e., number of nodes correctly identified as normal node) and the false positive 'FP' (i.e., the number of nodes incorrectly identified as influential node). A test with a high specificity value means that it is correctly classifies the nodes with the condition more often than a test with a lower specificity. Table 6 given below shows the specificity rate of the proposed TSKC-LAR and existing methods, CBIMA [1] and CFIN [2].

Table 6 Performance measure of specificity

Nodes	Specificity (%)		
	TSKC-LAR	CBIMA	CFIN
500	96.2	94.73	90.66
1000	95.25	93.15	88.15
1500	95.15	92.55	87.35
2000	94.85	91.35	86
2500	94.35	90.45	85.25
3000	94	89.25	84.15
3500	93.95	89	84
4000	93.75	88.15	83.85
4500	93.25	88.05	83
5000	93	87	82

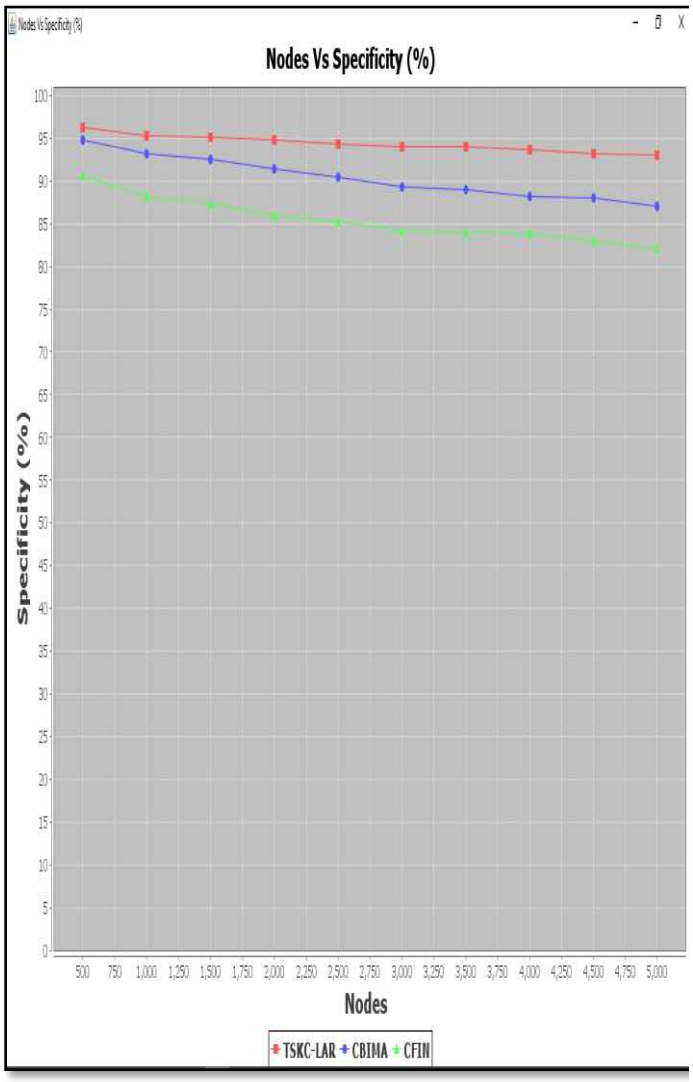


Figure 6 Graphical representation of specificity

Figure 6 given above illustrates the specificity rate for different numbers of nodes for accessing the internet through base station. With this specificity rate evaluation the service migration and service recommendation can be made in an efficient manner. With '500' nodes considered for simulation, '380' number of nodes correctly identified as normal node and '15' number of nodes incorrectly identified as influential node using TSKC-LAR and existing methods, '360' number of nodes correctly identified as normal node and '20' number of nodes incorrectly identified as influential node using CBIMA [1] and '340' number of nodes correctly identified as normal node and '35' number of nodes incorrectly identified as influential nodeusing CFIN [2], the overall specificity was observed to be 96.2%, 94.73% and 90.66% respectively. From the results it is inferred that the specificity rate using TSKC-LAR is better than [1] and [2]. The specificity rate improvement is due to the application of propagation probability that evaluates the distance on the basis of the neighborhood factor therefore assisting in removing the ineffective nodes. With this the specificity of TSKC-LAR method is said to be improved by 4% compared to [1] and 11% compared to [2].

5.3 Case 3: Accuracy

The third parameter used for analysis is the rate of accuracy. The accuracy of a test is its potentiality to distinguish the normal and influential nodes correctly. To evaluate the accuracy of a test, the ratio of influential nodes correctly predicted as it is and the nodes for evaluation has to be considered. Mathematically, this can be stated as:

$$Acc = \sum_{i=1}^{n} \frac{N_{CR}}{N_i} * 100 \qquad (15)$$

From the above equation (15), the accuracy rate 'Acc' is evaluated based on the sample nodes involved in influential node tracing 'N_i' and the nodes correctly traced 'N_{CR}'. It is expressed in terms of percentage (%).Table 7 given below shows the accuracy rate of the proposed TSKC-LAR and existing methods, CBIMA [1] and CFIN [2].

Table 7 Performance measure of accuracy

Nodes	Accuracy (%)		
	TSKC-LAR	CBIMA	CFIN
500	96	93	90
1000	95.85	92.15	89.35
1500	95.25	92	89.15
2000	95	89.85	88.15
2500	94.35	89.35	86.35
3000	94.15	89	86
3500	94	88.55	85.25
4000	93.75	88.25	85
4500	93.55	88	84.15
5000	93	87.15	84

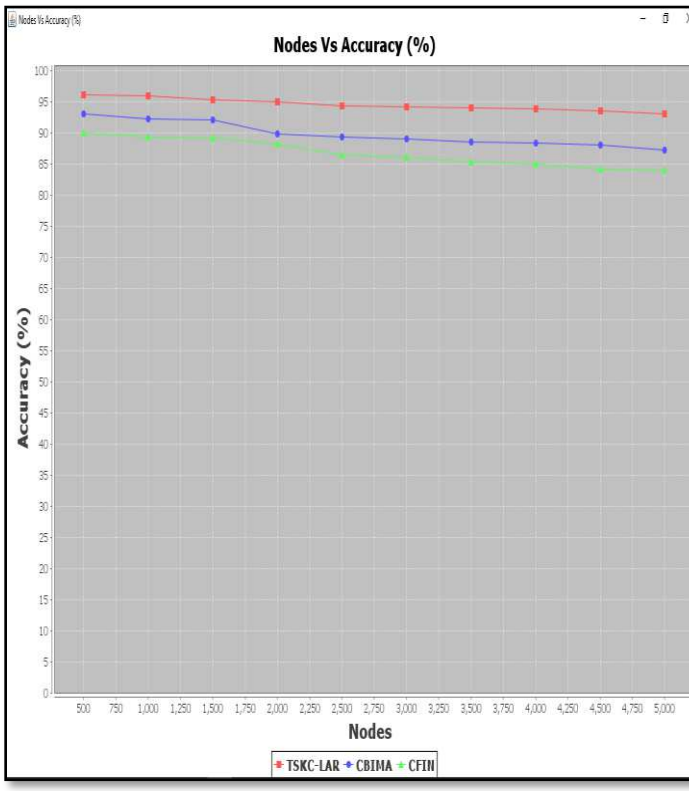


Figure 7 Graphical representation of accuracy

Figure 7 given above shows the accuracy involved in tracing the influential nodes in social network based on the trajectory of users. From the figure it is inferred that the accuracy rate is said to be inversely proportional to the nodes considered for simulation. In other words, increasing the nodes causes an increase in the overall nodes involved in influential node tracing and therefore reducing the accuracy rate. However, experimental simulation shows betterment using TSKC-LAR when compared to CBIMA [1] and CFIN [2].With '500' nodes considered for simulation and '480' nodes correctly traced using TSKC-LAR, '465' nodes correctly traced using [1] and '450' nodes correctly traced using [2], the overall accuracy was found to be 96%, 93% and 90% respectively. The reason behind the improvement in the accuracy rate is due to the application of Trilateral Statistical Node Extraction algorithm. By applying this algorithm, optimal nodes for influential node tracing is said to be extracted. Here, the point detected inside the transmission range along with the corresponding location via latitude and longitude is taken into account for obtaining optimal nodes. With this, the accuracy using TSKC-LAR is said to be improved by 5% compared to [1] and 9% compared to [2].

5.4 Case 4: Response Time

The response time in our work refers to the time consumed in identifying influential node in social network. Mathematically, this can be stated as:

$$RT = \sum_{i=1}^{n} N_i * Time[NTracing] \quad (16)$$

From the above equation (16), the response time 'RT' is measured based on the sample nodes involved in simulation 'N_i' and the time consumed in node tracing 'Time[NTracing]'. It is measured in terms of milliseconds (ms). Table 8 given below shows the response time of the proposed TSKC-LAR and existing methods, CBIMA [1] and CFIN [2].

Table 8 Tabulation for response time

Nodes	Response time (ms)		
	TSKC-LAR	CBIMA	CFIN
500	92.5	107.5	115
1000	103.25	120.35	130.35
1500	125.55	135.55	145.55
2000	155.85	190.25	200.15
2500	190.35	210.35	225.55
3000	215.55	220.54	235.55
3500	225.55	245.55	260.15
4000	240.75	280.25	315.55
4500	265.35	310.35	325.55
5000	280.25	325.55	340.15

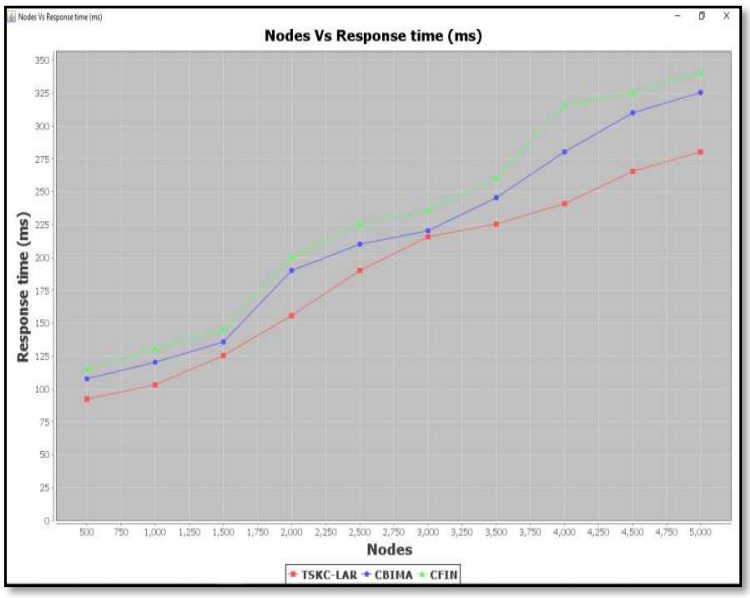


Figure 8 Graphical representation of response time

Figure 8 given above illustrates the response time involved in tracing the influential nodes in social network to the distinct user ID obtained at different time intervals. However, for fair comparison same numbers of nodes and user IDs are utilized in all the three methods for measuring the response time. From the figure the response time is directly proportional to the nodes, i.e., increasing the numbers of nodes causes increase in the node deployment and consequently a linear rise is said to be observed in the response time. However, with '500' nodes considered for experimentation and the time consumed in node trace with single node involved in the network framework being '0.185ms' using TSKC-LAR, the time consumed in node trace with single node involved in the network framework being '0.215ms' using [1] and '0.230ms' using [2]. From these results it is inferred that the response time is comparatively lesser using TSKC-LAR than [1] and [2]. The reason behind the improvement is due to the extraction of optimal nodes for influential node tracing via Trilateral Statistical function. By applying this function, three hypotheses are checked by using the location and node significance with respect to others, the most vital nodes required for identifying the influential node are extracted. With these nodes, the influential nodes are identified in minimum time using TSKC-LAR by 12% compared to [1] and 17% compared to [2].

5.5 Case 5: Influence Spread

Finally, the influence spread over the whole network is analyzed in this section. Influence maximization refers to the problem of identifying a small subset of nodes (seed nodes) in social network that maximize the spread of influence. Table 9 given below shows the influence spread using three methods, TSKC-LAR, CBIMA [1] and CFIN [2].

Table 9 Tabulation results of influence spread

Seed set size	Influence spread ()		
	TSKC-LAR	CBIMA	CFIN
10	900	750	600
20	1200	900	750
30	1400	1150	1000
40	1750	1300	1050
50	1900	1550	1250

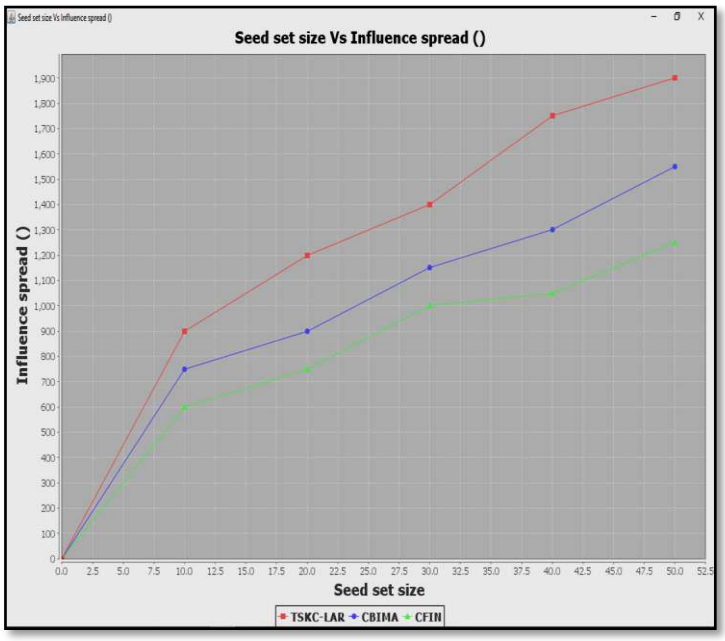


Figure 9 Graphical representation of influence spread

Figure 9 given above shows the influence spread for different seed set size in the range of 10 to 50. From the figure it is inferred that the influence spread is maximized in all the three methods and comparatively found better using TSKC-LAR method than CBIMA [1] and CFIN [2]. The influence spread maximization was achieved using TSKC-LAR method due to the application of Katz Centrality Least Angle Influence Node Tracing algorithm. By applying this algorithm, while retrieving the influential node in social network, the total squared error was minimized while selecting the predictors. Also, relationship between the predictor variable (i.e., nodes selected) and the response (i.e., the influential node) were estimated on the basis of regression factor. Finally, both the neighboring and non-neighboring node was involved while tracing the influential node based on the centrality factor. This in turn improved the influence spread using TSKC-LAR method by 26% compared to [1] and 54% compared to [2].

VI. Conclusion

An efficient Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR) for influential node tracing in social network is developed with objective of increasing the influence spread. The key objective of TSKC-LAR method is to ensure influence maximization and minimize response time during influence tracking. The objective of TSKC-LAR method is attained with application of Trilateral Statistical Node Extraction and Katz Centrality Least Angle Influence Node Tracing algorithm. First effective nodes were identified by means of Node Cover Preprocessor. Next, by employing Trilateral Statistical Node Extraction algorithm optimal nodes were extracted via three hypotheses. Moreover, to avoid premature convergence, normalized node selection was performed by applying normalization and spearman correlation. Finally, influential node in social network was identified by minimizing the total squared error. The efficiency of TSKC-LAR method is estimated in terms of

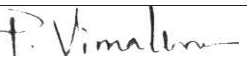
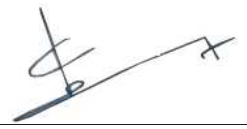
sensitivity, specificity, accuracy, influence spread and response time compared with state-of-the-art works. The simulation results shows that the TSKC-LAR method presents better performance with an enhancement of influence spread and minimization of response time when compared to the state-of-the-art works.

Declaration

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from <mevimal2016@gmail.com>

Author name	Signature	Date
P. Vimal kumar (Corresponding Author)		27/02/2021
C. Balasubramanian		27/02/2021

Funding: Not applicable

Conflicts of interest: No

References

[1] Zufan Zhang, Xieliang Li, Chenquan Gan, “Identifying influential nodes in social networks via community structure and influence distribution difference”, Digital Communications and Networks, Elsevier, Apr 2020 [Community Based Influence Maximization Algorithm (CBIMA)]

[2] Mohammad Mehdi Daliri Khomami, Alireza Rezvanian, Mohammad Reza Meybodi, Alireza Bagheri1, “CFIN: A community‑based algorithm for finding influential nodes in complex social networks”, The Journal of Supercomputing, Springer, Jun 2020

[3] Debadatta Naik, Ranjan Kumar Behera, Dharavath Ramesh, Santanu Kumar Rath, “Map‑Reduce‑Based Centrality Detection in Social Networks:An Algorithmic Approach”, Arabian Journal for Science and Engineering, Springer, Jun 2020

[4] Fabíola S. F. Pereira, João Gama, Sandra de Amo, Gina M. B. Oliveira, “On analyzing user preference dynamics with temporal socialnetworks”, Machine Learning, Springer, Jul 2018

[5] Ahmad Zareie, Amir Sheikhahmadi, Mahdi Jalili, Mohammad Sajjad Khaksar Fasaei, “Finding influential nodes in social networks based on neighborhoodcorrelation coefficient”, Knowledge-Based Systems, Elsevier, Jan 2020

[6] Nan Zhao, Jingjing Bao, and Nan Chen, “Ranking Influential Nodes in Complex Networks withInformation Entropy Method”, Complexity, Wiley, Jun 2020

[7] Shan Tian, Songsong Mo, Liwei Wang, Zhiyong Peng, “Deep Reinforcement Learning‑Based Approach to Tackle Topic‑AwareInfluence Maximization”, Data Science and Engineering, Springer, Feb 2020

[8] Gouheng Zhao, Peng Jia, Cheng Huang, Anmin Zhou, Yong Fang, “A Machine Learning Based Framework for Identifying Influential Nodes in Complex Networks”, IEEE Access, Mar 2020

[9] Weimin Li, Yuting Fan, Jun Mo, Wei Liu, Can Wang, Minjun Xin, Qun Jin, “Three-hop velocity attenuation propagation modelfor influence maximization in social networks”, World Wide Web, Springer, Oct 2019

[10] Mohammed Alshahrani, Zhu Fuxi, Ahmed Sameh, Soufiana Mekouar, Sheng Huang, “Efficient algorithms based on centrality measures for identification of top-K influential users in social networks”, Information Sciences, Elsevier, Mar 2020

[11] Gouheng Zhao, Peng Jia, Anmin Zhou, Bing Zhang, “InfGCN: Identifying influential nodes in complex networks with graphconvolutional networks”, Neurocomputing, Elsevier, Jul 2020

[12] Tinghuai Ma, Hongmei Wang, Lejun Zhang, Yuan Tian, Najla Al-Nabhan, “Graph classification based on structural features of significant nodes andspatial convolutional neural networks”, Neurocomputing, Elsevier, Oct 2020

[13] Jung Hyuk Seo, Myoung Ho Kim, “Finding influential communities in networks with multipleinfluence types”, Information Sciences, Elsevier, Oct 2020

[14] Shan Tian, Songsong Mo, Liwei Wang, Zhiyong Peng, “Deep Reinforcement Learning‑Based Approach to Tackle Topic‑AwareInfluence Maximization”, Data Science and Engineering, Springer, Feb 2020

[15] Huiyu Min, Jiuxin Cao, Tangfei Yuan, Bo Liu, “Topic based time-sensitive influence maximizationin online social networks”, World Wide Web, Springer Nature, Jan 2020

[16] Guojie Song, Yuanhao Li, Xiaodong Chen, and Xinran He, “Influential Node Tracking on Dynamic SocialNetwork: An Interchange Greedy Approach”, IEEE. Translations and content mining, Sepl 2016

[17] Wenjun Wang and W. Nick Street, “Modeling and maximizing influencediffusion in social networks for viral marketing”, Applied Network Science, Springer, Oct 2018

[18] Hao Wei, Zhisong Pan, Guyu Hu, Liangliang Zhang, Haimin Yang, Xin Li, Xingyu Zhou, “Identifying influential nodes based onnetwork representation learning in complexnetworks”, PLOS ONE | https://doi.org/10.1371/journal.pone.0200091 July 9, 2018

[19] Rodrigo Olivares, Francisco Muñoz, Fabián Riquelme, “A multi-objective linear threshold influence spread model solved byswarm intelligence-based methods”, Knowledge-Based Systems, Elsevier, Dec 2020

[20] S. Wang, Y. Zhao, J. Xu, J. Yuan, C. Hsu, Edge Server Placement in Mobile Edge Computing, Journal of Parallel and Distributed Computing, vol, 127, pp. 160-168, 2019

[21] S. Wang, Y. Zhao, L. Huang, J. Xu, C. Hsu. QoS Prediction for Service Recommendations in Mobile Edge Computing, Journal of Parallel and Distributed Computing,vol. 127, pp.134-144, 2019

[22] Y. Guo, S. Wang, A. Zhou, J. Xu, J. Yuan, C. Hsu. User Allocation-aware Edge Cloud Placement in Mobile Edge Computing, Software: Practice and Experience, vol. 50, no. 5, pp. 489-502, 2020

[23] S. Wang, Y. Guo, N. Zhang, P. Yang, A. Zhou, X. Shen. Delay-aware Microservice Coordination in Mobile Edge Computing: A Reinforcement Learning Approach, IEEE Transactions on Mobile Computing, https://ieeexplore.ieee.org/document/8924682, 2019

Figures

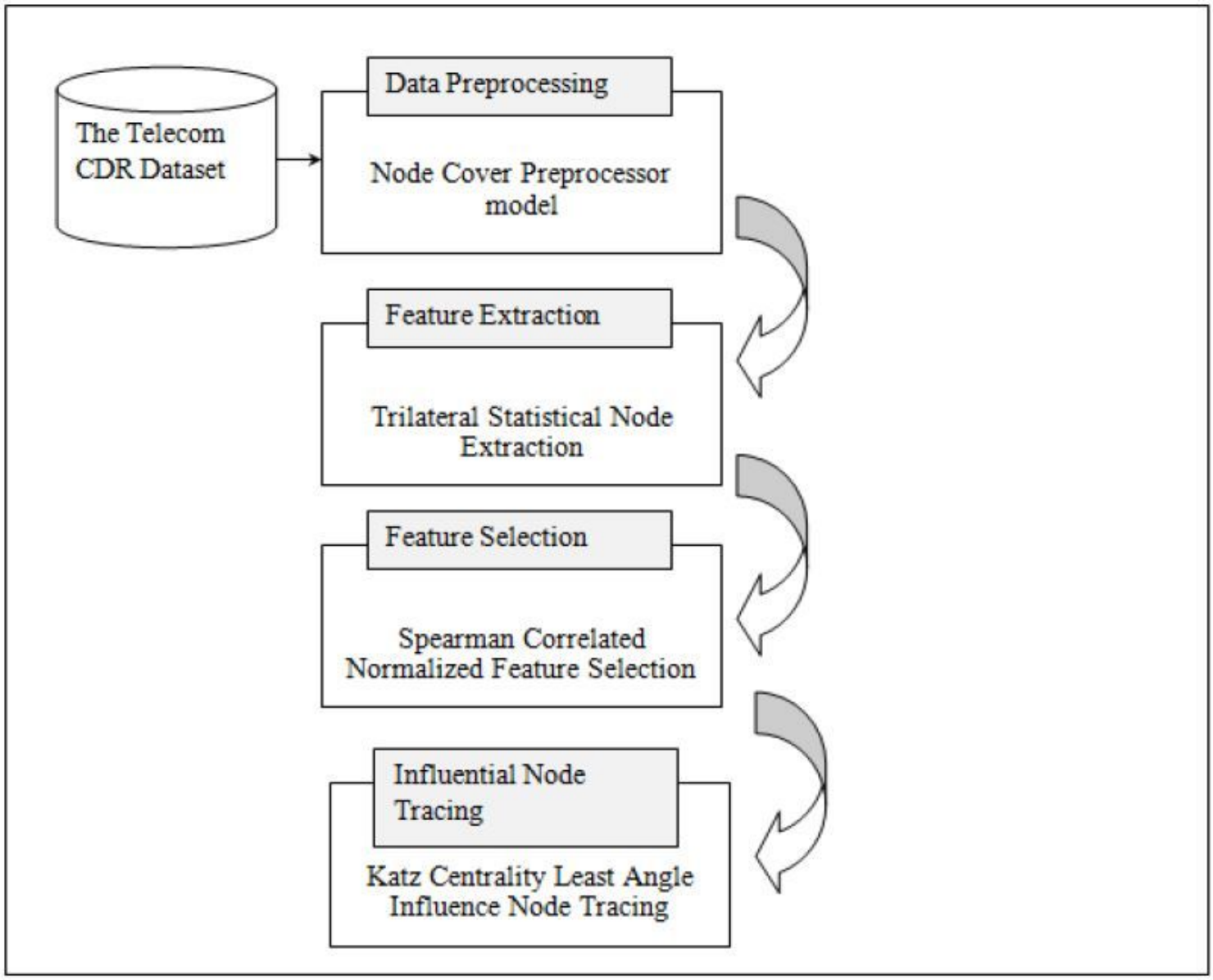


Figure 1

Block diagram of Trilateral Spearman Katz Centrality-based Least Angle Regression (TSKC-LAR)

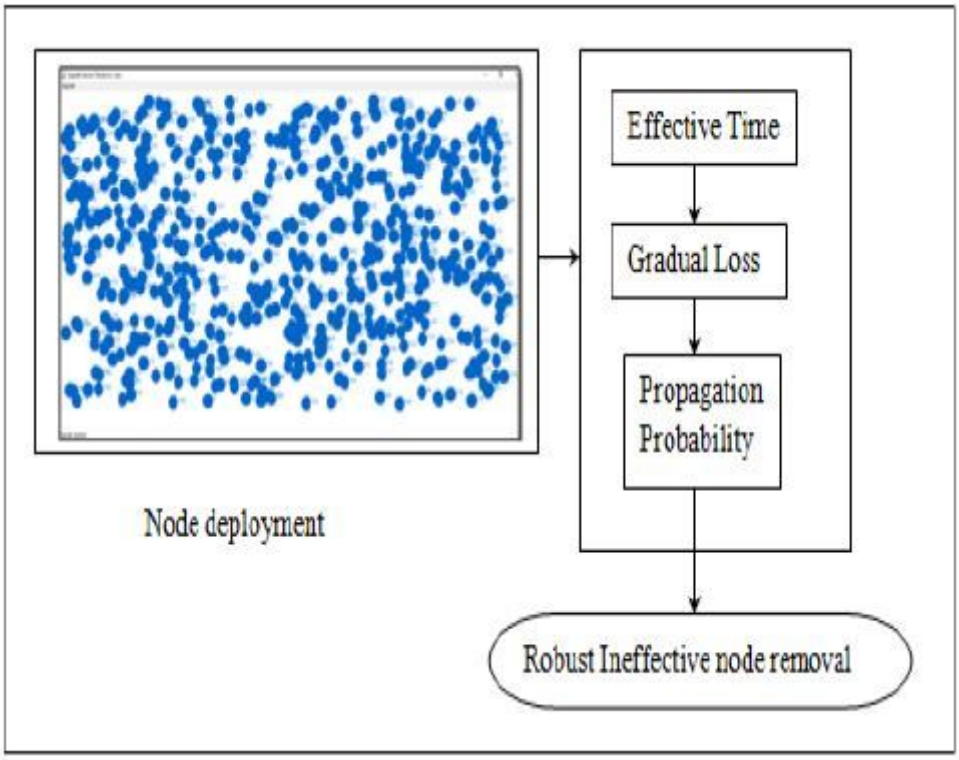


Figure 2

Block diagram of Node Cover Preprocessor model

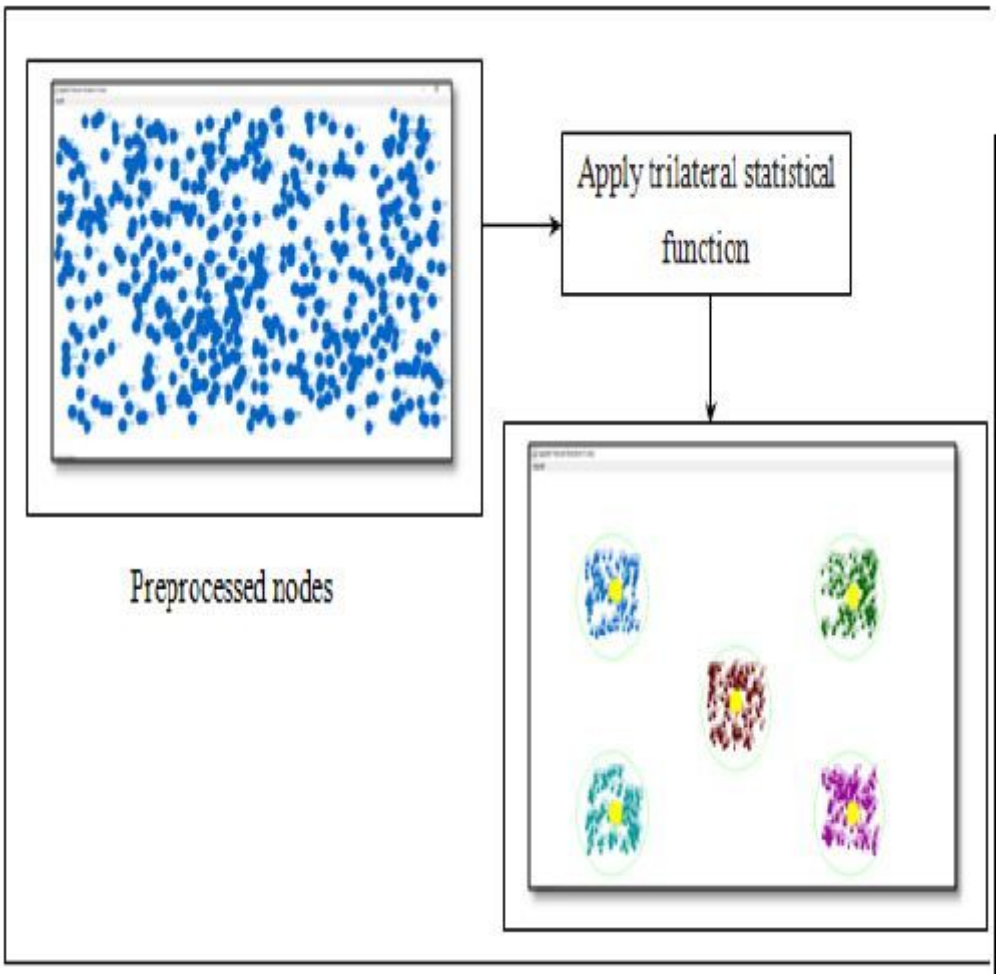


Figure 3

Block diagram of Trilateral Statistical Node Extraction model

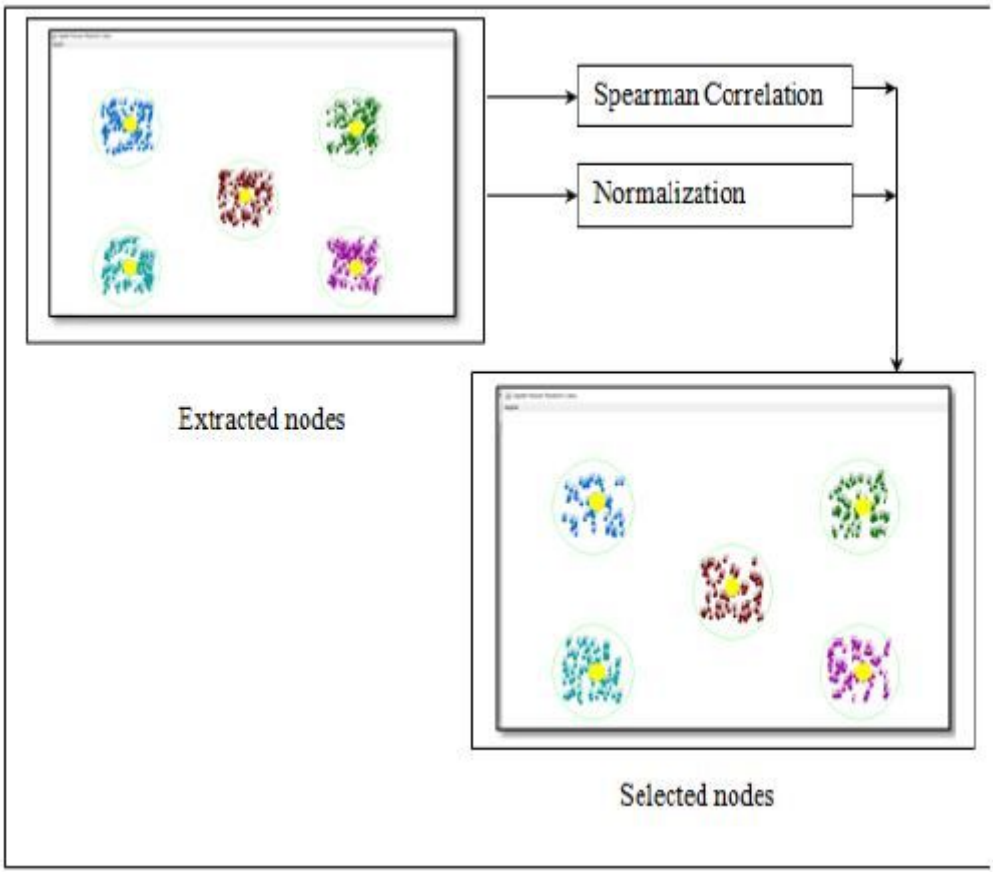


Figure 4

Block diagram of Spearman Correlated Normalized Feature Selection model

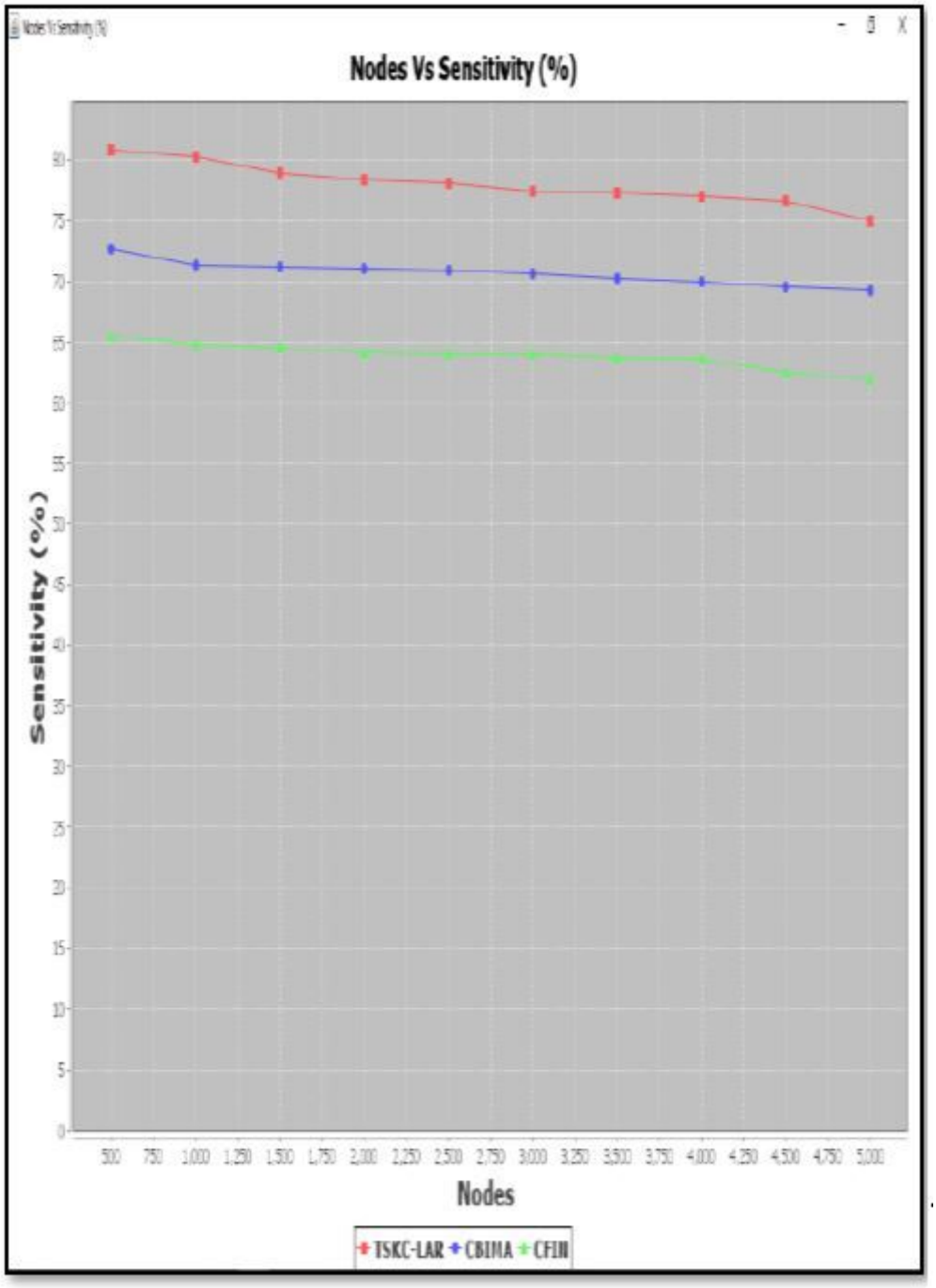


Figure 5

Graphical representation of sensitivity

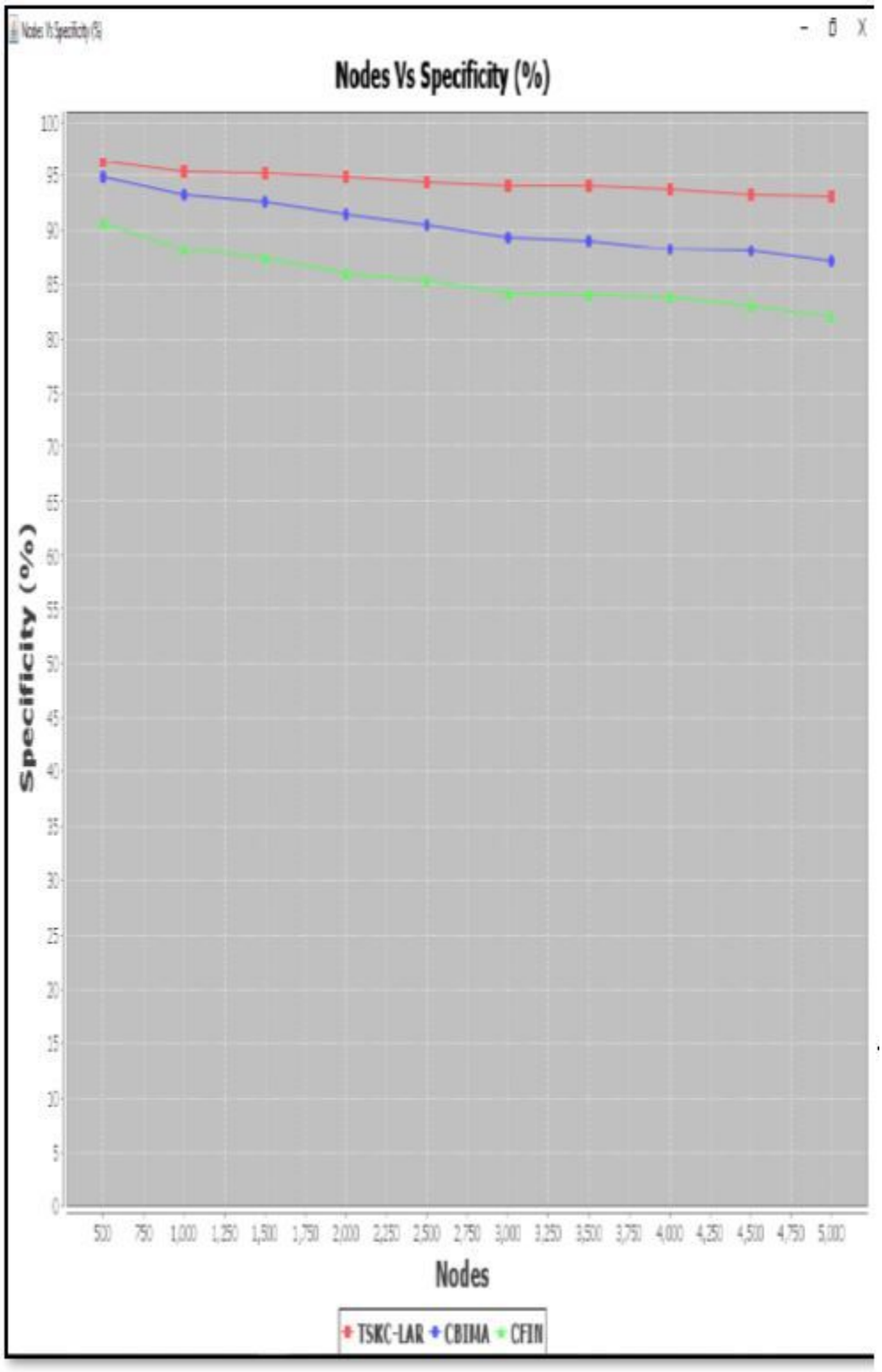


Figure 6

Graphical representation of specificity

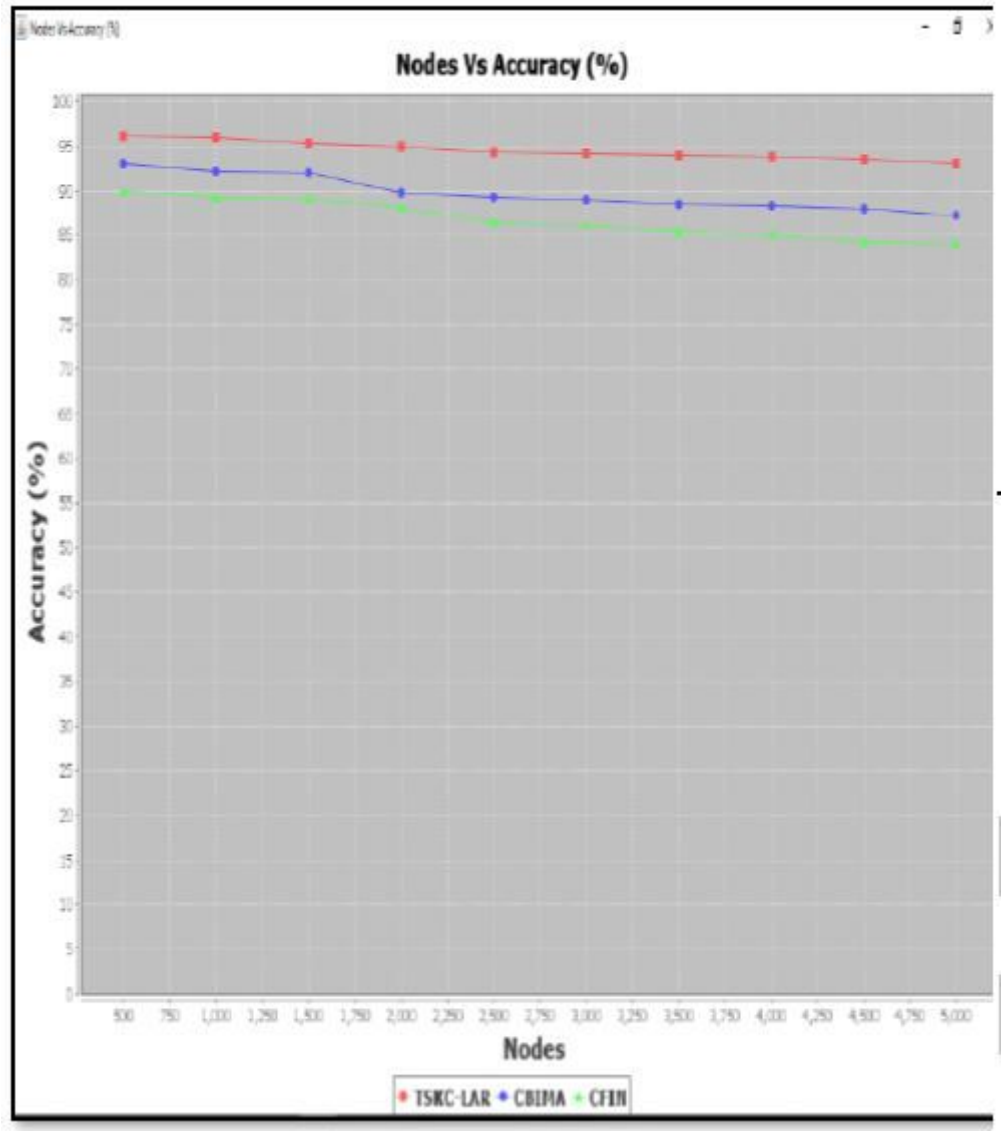


Figure 7

Graphical representation of accuracy

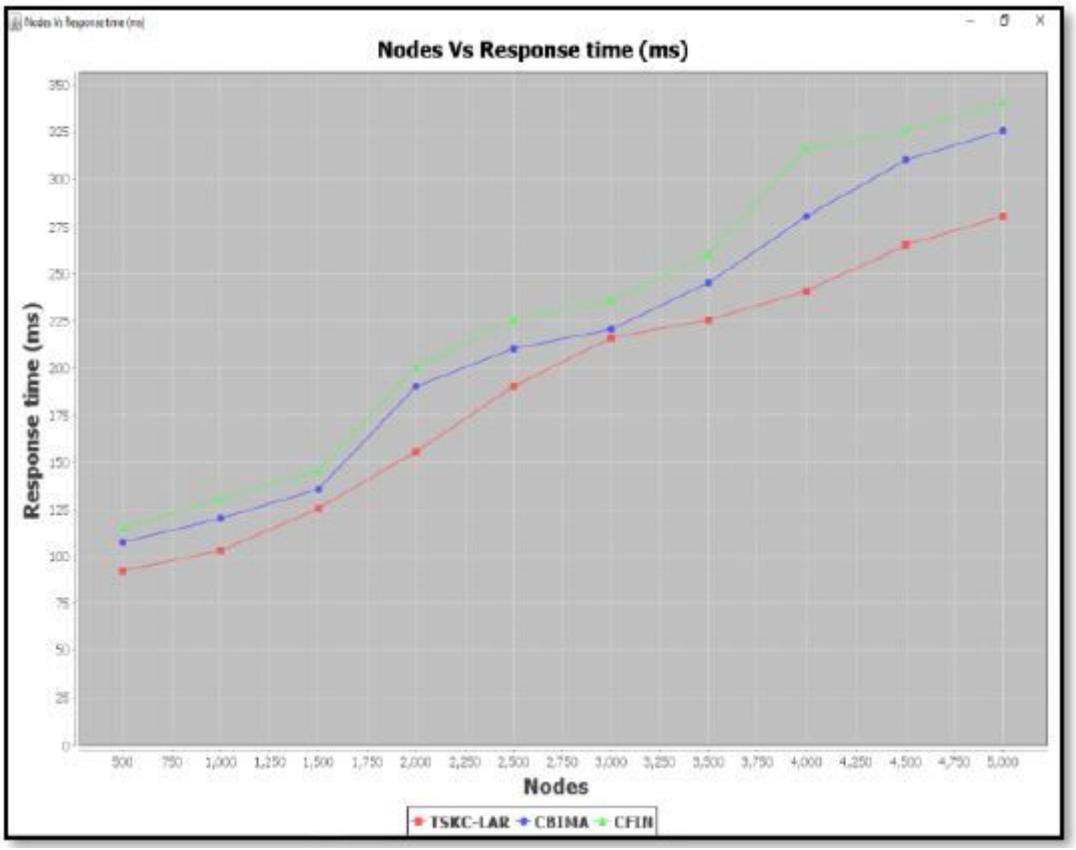


Figure 8

Graphical representation of response time

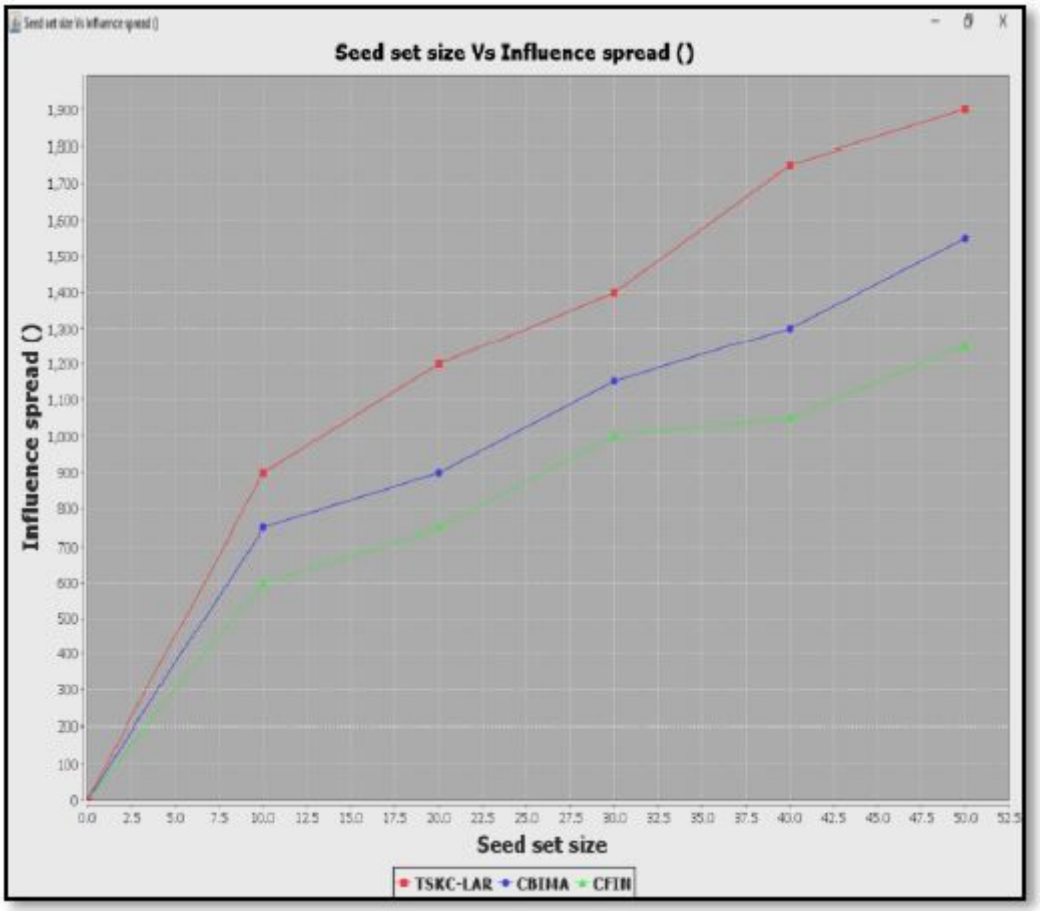


Figure 9

Graphical representation of influence spread