Efficient Algorithm for Maximal Clique Size Evaluation

Broad Learning of its Relation with Centrality Metrics for Large Dataset Networks

Ubaida Fatima[1], Dr. Saman Hina[2]
Department of Mathematics[1]
Department of Computer Science and IT[2]
NED University of Engineering and Technology, Karachi, Pakistan[1, 2]

***Abstract*—A large dataset network is considered for computation of maximal clique size (MC). Additionally, its link with popular centrality metrics to decrease uncertainty and complexity and for finding influential points of any network has also been investigated. Previous studies focus on centrality metrics like degree centrality (DC), closeness centrality (CC), betweenness centrality (BC) and Eigenvector centrality (EVC) and compare them with maximal clique size however, in this study Katz centrality measure is also considered and shows a pretty robust relation with maximal clique size (MC). Secondly, maximal clique size (MC) algorithm is also revised for network analysis to avoid complexity in computation. Association between MC and five centrality metrics has been evaluated through recognized methods that are Pearson's correlation coefficient (PCC), Spearman's correlation coefficient (SCC) and Kendall's correlation coefficient (KCC). The strong strength of association between them is seen through all three correlation coefficients measure.**

Keywords*—Centrality measures; network analysis; maximal clique size

I. Introduction

This Network analysis has become a crucial tool in studying the patterns involved in branched systems and graphs. From its initial journey of solving bridges by Euler all the way back in 1735, network analysis and graph theory have greatly evolved and found applications in nearly every area of study. Since, these analyses involving the exchange of information/resources between 'actors' (nodes) fields like big data science, health care, finance, computer science, social sciences, etc. have grown as a result of efficient use of networking techniques [1].

More specifically Complex Network Analysis has emerged as a major area of research in big data science. The aim of this approach is to analyze real life complex network models using the approaches of graph theory. Numerous approaches have been developed for the analysis of networks; centrality measures have really contributed to the understanding of these networks. Node centrality is a prominently used measure, it links one node with others in the network based on a statistical quantitative measuring of the topological importance of the node with respect to the others [2]. In general, node centrality can help study a wide range of measures ranging from sports associated patterns of play, to identify user preferences in social networks, the most used clinics in urban and rural settings, to even the super-spreaders of a disease, etc. The existing techniques for evaluating the centrality measures involve a neighborhood-based approach and a shortest path algorithm approach. The neighborhood approach makes use of the key features of a node such as the degree centrality (DC) and Eigenvector centrality (EVC), while the shortest path approach utilizes the betweenness centrality (BC) and the closeness centrality (CC) measures [3].

Due to being computationally easier to manage, numerous variations (spatial and temporal) of the algorithms for determining centrality metrics have been developed. However, one question associated with centrality of a node is usually the allowable size of a 'clique' for a node. A graph contains a "clique" that is a set of some nodes such that each two different nodes are adjacent. The size of a clique is defined as the count of nodes that are present in the clique. Every node of a graph might be a piece of one or more than one cliques of different sizes. In networks which are highly linked and have complex interactions, this maximum size of a clique can help identify whether a node in particular is of importance in a community or not based on its modular score. The modularity score is a measure of effectiveness of a networks partitioning into communities. A larger modularity score means a highly inter related community with a high number of vertices within it. Hence, it becomes imperative to identify the vertices that are scored high on the modular scale and design algorithms for the detection of a community using these vertices [2].

The paper is oriented as follows: literature review on network analysis is mentioned in Section II. Section III of this paper contains network analysis through centrality metrics. Revised maximal clique size algorithm evaluation for small network was done in Section IV. Results of centrality metrics and maximal clique size for large product network data is discussed in Section V. Conclusion and future work is presented in Section VI.

II. Literature Review

Stattner and Vidot (2011) presented new favorable circumstances in the area of social networks to comprehend the outbreak of infectious diseases as these events have been increasing rapidly such as the spread of H1N1 influenza virus. Hence the hindrance and regulation of outbreaks have become a health problem of fundamental importance. In this study, the methods already used in epidemiology and those which are recent both are focused in order to apply modeling on disease spreads and overviewed possible future implementations on social network analysis [2].

Zhnag et al. (2015) highlighted the point that common models such as SIR model overlooks the flocking or protection consequences and thus may have some improbable assumptions. Therefore, in this study an improved SIR model is proposed in which these consequences are considered. Both stochastic as well as deterministic models are used to identify the outbreaks on social networks. The results obtained from both of the simulations show that diseases spread even more in social contact networks having greater average of degree. Some dormant immunization strategies have been presented in this work as well to support the findings [3].

Lawyer (2015) mentioned that the spreading power of all nodes in a network should be identified as every vertex in a network generates some force for the distribution of infection, and the recently used centrality measures like eigenvalue, degree or k-shell centrality can be used to accurately identify the nodes that are most influential but not for the nodes that are not much influential. It was concluded that the resulted metric and expected force accurately evaluates the spreading power of all nodes in social contact networks. The force may be estimated independently for each vertex that may be applicable for networks with dynamic or very large adjacency matrix [4].

Yin et al. (2017) proposed a modified SIS model, which contains the property that in social contact networks a vertex along with its neighbor nodes also contacts to the other ones randomly that do not have direct connections that may be called as stranger contacts. This modified model is implemented on a scale-free network and the impact of these different contact patterns are studied on the dynamics of epidemics. This study concluded that the more partiality for direct contacts, the less likely would be the outbreak of disease. Furthermore, the finest strategy of disease control is to adjust both of the number of contact patterns [5].

Meghanathan (2017) explained betweenness centrality metric for complex graphs. Association among betweenness and Local clustering coefficient was discussed. Local clustering coefficient- based degree centrality measure was stated and studied with betweenness centrality on real-world datasets [6].

Meghanathan (2018) identified the relationship among vital centrality measures that are easily computed and maximal clique size which is complex in computation. The association was studied on 10 real-world datasets between centrality metrics and maximal clique size through three well known correlation coefficients that are Pearson's, Spearman's and Kendall's [2].

III. Network Graph and Important Metrics in Network Analysis

Key nodes can be recognized in a given social network by looking forward to the following metrics:

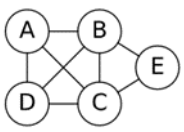

Fig. 1. Undirected Network of 5 Nodes [7].

Adjacency matrix (Ad) for the undirected network containing 5 vertices (nodes) of Fig. 1 is demonstrated as:

$$Ad = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad (1)$$

A. Degree Centrality (DC)

Degree Centrality (DC) is a parameter to measure number of contacts that a node have in a graph or a network, contacts are represented by edges. If the communication (edges) is directed among nodes in a graph or a network then the DC is divided into two terminologies that are indegree centrality or outdegree centrality. Indegree and outdegree centrality of a node refers to forward and backward connection towards other nodes present in a network [1].

Degree centrality (DC) measure of network graph present in Fig. 1 is like that node id E minimum value of DC that is 3 whereas node ids B and C have maximum value of DC that is 4 [1].

B. Betweenness Centrality (BC)

The betweenness centrality (BC) of a node is the total count of shortest walks passing through a node among any two nodes by seeing all sets of nodes in the graph. The count of shortest walks from node **m** to node **n** that is passing through a node g (represented as $\boldsymbol{sw_{mn}(g)}$) is the maximum of the count of shortest walks from node **m** to node $\boldsymbol{g}$ in the shortest walk tree rooted at node **m** and the count of shortest walks from node **n** to node $\boldsymbol{g}$ in the shortest walk tree rooted at node **n**. The formula for computation of BC is given by equation (2) and computation of BC for Fig. 1 is mentioned in Table I [2].

$$BC(g) = \sum_{\substack{g \neq m \\ g \neq n}} \frac{sw_{mn}(g)}{sw_{mn}} \quad (2)$$

The equation (2) can compute a betweenness centrality (BC) of a node in any network graph. BC determines the influence of a node in a graph for network analysis in a way that how important a node (vertex) is in between a communication of any other two nodes of the same graph.

From Fig. 1, it is observed clearly that node B and node C are pretty important for communication as they are lying on shortest walk between node A and node E and similarly between node D and node E. Table I demonstrate a fine picture of BC measure.

TABLE. I. BC Measure For a Network in Fig. 1

Node Id A	**Node Id B**	**Node Id C**	**Node Id D**	**Node Id E**
BC of node id A is **0.**	Set$[A,E] \rightarrow 1/2$ Set$[E,A] \rightarrow 1/2$ Set$[D,E] \rightarrow 1/2$ Set$[E,D] \rightarrow 1/2$ BC of node id B is sum of all above sets that is equal to 2.	Set$[A,E] \rightarrow 1/2$ Set$[E,A] \rightarrow 1/2$ Set$[D,E] \rightarrow 1/2$ Set$[E,D] \rightarrow 1/2$ BC of node id B is sum of all above sets that is equal to 2.	BC of node id D is **0.**	BC of node id E is **0.**

C. Closeness Centrality (CC)

Node's closeness centrality (CC) is defined as the reciprocal of the sum of the count of shortest walks from a node to all other nodes present in a network [8]. The formula for CC is given by equation (3) for network analysis. The node which carries a largest amount of CC is nearest to rest of the nodes in a network that aids a node in communicating and developing a relation with other nodes in that graph. Similarly a node with lowest CC is far from other nodes present in a network and that node may face difficulty in communicating and developing a relation with the rest of the nodes in a graph.

$$cc = \frac{1}{\sum count\ of\ shortest\ walks\ between\ a\ node\ to\ all\ other\ nodes} \quad (3)$$

For computation of CC by equation (3), we have to first evaluate shortest walk between every two nodes in a graph. For Fig. 1 shortest walk between every two nodes is represented by shortest walk distance matrix (SWDM) in equation (4).

$$SWDM = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 2 \\ 2 & 1 & 1 & 2 & 0 \end{bmatrix} \quad (4)$$

It is seen in Fig. 1 that there is no loop present therefore diagonal element of SWDM are zero shown in equation (4). Sum of shortest walk for each node is represented in equation (5).

$$\sum SWDM = \begin{bmatrix} 5 \\ 4 \\ 4 \\ 5 \\ 6 \end{bmatrix} \quad (5)$$

Finally CC for each node i in Fig. 1 is computed in equation (6) by following the definition of CC measure.

$$CC(i) = \frac{1}{\sum SWDM(i)} = \begin{bmatrix} 0.200 \\ 0.250 \\ 0.250 \\ 0.200 \\ 0.167 \end{bmatrix} \quad (6)$$

Outcomes in equation (6) shows the significance of node B and node C that have highest amount of CC which means node B and node C are nearest to rest of the nodes in a graph presented in Fig. 1.

D. Eigenvector Centrality (EVC)

Eigenvector centrality (EVC) is a measurement of amount that indicates key nodes in a graph. EVC explains the role of neighboring nodes in a way that all those nodes are essential in a network which are linked with useful nodes. Dominant eigenvector of adjacency matrix (Ad) is EVC. The EVC amount of the nodes in a network corresponds to the input for the nodes in the principal eigenvector of the network represented by $\boldsymbol{Ad}$. The **n** eigenvalues and the corresponding eigenvector is extracted from **n×n** Ad. Power method is used for evaluation of EVC from $\boldsymbol{Ad}$ of the network. For this method, we initiate from the ones vector that is $\boldsymbol{X_o} = [\mathbf{1\ 1\ 1} \ldots . \mathbf{1\ 1\ 1}]$ corresponding to the count of nodes in the network and passes through a number of iterations [2, 9, 10]. The preliminary eigenvector evaluated during the$(k+1)^{th}$ iteration is given as follows:

$$EVC = \frac{(Ad)*X_k}{\|(Ad)*X_k\|} \quad (7)$$

Where $\|(Ad) * X_k\|$ is the normalized amount of the EVC obtained in proceeding of k^{th}iteration. Power method is applied and repeated till normalized values becomes same and converges as seen in Table II.

Table II also points the importance of node id B and node id C in the considered graph network (that is Fig. 1). In 7th iteration of power method we obtained a dominant eigenvalue and corresponding eigenvector for a matrix graph Ad.

E. Katz Centrality (KC)

The Katz centrality (KC) evaluates the centrality of a vertex (node) that depends on the centrality of its adjacent nodes relatively than considering shortest walks between nodes. It is a broad view of EVC [11]. The Katz centrality (KC) for node $\boldsymbol{g}$ is computed by formula mentioned in equation (8).

$$KC(g) = \alpha \sum_{j=1}^{n} Ad_{ji} KC(g) + \beta \quad (8)$$

Where, the parameter $\boldsymbol{\beta}$ controls the centrality not to become zero. First term of equation (8) arrows to eigenvector centrality (EVC).

$$KC = \alpha Ad^T KC + \beta.\mathbb{I} \quad (9)$$

In equation (9), $\mathbb{I}$ is a unit column vector.

$$KC - \alpha Ad^T KC = \beta.\mathbb{I} \quad (10)$$

$$KC(1 - \alpha Ad^T) = \beta.\mathbb{I} \quad (11)$$

$$KC = \beta(1 - \alpha Ad^T)^{-1}.\mathbb{I} \quad (12)$$

For computation of Katz centrality (KC), always suppose value of alpha (α) less than the reciprocal of dominant eigenvalue (λ) for convergence. As dominant eigenvalue is obtained for network presented in Fig. 3 is of amount **3.3234;** therefore, in equation (12) considering $\alpha = 0.2$ and $\beta = 1$ for Katz centrality (KC) computation. Equation (13), (14), (15) and (16) demonstrated the complete evaluation details of Katz centrality for a network in Fig. 1.

$$KC = 1 \times \left(\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} - 0.2 \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (13)$$

$$KC = \left(\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.0 & 0.2 & 0.2 & 0.2 & 0.0 \\ 0.2 & 0.0 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.0 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.2 & 0.0 & 0.0 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (14)$$

$$KC = \left(\begin{bmatrix} 1.00 & -0.2 & -0.2 & -0.2 & 0.00 \\ -0.2 & 1.00 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & 1.00 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & 1.00 & 0.00 \\ 0.00 & -0.2 & -0.2 & 0.00 & 1.00 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (15)$$

TABLE. II. EVC MEASURE FOR A NETWORK IN FIG. 1

$$EV1 = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 4 \\ 3 \\ 2 \end{bmatrix}$$

$$Normalized\ value = n1 = \sqrt{3^2 + 4^2 + 4^2 + 3^2 + 2^2} = 7.3485$$

$$EVC1 = \frac{EV1}{n1} = \begin{bmatrix} 0.4082 \\ 0.5443 \\ 0.5443 \\ 0.4082 \\ 0.2722 \end{bmatrix} \rightarrow Iteration\ \#\ 1$$

$$EV2 = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.4082 \\ 0.5443 \\ 0.5443 \\ 0.4082 \\ 0.2722 \end{bmatrix} = \begin{bmatrix} 1.4969 \\ 1.6330 \\ 1.6330 \\ 1.4969 \\ 1.0887 \end{bmatrix}$$

$$n2 = \sqrt{1.4969^2 + 1.6330^2 + 1.6330^2 + 1.4969^2 + 1.0887^2} = 3.3166$$

$$EVC2 = \frac{EV2}{n2} = \begin{bmatrix} 0.4513 \\ 0.4924 \\ 0.4924 \\ 0.4513 \\ 0.3282 \end{bmatrix} \rightarrow Iteration\ \#\ 2$$

$$EV4 = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.4323 \\ 0.5188 \\ 0.5188 \\ 0.4323 \\ 0.2965 \end{bmatrix} = \begin{bmatrix} 1.4699 \\ 1.6799 \\ 1.6799 \\ 1.4699 \\ 1.0376 \end{bmatrix}$$

$$n4 = \sqrt{1.4699^2 + 1.6799^2 + 1.6799^2 + 1.4699^2 + 1.0376^2} = 3.3230$$

$$EVC4 = \frac{EV4}{n4} = \begin{bmatrix} 0.4424 \\ 0.5055 \\ 0.5055 \\ 0.4424 \\ 0.3122 \end{bmatrix} \rightarrow Iteration\ \#\ 4$$

$$EV5 = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.4424 \\ 0.5055 \\ 0.5055 \\ 0.4424 \\ 0.3122 \end{bmatrix} = \begin{bmatrix} 1.4534 \\ 1.7025 \\ 1.7025 \\ 1.4534 \\ 1.0111 \end{bmatrix}$$

$$n5 = \sqrt{1.4534^2 + 1.7025^2 + 1.7025^2 + 1.4534^2 + 1.0111^2} = 3.3233$$

$$EVC5 = \frac{EV5}{n5} = \begin{bmatrix} 0.4373 \\ 0.5123 \\ 0.5123 \\ 0.4373 \\ 0.3042 \end{bmatrix} \rightarrow Iteration\ \#\ 5$$

$$EV7 = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.4399 \\ 0.5089 \\ 0.5089 \\ 0.4399 \\ 0.3083 \end{bmatrix} = \begin{bmatrix} 1.3442 \\ 1.0972 \\ 1.0972 \\ 0.7531 \\ 0.3482 \end{bmatrix}$$

$$n7 = \sqrt{1.3442^2 + 1.0972^2 + 1.0972^2 + 0.7531^2 + 0.3482^2} = 3.3234 \rightarrow \lambda\ (Principal\ Eigenvalue)$$

$$EVC7 = \frac{EV7}{n7} = \begin{bmatrix} 0.4386 \\ 0.5106 \\ 0.5106 \\ 0.4386 \\ 0.3062 \end{bmatrix} \rightarrow Iteration\ \#\ 7$$

$$KC = \begin{bmatrix} 1.2821 & 0.4808 & 0.4808 & 0.4487 & 0.1923 \\ 0.4808 & 1.3782 & 0.5449 & 0.4808 & 0.3846 \\ 0.4808 & 0.5449 & 1.3782 & 0.4808 & 0.3846 \\ 0.4487 & 0.4808 & 0.4808 & 1.2821 & 0.1923 \\ 0.1923 & 0.3846 & 0.3846 & 0.1923 & 1.1538 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (16)$$

$$KC = \begin{bmatrix} 2.8846 \\ 3.2692 \\ 3.2692 \\ 2.8846 \\ 2.3077 \end{bmatrix} \quad (17)$$

Equation (17) also arrows a same result of influential nodes in a network as all four previous centrality measure indicates that node id B and node id C plays a vital role in communication and developing relation with other remaining nodes. Now we move towards Section IV for a concept of maximal clique size (MC) evaluation and its connection with centrality metrics.

IV. MAXIMAL CLIQUE SIZE (MC) AND ITS ASSOCIATION WITH KEY CENTRALITY METRICS

The concept behind maximal clique size (MC) of a graph for any node is that the node $\boldsymbol{g}$ is assigned a value that belongs to the presence of node $\boldsymbol{g}$ in a maximum clique size (MC) of that graph. The MC of a node is a determination of amount of modularity of a node and that can be used to recognize seed nodes about which communities can develop. The evaluation of modular node in large dataset networks have very significance for network analysis but it is seen from previous literature that preference was given to linked measure called maximal clique size (MC) over count of modular nodes. In addition to previous work, we modify previous methods for determination of maximal clique size (MC) for large dataset networks which was pretty difficult to compute, aim is to decrease complexity in computation. Modified algorithm of MC is mentioned in Table IX. Also small network example (that is Fig. 1) is considered to explain the concept of MC.

There are in total two maximal cliques present that are demonstrated in Fig. 2 with two different colors. One maximal clique is {A, B, C, D} shown by yellow lines and second one is {B, C, E} shown by green color. MC value for each node is represented in Table III.

The link between maximal clique size (MC) and all five centrality metrics is measured through renowned correlation coefficients that are Pearson's, Spearman's and Kendall's which are discussed briefly and determined in this section. The association between MC and all five centrality metrics is important in the way that if strength of association is strong and positive then we can go for centrality metrics in network analysis rather than to compute MC. Results has shown strength of positive association between them which are mentioned in Table XV.

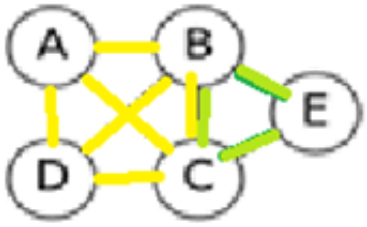

Fig. 2. Marking of Maximal Cliques on a Network Considered in Fig. 1.

TABLE. III. MAXIMAL CLIQUE SIZE FOR A NETWORK IN FIG. 1

Node ID	A	B	C	D	E
MC	4	4	4	4	3

A. Pearson's Product Moment-Based Correlation Coefficient (PCC)

The Pearson's correlation coefficient (PCC) is stated for any two data's as the ratio of covariance and the product of standard deviations. Suppose mean of maximal clique size and degree centrality are demonstrated by $\boldsymbol{MC_{avg}}$ and $\boldsymbol{DC_{avg}}$respectively for a network of **n** number of nodes. Suppose that each input corresponding to **n** number of nodes for maximal clique size and degree centrality is demonstrated by $\boldsymbol{MC_i}$ **and** $\boldsymbol{DC_i}$ respectively [6]. The evaluation for Fig. 1 through equation (18) is computed in Table IV.

$$PCC(MC,DC) = \frac{\sum_{i=1}^{n}(MC_i - MC_{avg})(DC_i - DC_{avg})}{\sqrt{\sum_{i=1}^{n}(MC_i - MC_{avg})^2 \sum_{i=1}^{n}(DC_i - DC_{avg})^2}} \quad (18)$$

$$PCC(MC,DC) = \frac{1.2}{\sqrt{0.8 \times 2.8}} = 0.8017 \quad (19)$$

From equation (19) it is clear that there exists a pretty strong positive association between maximal clique size (MC) and degree centrality (DC) that is of amount **0.8017**. The outcome indicates that for recognition of influential nodes in a network analysis one may use degree centrality metric (DC) as compared to maximal clique size (MC) that is difficult and time consuming to evaluate. Now we move on to other method of finding link between these two measures that is Spearman's correlation coefficient.

B. Spearman's Rank-Based Correlation Coefficient (SCC)

The Spearman's correlation coefficient (SCC) is stated for two data's as the determination of association by considering the ranks of the values rather than their exact values. To find the link between two variables MC and DC, we transform the two data's $\boldsymbol{MC_i}$ **and** $\boldsymbol{DC_i}$into rank data that is $\boldsymbol{m_i}$ **and** $\boldsymbol{dc_i}$ respectively. SSC can be evaluate through formula presented in equation (20) [6].

$$SCC(MC,DC) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (20)$$

Where, $d_{i=}m_i - dc_i$ is the difference of ranks between two variables. The evaluation of SCC between MC and DC for network in Fig. 1 is done through equation (20) presented in Table V.

$$SCC(MC,DC) = 1 - \frac{6*4}{5(5^2-1)} = 1 - \frac{24}{120} = 0.8 \quad (21)$$

Equation (21) demonstrates the output of SCC computation that also shows a pretty strong positive association between MC and DC that is of amount **0.8**. In few words, one may prefer DC over MC evaluation for a network analysis.

C. Kendall's Concordance-based Correlation Coefficient (KCC)

Kendall's correlation coefficient (KCC) is stated for two data's as the count of similarity in the arrangement of the values for the variables (data's) acquired by the nodes in the network. The set of nodes $\boldsymbol{n_i}$ and $\boldsymbol{n_j}$ are said to be concordant sets (conc.sets) if either of these 2 nodes rigorously has a greater value or a smaller value for two variables MC and DC. Similarly, the set of two $\boldsymbol{n_i}$ and $\boldsymbol{n_j}$ is said to be discordant sets (disc.sets) if a node has a greater value or smaller value for at least one out of two variables. The set of nodes $\boldsymbol{n_i}$ and $\boldsymbol{n_j}$ is neither said to be concordant set nor to be discordant set if either of the set have equal values for MC and DC (shown in Table VI) [2]. The KCC is evaluated by formula presented in equation (22) and evaluation is presented in Table VII.

$$KCC(MC,DC) = \frac{no.of\ conc.sets - no.of\ disc.sets}{\frac{1}{2}n(n-1)} \quad (22)$$

$$no.of\ conc.sets = 4 \quad (23)$$

$$no.of\ disc.sets = 0 \quad (24)$$

$$Total\ no.of\ sets = \frac{5(5-1)}{2} = 10 \quad (25)$$

$$KCC(MC,DC) = \frac{4-0}{10} = 0.4 \quad (26)$$

Count of concordant sets (conc.sets) is given by equation (23), count of discordant sets (disc.sets) is given by equation (24) and equation (25) represents the total number of sets. The outcome of the Kendall's correlation coefficient (KCC) is given by equation (26) that arrows positive link between maximal clique size (MC) and degree centrality (DC) that is of amount **0.4**. This amount also shows the preference of degree centrality (DC) over maximal clique size (MC).

TABLE. IV. FINDING ASSOCIATION BETWEEN MC AND DC THROUGH PCC

n_i	MC	DC	$MC - MC_{avg}$	$DC - DC_{avg}$	$(MC - MC_{avg}) * (DC - DC_{avg})$	$(MC - MC_{avg})^2$	$(DC - DC_{avg})^2$
A	4	3	0.2	-0.2	-0.04	0.04	0.04
B	4	4	0.2	0.8	0.16	0.04	0.64
C	4	4	0.2	0.8	0.16	0.04	0.64
D	4	3	0.2	-0.2	-0.04	0.04	0.04
E	3	2	-0.8	-1.2	0.96	0.64	1.44
Avg	3.8	3.2		Sum	1.2	0.8	2.8

TABLE. V. FINDING ASSOCIATION BETWEEN MC AND DC THROUGH SCC

n_i	MC	Trial Rank: MC	Final Rank: m_i	DC	Trial Rank: DC	Final Rank: dc_i	$d_i = m_i - dc_i$	d_i^2
A	4	2	3.5	3	2	2.5	1	1
B	4	3	3.5	4	4	4.5	-1	1
C	4	4	3.5	4	5	4.5	-1	1
D	4	5	3.5	3	3	2.5	1	1
E	3	1	1	2	1	1	0	0
							Sum	4

TABLE. VI. VALUES OF MC AND DC FOR EACH NODE PRESENT IN FIG. 1

Node Id	A	B	C	D	E
MC	4	4	4	4	3
DC	3	4	4	3	2

TABLE. VII. FINDING ASSOCIATION BETWEEN MC AND DC THROUGH KENDALL'S CORRELATION COEFFICIENT

Node Sets (n_i, n_j)	(A, B)	(A, C)	(A, D)	(A, E)	(B, C)	(B, D)	(B, E)	(C, D)	(C, E)	(D, E)
MC_i, DC_i	(4,3)	(4,3)	(4,3)	(4,3)	(4,4)	(4,4)	(4,4)	(4,4)	(4,4)	(4,3)
MC_j, DC_j	(4,4)	(4,4)	(4,3)	(3,2)	(4,4)	(4,3)	(3,2)	(4,3)	(3,2)	(3,2)
Class of Sets	**N/A**	**N/A**	**N/A**	**Conc.set**	**N/A**	**N/A**	**Conc.set**	**N/A**	**Conc.set**	**Conc.set**

V. AMAZON PRODUCT NETWORK DATA

The dataset of amazon product co-purchasing is of June 2003 containing 403394 nodes and 3387388 edges is evaluated. This data informs the consumer's pattern of buying which kind of products are usually bought in combination [12]. This 403394 nodes (products) data is converted into adjacency matrix $'Ad'$ and then graph is formed for further network analysis as demonstrated in Fig. 3 and Fig. 4 where nodes (products) are in blue color and edges are represented by green color.

Amazon co-purchasing data of 1001 products are extracted for computation to identify influential products through centrality metrics. All these five metrics are strongly linked with each other. It is captured in Table VIII clearly like node id 4 is highest in all 5 metrics.

Table VIII shows all five centrality metrics of nodes present in large product data set. Evaluation is done on extracted data of 1001 nodes from amazon website. The target is to find influential nodes through centrality metrics and maximal clique size (MC). Node id **5** is found as vital node (product) in network analysis through these 5 measures, as node id **5** have highest values in all most all 5 measures that indicates its importance in terms of profit in marketing as these outcomes are evaluated through amazon product network data. Secondly node id **29** has a second highest measure which is also arrows its importance in marketing of amazon products. Now we move towards another significant measure in network analysis that is maximum clique size (MC).

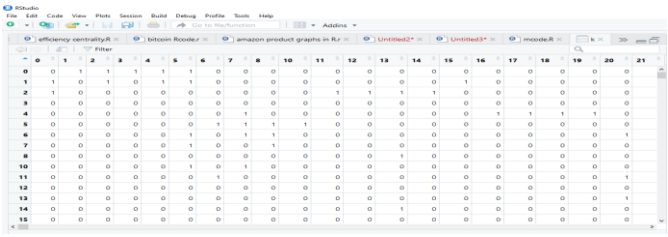

Fig. 3. Adjacency Matrix Formation for Amazon Product Dataset.

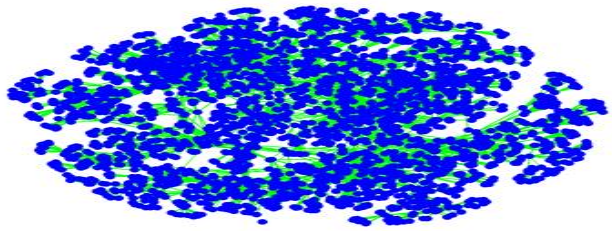

Fig. 4. Plot of Amazon Product Network from Adjacency Matrix.

TABLE. VIII. DETERMINATION OF INFLUENTIAL NODES IN AMAZON DATA THROUGH CENTRALITY METRICS

Products (Nodes Id)	**Degree (DC)**	**Betweenness (BC)**	**Closeness (CC)**	**Eigenvector (EVC)**	**Katz Centrality (KC)**
0	10	452.448	5.1939e-05	0.08630	1.1244
1	10	150.285	4.7123e-05	0.05577	1.1166
2	10	122.528	4.6759e-05	0.07234	1.1190
3	10	1150.530	5.2952e-05	0.08552	1.1272
4	21	9008.548	5.3205e-05	0.17034	1.2619
5	74	311765.340	6.0335e-05	0.5	1.8879
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
29	31	858717.100	5.0241e-05	0.00122	1.3649
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
499	11	563.207	4.7703e-05	0.00068	1.1250
500	10	3816.174	4.7700e-05	0.00067	1.1140
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
999	10	4506.667	3.8430e-05	2.4901e-07	1.0942
1000	10	0	1.2411e-07	1.0142e-17	1.0000

TABLE. IX. PSEUDO-ALGORITHM FOR COMPUTATION OF MAXIMAL CLIQUE SIZE IN LARGE DATASET

```
> Import. Dataset (data)
>A=Adjacency. Matrix (data) #formation of Adjacency matrix from a network data
> Library (ggnet); library (network); library (sna); library (ggplot2)
>graph=network (A)
>g=ggnet2 (graph)
>g=plot (A)            # graph formation
>largest. Cliques (g) # display maximal cliques and nodes that are present in largest clique size
>n=clique. Number (g)                    # node count in maximal clique
>m=maximal. Cliques (g)
>r=zeros (total nodes, 1)                # null matrix of total nodes by 1
>    if    (n>=1) then
          for    i : total nodes do
                  v= unlist (m[n])
                  display (v)   }
                Nv =length (v)
              for   (i in 1: Nv) do
                    j ←v[i]
                if    (r[j] <= Nv) then
                      r[j]   ← length (v)
                      display (r)
                            else   {
                                  display (r)    }  }
                    n=n-1     }  }       # return a vector r of total nodes by 1 dimension containing a maximal clique size for each node
```

The maximal clique size (MC) is evaluated for amazon product network of 1001 nodes by improved algorithm as mentioned in Table IX and outcomes for each node in a network are represented in Table X. This improved algorithm decreases the complexities in computation and demonstrates the appropriate result of MC for each and every node in the data.

The maximal clique size (MC) is evaluated for amazon product network of 1001 nodes that is mentioned in Table X. The association between MC and DC by three renowned measures are discussed and evaluated that are represented in Tables XI, XII, XIII and XIV by correlation coefficients Pearson's, Spearman's and Kendall's, respectively.

$$PCC(MC,DC) = \frac{3313.17}{\sqrt{6228.476 \times 26079.77}} = 0.2599566 \qquad (27)$$

The outcome of link between maximal clique size (MC) and degree centrality (DC) through PCC for large dataset of amazon network is equals to **0.2599566** which shows positive association between these two variables as discussed in previous section for small network example.

$$SCC(MC,DC) = 1 - \frac{6*115461200}{1001\times(1001^2-1)} = 0.3093 \qquad (28)$$

Similarly, the result of association between maximal clique size (MC) and degree centrality (DC) of amazon product network through SCC measure also indicates a positive link that is of amount **0.3039**. It is also pretty clear from Table XIII that node id 5 which carries all five highest centrality metrics specially highest DC i.e. 74 also contains third larger maximal clique size (MC) in amazon network which shows strength of link between them.

TABLE. X. MAXIMAL CLIQUE SIZE COMPUTATION FOR AMAZON PRODUCT DATA

Node Id	0	1	2	3	4	5	.	499	500	.	999	1000
MC	5	5	5	5	6	8	.	9	9	.	2	2

TABLE. XI. EVALUATING RELATION AMONG MC AND DC FOR AMAZON DATASET THROUGH PCC

n_i	**MC**	**DC**	$MC - MC_{avg}$	$DC - DC_{avg}$	$(MC - MC_{avg}) * (DC - DC_{avg})$	$(MC - MC_{avg})^2$	$(DC - DC_{avg})^2$
0	5	10	-0.718	-2.685	1.928	0.515	7.209
1	5	10	-0.718	-2.685	1.928	0.515	7.209
2	5	10	-0.718	-2.685	1.928	0.515	7.209
3	5	10	-0.718	-2.685	1.928	0.515	7.209
4	6	21	0.282	8.315	2.345	0.079	69.139
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
499	9	11	3.282	-1.685	-2.24817	10.771	2.839
500	9	10	3.282	-2.685	-5.53017	10.771	7.209
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
999	2	10	-3.718	-2.685	9.983	13.823	7.209
1000	2	10	-3.718	-2.685	9.983	13.823	7.209
Avg	5.72	12.7		**Sum**	3313.17	6228.476	26079.77

TABLE. XII. EVALUATING RELATION AMONG MC AND DC FOR AMAZON DATASET THROUGH SCC

n_i	MC	Final Rank: m_i	DC	Final Rank: dc_i	$d_i = m_i - dc_i$	d_i^2
0	5	338	10	203	135	18225
1	5	338	10	203	135	18225
2	5	338	10	203	135	18225
3	5	338	10	203	135	18225
4	6	473	21	942	-469	219961
.	.	.	.	.	.	.
.	.	.	.	.	.	.
499	9	924	11	478	446	86142.25
500	9	924	10	203	721	198916.00
.	.	.	.	.	.	.
.	.	.	.	.	.	.
999	2	122	10	203	-81	6561
1000	2	122	10	203	-81	6561
					Sum	115461200

TABLE. XIII. COUNT OF MC AND DC MEASURE FOR EACH NODE IN AMAZON PRODUCT NETWORK

Node Id	0	1	2	3	4	5	.	499	500	.	999	1000
MC	5	5	5	5	6	8	.	9	9	.	2	2
DC	10	10	10	10	21	74	.	11	10	.	10	10

TABLE. XIV. EVALUATING RELATION BETWEEN MC AND DC FOR AMAZON PRODUCT NETWORK

Node Sets (n_i, n_j)	(0,1)	(0,2)	(0,3)	(0,4)	(0,5)	.	(499,500)	.	(999,1000)
MC_i, DC_i	(5,10)	(5,10)	(5,10)	(5,10)	(5,10)	.	(9,11)	.	(2,10)
MC_j, DC_j	(5,10)	(5,10)	(5,10)	(6,21)	(8,74)	.	(9,10)	.	(2,10)
Class of Sets	**N/A**	**N/A**	**N/A**	**Conc.set**	**Conc.set**	**.**	**N/A**	**.**	**N/A**

TABLE. XV. RELATIONSHIP BETWEEN KEY CENTRALITY METRICS AND MC USING PCC, SCC AND KCC

	DC and MC			**BC and MC**			**CC and MC**			**EVC and MC**			**KC and MC**		
	PCC	**SCC**	**KCC**	**PCC**	**SCC**	**KCC**	**PCC**	**SCC**	**KCC**	**PCC**	**SCC**	**KCC**	**PCC**	**SCC**	**KCC**
5 nodes network in Fig. 1	0.801	0.8	0.4	0.408	0.408	0.408	0.721	0.745	0.707	0.904	0.707	0.632	0.873	0.725	0.667
Amazon Product network	0.259	0.309	0.177	0.057	-0.18	-0.13	-0.07	-0.12	-0.082	0.122	0.085	0.065	0.299	0.415	0.301

$$\text{no. of conc. sets} = 214698 \quad (29)$$

$$\text{no. of disc. sets} = 125986 \quad (30)$$

$$Total\ no.of\ sets = \frac{1001(1001-1)}{2} = 500500 \quad (31)$$

$$KCC(MC, DC) = \frac{214698-125986}{500500} = 0.17724 \quad (32)$$

Equation (32) represents the outcome of KCC for maximal clique size (MC) and degree centrality (DC) that again conveys a positive link between them for amazon product data of 1001 nodes. It is observed from previous literature and present study that amount of KCC measure is small as compared to PCC and SCC measures but delivers a same picture of concept that they have positive connection between maximal clique size (MC) and degree centrality (DC) in network analysis.

The bond between fundamental centrality metrics (like DC, BC, CC, EVC and KC) and maximal clique size (MC) is demonstrated in Table XV through PCC, SCC and KCC measures. It is seen that DC, EVC and KC have strong positive relation with MC in network analysis. Katz centrality (KC)

metric also shows pretty strong positive association with maximal clique size (MC) as it is observed from present study. For determination of significant nodes in large datasets one may prefer degree centrality (DC), eigenvector centrality (EVC) and Katz centrality (KC) measures over maximal clique size (MC) computation which is difficult to measure. Secondly, betweenness and closeness centrality metrics shows least association with MC.

VI. Conclusion

The complete work of this paper addressed an amount of modularity and use of improved method of maximal clique size (MC) in large network datasets. Although it is hard to measure MC for big datasets and finding its connection with centrality metrics, the improved algorithm has been introduced to decrease complexity for large networks and results have been computed for Amazon large product network data and also for a small network example. Strong connection of maximal clique size (MC) with degree centrality (DC), eigenvector centrality (EVC) and Katz centrality (KC) was seen by Pearson's correlation (PCC), Spearman's correlation coefficient (SCC) and Kendall's correlation coefficient (KCC). The strength of association between them indicates that these three centrality measures can be favored over maximal clique size (MC) computation for network analysis. It is also seen that Pearson's and Spearman's correlation coefficients measure outcomes are almost same as compared to Kendall's correlation coefficient measure which shows small values in their comparison but picture of outcome is same that is quality of association between variables.

References

[1] Kosorukoff, A. and D.L. Passmore, Social Network Analysis: Theory and Applications. 2011: Passmore, D. L.

[2] NatarajanMeghanathan. A comprehensive analysis of the correlation betweenmaximal clique size and centrality metricsfor complex network graphs. in 16th IEEE International Conference on Emerging eLearning Technologies and Applications, Proceedings. 2018.

[3] Zhang, Z., et al., Modeling Epidemics Spreading on Social Contact Networks. IEEE transactions on emerging topics in computing, 2015. 3(3): p. 410-419.

[4] Lawyer, G., Understanding the influence of all nodes in a network. Scientific Reports, 2015. 5: p. 8665.

[5] Yin, Q., et al., The impact of contact patterns on epidemic dynamics. PLOS ONE, 2017. 12(3): p. e0173411.

[6] Meghanathan, N., A computationally lightweight and localized centrality metric in lieu of betweenness centrality for complex network analysis. Vietnam Journal of Computer Science, 2017. 4(1): p. 23-38.

[7] Graph. [cited 2019 17 January 2019]; Available from:https://frnsys.com/ai_notes/foundations/graphs.html.

[8] Borgatti, S. Centrality. 2005 [cited 2015 10 October 2015]; Available from: http://www.analytictech.com/essex/Lectures/centrality.pdf.

[9] Corporation, L. Social Network Analysis (SNA). 2019 [cited 2019; Available from: https://www.slideshare.net/gcheliotis/social-network-analysis-3273045.

[10] Bihari, A. and M. Kumar Pandia, Eigenvector centrality and its application in research professionals' relationship network. 2015.

[11] Developers, N. Katz Centrality. 2015 26 October 2015 [cited 2018 15 December 2018]; Available from:https://networkx.github.io/documentation/networkx1.10/reference/generated/networkx.algorithms.centrality.katz_centrality.html.

[12] Leskovec, J. Stanford Large Network Dataset Collection. [cited 2016 1 January 2016]; Available from: http://snap.stanford.edu/data/.