Determination and Identification of Important and Influential Nodes Involved in the Pathology of Escherichia Coli Using Improved TOPSIS Method

zohreh minaei (zohreh.minaei@gmail.com)

Research article

Keywords: E. Coli, TOPSIS, Network Analysis, Centrality

Posted Date: December 10th, 2019

DOI: https://doi.org/10.21203/rs.2.18279/v1

Determination and Identification of Important and Influential Nodes Involved in the Pathology of Escherichia Coli Using Improved TOPSIS Method

Zohreh Minaei[1,*]

[1]Department of Engineering, Islamic Azad University, Bushehr, Iran.

* zohreh.minaei@gmail.com

Abstract

Various disciplines are trying to solve one of the most noteworthy queries and broadly used concepts in biology, essentiality. Centrality is a primary index and a promising method for identifying essential nodes, particularly in biological networks. Thus, important nodes of the network can be identified by analyzing some of the centrality extracted from the network. In this paper, we aim to identify the important proteins in the Escherichia Coli (E.Coli) network based on extraction of centralities. During these operations, centralities such as degree of centrality, betweeness, laplacian and closeness, are considered as node's important indicators. Finally the important nodes will be determined based on the centrality and Technique for order performance by similarity (TOPSIS) method. After performing the weighted TOPSIS simulation and obtaining the output result, it was found that the proposed hybrid system is able to place 74 and 99 important nodes between the top 100 and 150 nodes, respectively. Finally, the results of this study are compared with other similar studies.

Keywords: E.Coli; TOPSIS; Network Analysis; Centrality

1 Background

There are many intricate networks in nature, each of which has important and influential nodes in terms of their size, Therefore, there are many ways to determine and identify the important and effective nodes of these networks. These methods are often used to investigate the chemical network (The interaction network between proteins) or the network of human interactions (Social Networks) [10]. In general, as we said at the beginning, the main challenge in modeling complex networks is the challenge of how these networks grow. One solution to this problem is to identify the nodes in these networks. Because knowing the important and powerful nodes in the networks can increase the possibility of communication between the nodes [1].

In general, all human beings are also part of the existing networks and networks of various social relationships. The science of identifying the graph and the important relationships between its nodes is due to the efforts of a

scientist named Obler and And based on Paul's problem. As a result, the science of identifying graphs and their relationships is known as graph theory. This theory focuses on issues such as finding the shortest path between two nodes as well as finding the highest flow rate between the source and destination nodes [10]. The result of scientists efforts in graph theory is the production of a system called the network analysis system. This system is one of the most important network related topics introduced in 1920. Since the network analysis system is an interconnection system across all networks, scientists from various sciences, including sociology, economics, and computer science, have worked on are very difficult to obtain. As a result, the researchers were able to describe how networks use different topologies [11].

In a study of identifying important nodes, the researchers were able to refine the method of ranking nodes that can be expanded. In general, this new method, known as weighting technique to increase efficiency, has been suggested by the weighted TOPSIS method. Finally, the researchers used the SIR model to simulate four types of real networks to evaluate the method [4].

It has also [6] been analyzed and evaluated in the methods of network evaluation based on the extracted centralities. According to research, biological networks can be evaluated in 4 ways: 1) Use of classical centralities. 2) Combining classical centralities. 3) Providing new centralities, and finally 4) Merging data using existing centralities.

In [5], a system has been proposed that can extract the centralities of biological networks in both directional and non-directional states. Based on the extracted centralities for E.Coli, the researchers concluded that the laplacian centrality had the best output, being able to identify 74 of the 605 important nodes among the top 100 nodes (In terms of the value calculated by the laplacian centrality per node). Centralities such as degree, closeness, infocent, decay are also next in line, capable of identifying 71 nodes. They also found that centralities such as local cluster, coefficient, and DMNC were identified at the bottom of the list of E.Coli disease network centralities, with fewer than 10 nodes identified.

2 The Basic theory of E.Coli, TOPSIS, and centrality measures for node influence

2.1 The basic theory of E.Coli

Escherichia Coli (E.Coli) is caused by a bacterium that normally lives in the gut of healthy individuals and animals. Most forms of the disease are harmful and cause mild diarrhea. But some of its stronger bacteria such as E.Coli O157:H7 can cause severe abdominal contractions, bloody diarrhea and nausea. Healthy adults generally

recover after a week, but children and the elderly are more likely to develop a dangerous condition called uramic syndrome.

2.2 The basic theory of TOPSIS

Optimization exists every where in our real life. Among so many optimization methods, Technique for order performance by similarity (TOPSIS) was developed by Huang and Yon in 1981 which is a simple but effective ranking method. TOPSIS as one of the classical compensatory methods in multi-criteria decision making to solve similarity-based prioritization problems. The TOPSIS method is a highly technical and robust decision-making method for prioritizing options by simulating the ideal solution. The choice of this method should have the shortest distance from the positive ideal and on the other hand the longest distance from the negative ideal.

In general, the philosophy of the TOPSIS method is that in the method described above, two hypothetical options are defined. One of these options, which is a set of the best values observed in the decision matrix, is called the positive ideal. Meanwhile, another hypothetical option is defined, which includes the worst case scenario. This option is called negative ideal [13].

Criteria can be positive or negative in nature, and their unit of measurement may also be different.This method performs the identification of important and influential nodes through seven stages of operation (Extrecting centralities, Preparation of decision matrix, Decision normalization, Calculate the normalized matrix, Determine positive and negative ideals, Determine the distance, Calculate scores). Among the benefits of the TOPSIS method are the following:

- In the TOPSIS method, the selected option must have the shortest distance from the ideal solution and the longest distance from the most inefficient answer.
- In this method, the matrix n*m with m option and n index is evaluated.
- One of the important advantages of this method is that both objective and subjective indicators and criteria can be used simultaneously.
- It incorporates both quantitative and qualitative criteria into the topic of location.
- Its output can specify the order of priority of the options and quantify this priority.
- Takes into account the contradiction between indicators.
- The method is simple and fast.

2.3 Centrality measures

To this end, various centralities such as laplacian centrality, degree centrality, closeness centrality, stresscent centrality, and betweenness centrality have been proposed to obtain information on the nodes forming the networks. Therefore, in this study we intend to use some of these centralities (According to the reviews, the best centralities to combine and use in the TOPSIS method are laplacian , betweenness, degree, and stresscent centralities [5]).

Considering a simple network G = (V, E) with n = |V| nodes and m = |E| links. We use an adjacent matrix $A = \{a_{uv}\} \in R^{n,n}$ to describe G, where $a_{uv} = 1$ if node u is connected with node v and $a_{uv} = 0$ if node u is not connected with node v.We use $\Gamma_h(v)$ to represent the set of neighbors within h-hops from node v [4].

Definition 3.1 The Degree Centrality (DC) , $C_D(v)$, of node v can be calculated as:

$$C_D(v) = \sum_{u=1}^{n} a_{uv} = |\Gamma_1(v)| \quad (1)$$

Definition 3.2 The Betweenness Centrality (BC), $C_B(v)$, of node v is define as the fraction of shortest paths between all node pairs that pass through node v which is given by:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}V}{\sigma_{st}} \quad (2)$$

where σ_{st} represents the number of all possible shortest paths between s and t, and σ_{st} (v) represents the number of shortest paths between s and t which pass through v.

Definition 3.3 The Closeness Centrality (CC), $C_C(v)$, of node v is defined as the reciprocal of the sum of the shortest distances to all other nodes of V.

$$C_C(v) = \frac{1}{\sum_{u \in V \setminus v} d_{uv}} \quad (3)$$

where d_{uv} represents the shortest distance between u and v.

3 Proposed Method

As stated from the beginning, the main purpose of this study is to determine the important and effective nodes in the network of interactions between E.Coli protein.Therefore, in this study, based on the hybridization of the centroid compositions presented in [6], we propose a central compositional method to identify important nodes. The process of identifying the major nodes in this method is based on three steps: generating a network of

interactions between E.Coli proteins, calculating the value of Centralities, and finally analyzing the Centralities values by the TOPSIS method. So first we explain how to generate the network, then how to extract the Centralities and finally how the TOPSIS method works.

3.1 Construct a Network

The network of interactions between E.Coli proteins, contains 3138 proteins have been identified, of which only 605 proteins of this network are more valuable.

Note: It should be noted that this network is already generated and it doesn't need to be reproduced.

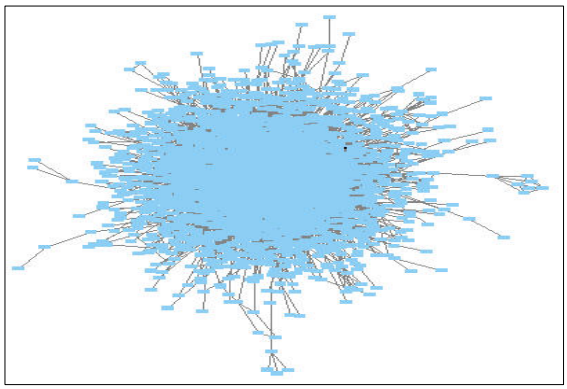

Figure. 1. E.Coli network

3-2. The algorithm of the proposed method

By definition, the TOPSIS method performs its seven steps based on a number of centralities and a weighting system. We also said that the existing centralities are all obtained from the matrix (Proximity Matrix) produced by the E.Coli network. Hence, we proceed to describe the seven steps of the TOPSIS method.

Step1: Extracting centralities

To this end, various centralities such as laplacian centrality, degree centrality, interstitial centrality, stresscent centrality, and eigenvector centrality have been proposed to obtain information on the nodes forming the networks. Therefore, in this study we intend to use some of these centralities (According to the reviews, the best Centralities to combine and use in the TOPSIS method are laplacian , betweenness, degree, and stresscent [5]).

Step2: Generating decision matrix

At the production stage of the decision matrix, also known as the zero matrix, a new matrix is constructed by aggregating the values corresponding to the centralities obtained from the adjacent matrix of the E.Coli network. In this matrix the number of columns is equal to the number of extracted centralities (In this study 4 centrality were used) and the number of rows is equal to the number of nodes in the main network (3138 nodes).

$$X = \begin{bmatrix} A_1(Q_1) & \dots\dots\dots\dots & A_1(Q_M) \\ . & . & . \\ . & . & . \\ . & . & . \\ A_N(Q_1) & \dots\dots\dots\dots & A_N(Q_M) \end{bmatrix} \quad (4)$$

$$\begin{cases} i = 1, 2, 3, \dots, N \\ j = 1, 2, 3, \dots, M \end{cases} --> A_i(Q_j)$$

Step3: Normalized the decision matrix

After generating the decision matrix in the previous step, and since the values in this matrix may not have the same representation and format, we need to normalize all inputs of the decision matrix at this stage (Convert all the numbers in the matrix into one identical display format. In this method, all numbers between zero and one are shown). Normalization is actually done in two types based on the input weight mode (Positive or negative weight).

$$r_{ij} = A_i(Q_j)^{min} / A_i(Q_j) \qquad w < 0 \quad (5)$$

$$r_{ij} = A_i(Q_j) / A_i(Q_j)^{max} \qquad w > 0 \quad (6)$$

where r_{ij} represents the points earned by option i in criterion j, $A_i(Q_j)$ represents the Each of the values contained in each decision matrix column, $A_i(Q_j)$min represents the smallest value in each column and $A_i(Q_j)$ max represents the largest value in each column.

Note: The following matrix shows the output of these two formulas

$$NM = \begin{bmatrix} r_{11} & \dots\dots\dots\dots & r_{1M} \\ . & . & . \\ . & . & . \\ . & . & . \\ r_{N1} & \dots\dots\dots\dots & r_{NM} \end{bmatrix}$$

$$\begin{cases} i = 1, 2, 3, \dots, N \\ j = 1, 2, 3, \dots, M \end{cases} --> r_{ij}$$

Step4: Compute the weighted normalized decision matrix

At this point we have to produce weight for the number of input matrix columns (Number of used centrality). But each criterion must have a weight (The positive or negative weight is determined by the type of values). It should be noted that the sum of the weights assigned to the criteria should not exceed 1 & 0 [14].

$$W_j \in [1,0] \qquad \sum W_j = 1 \quad (7)$$

So far different weighting methods have been presented in TOPSIS method. In this research we have used entropy

mapping based on weighting method.

$$W_j = \frac{d_j}{\sum d_i} \qquad i, j = 1,2, \dots, n \tag{8}$$

A) $d_j = (1 - e_j)$

B) $e_j = -k \sum_{i=1}^{m} P_{ij} \left(lnP_{ij}\right) \quad j = 1,2,3, \dots, n$

$$k = \frac{1}{\ln(M)}$$

C) $P_{ij} = \frac{x_{ij}}{\sum x_{ij}}$

Where P_{ij} represents the normalization state of the values in each row and column (The value of centrality per node), X_{ij} represents the amount of centrality per node in each row and column, e_j represents the entropy mapping represents the value of each node. This mapping is calculated on the value of k and P_{ij}. K This value is used to generate a small coefficient and has a negative sign when used.

The reason for the negative sign is to convert the negative value of the sum of the entropy mapping to positive. So in the next step, the amount of weight produced will be less than one. M represents the total number of nodes in the decision matrix and d_j represents the Initial weight per centrality. After determining the weights (The general formula for weight production is given below), we have to create a matrix called unbalanced matrix (V).

This matrix is derived from the multiplication of created weight for each criterion in the normalized matrix.

$$V = NM * W_j \tag{9}$$

$$V = \begin{bmatrix} W_1 r_{11} & \dots\dots\dots\dots & W_j r_{1M} \\ . & . & . \\ . & . & . \\ . & . & . \\ W_1 r_{N1} & \dots\dots\dots\dots & W_j r_{NM} \end{bmatrix}$$

Where V represents the unbalanced matrix, NM represents the normalized matrix and W_j represents the Weight produced per criterion.

Step5: Determine the ideal of positive and negative

In order to form a positive ideal, the best value must be chosen in each of the unbalanced matrix columns .That is, if the index corresponding to the column is negative (such as cost factor), then the lowest value in that column should be selected. In order to determine the negative ideal, we act the opposite of the positive ideal.

$$A^{+} = \{v_{i1}^{max}, v_{i2}^{max}, \dots, v_{im}^{max}\} \quad (10)$$

v_{im}^{max} represents the maximum amount of centrality per node in each column.

$$A^{-} = \{v_{i1}^{min}, V_{i2}^{min}, \dots, v_{im}^{min}\} \quad (11)$$

v_{im}^{min} represents the minimum amount of centrality per node in each column.

Step6: Determine the distance between each node and the ideal

In this section, we calculate the distance of each node for positive and negative ideals. The purpose of calculating this distance is to determine the importance of each node. In general, this distance is calculated in positive and negative states.

$$D_i^{+} = \left[\sum_{j=1}^{m}\left(Y_{ij} - v_j^{max}\right)^2\right]^{1/2} \quad (12)$$

Where D_i^{+} represents the distance in positive, Y_{ij} represents the values in each column of the unbalanced matrix and V_j^{max} represents the maximum value per column.

$$D_i^{-} = \left[\sum_{j=1}^{m}\left(Y_{ij} - v_j^{min}\right)^2\right]^{1/2} \quad (13)$$

Where D_i^{-} represents the distance in negative, Y_{ij} represents the values in each column of the unbalanced matrix and V_j^{min} represents the minimum value per column.

Step7: Ranking nodes

In this section, after determining the positive and negative distance of the nodes in each column, we determine the size and rank of each node. Generally, this step determines the important nodes in the network.

$$Z_i = \frac{D_i^{-}}{D_i^{-} + D_i^{+}} \qquad 0 \leq Z_i \leq 1 \quad (14)$$

Where Z_i represents the size of each node, D_i^{-} represents the negative distance of each node and D_i^{+} represents the positive distance of each node.

4 Results

In order to perform the proposed model process, we first need to define the simulation parameters. Therefore, Table 1 provides information about the simulation of this study. As mentioned in the previous sections, we need to extract some basic data from these networks in order to investigate complex networks, such as social networks,

biological networks, and so on. This data, which is extracted from networks, is called centrality. Until now, there have been numerous centralities with different degrees of importance for analyzing networks and graphs.

Table 1. simulation parameters

Items	Data set
Type of network	E.Coli biological network
Number of node's	3138
Number of essential nod's	605
Method/Algorithm	TOPSIS
Number of centrality	4
Weighting method	Entroppy maping

According to [5], the best centralities for investigating the biological network of E.Coli are: laplacian , betweenness, degree and stresscent. Simulation software such as Paython, R, MATLAB can be used to extract these centralities from the target network. But on the [5], a web site called centiserver has been introduced that is able to extract all the centralities provided for any type of network, both directional and non-directional.

After extracting the desired centralities, sorting the values of each node in a large state to a small one and finally examining it with a list of important nodes, it was found that among the extracted centralities, the laplacian centrality contracted 74 important nodes in the top 100 nodes, It has the best performance among E.Coli biological network centralities. It also managed to place 95 important nodes among the top 150 nodes, Which came in second.

As we have stated from the beginning, the purpose of this study is to present a hybrid system based on applying the best extracted centrals from the E.Coli biological network. In this regard, we have introduced a hybrid system called Weighted TOPSIS . Initially, we hypothesized that the system would be able to place the number of more important nodes between its top 100 and top 150 nodes using its seven-step process as well as the ranking process of top nodes. This method, by utilizing an optimal weighting system (entropy mapping weighting system), enables this hypothesis to be realized. Table 2 shows the weights obtained for each centrality.

Table 2. weights obtained for each centrality

Centrality	Weight
Laplacian	0.2438
Degree	0.1451
Betweenness	0.3032
Stresscent	0.3034

After performing the weighted TOPSIS simulation and obtaining the output result, it was found that the proposed hybrid system is able to place 74 and 99 important nodes between the top 100 and 150 nodes, respectively. Table 3 shows the result's of the each centrality [5] and proposed method.

Figuer.2 illustrates these results.

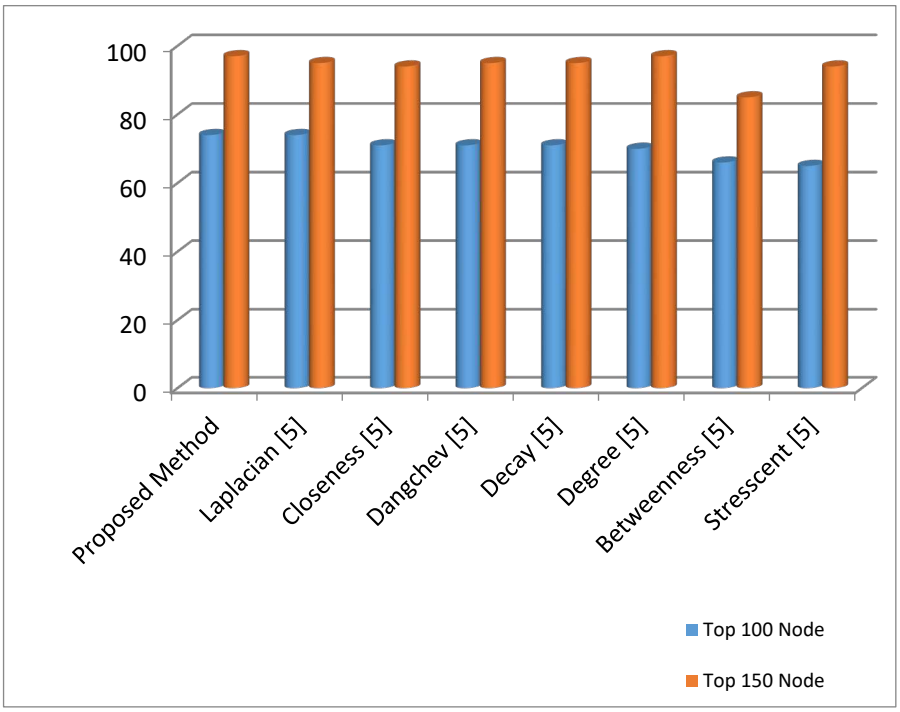

Figure. 2. Illustrates these results

Table.3 illustrates these results.

Table 3. results of the centralities and proposed method

Centrality	Top 150 Node	Top 100 Node
Proposed Method	99	74
Laplacian Centrality	95	74
Degree Centrality	97	70
Closeness Centrality	94	71
Dangchev Centrality	95	71
Decay Centrality	95	71
Betweenness Centrality	85	66
Stresscent Centrality	94	65

5 Conclusions

Nowadays various methods have been proposed for analyzing complex networks such as social networks and biological networks. In the meantime, the centralities based approach is more applicable. So in our research we also used a hybrid approach based on these same centralities. In this regard, we first extracted the different centralities of our network and selected the best ones (Table 3). As a result, by creating a new input network (Based on superior centralities), we were able to implement the seven steps of our proposed method and examine the results. Finally, it was found that the proposed method is able to accommodate more important nodes between its top 100 and top 150 nodes (Fig. 2). We are currently studying Support Vector Machine (SVM) network to improve the output. SVM or back-vector machine-based model is able to identify important nodes more accurately because of the benefits of artificial intelligence.

Abbreviations

E.Coli: Escherichia Coli

TOPSIS: Technique for order performance by similarity

SVM: Support Vector Machine

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

- The datasets generated and/or analysed during the current study are available in the [centiserver] repository, [www.centiserver.org]

Competing interests

I have no competing interests.

Funding

Not applicable.

Authors' contributions

Not applicable.

Acknowledgements

Not applicable.

References

1. Bertrand, A. : Comments on “Distributed identification of the most critical node for average consensus”. IEEE, **1**, 1-2 (2016)

2. Deng, Y. : An improved genetic algorithm with initial population strategy for symmetric TSP. ELSEVIER., **4**, 68-71 (2015)

3. Gao, Z., Shi, Y., Chen, S. : Identifying Influential Nodes for Efficient Routing in Opportunistic Networks. Journal of Communications., **10**, 48-54 (2015)

4. Hu, J., Du, Y., Mo, H., Wei, D., Deng, Y. : A modified weighted TOPSIS to identify influential nodes in complex networks. ELSEVIER., **1**, 1-13 (2015)

5. Jalili, M., Salehzadeh-Yazdi, A., Asgari, Y., Arab, S.S., Yaghmaei, M., Ghavamzadeh, A., Alimoghaddam, K. : CentiServer: A Comprehensive Resource,Web-Based Application and R Package for Centrality Analysis. PLOS ONE., doi:10.1371/journal.pone.0143111, 1-8 (2015)

6. Jalili, M., Salehzadeh-Yazdi, A., Gupta, S., Wolkenhauer, O., Yaghmaei, M., Resendis-Antonio, O., Alimoghaddam, K. : Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. Frontiers in Physiology., **7**, 1-4 (2016)

7. Lee, S., Huang, E.J. : Modeling ALS and FTD with iPSC-derived neurons. ELSEVIER, Brain Research., **1656**, 88-97 (2017)

8. Li, J., Duenas-Osorio, L., Chen, C., Berryhill, B., Yazdani., A.: Characterizing the topological and controllability features of U.S. power transmission network. ELSEVIER., **453**, 84-98 (2016)

9. Liu, M.L., Zang, T., Zhang, C.L. : Direct Lineage Reprogramming Reveals Disease-Specific Phenotypes of Motor Neurons from Human ALS Patients. Cell Press., **4**, 115-128 (2016)

10. Nie, T., Guo, Z., Zhao, K., Lu, Z.M. : Using mapping entropy to identify node centrality in complex networks. ELSEVIER., **453**, 290-297 (2016)

11. Shirdel, G.H., Abdilhosseinzadeh, M. : The critical node problem in stochastic networks with discrete-time Markov chain. CROOR., **3**, 33-46 (2016)

12. Ventrsca, M., Aleman, D. : Efficiently identifying critical nodes in large complex networks. Computational Social Networks, Springer., **2**, 1-16 (2015)

13. Yang, Y., Xie, G. : Efficient identification of node importance in social networks. ELSEVIER., **17**, 1-12 (2016)

14. Zhao, J., Fang, Z. : Research on Campus Bike Path Planning Scheme Evaluation Based on TOPSIS Method: Wei'shui Campus Bike Path Planning as an Example. ELSEVIER., 137, 858-866 (2016)

Figures

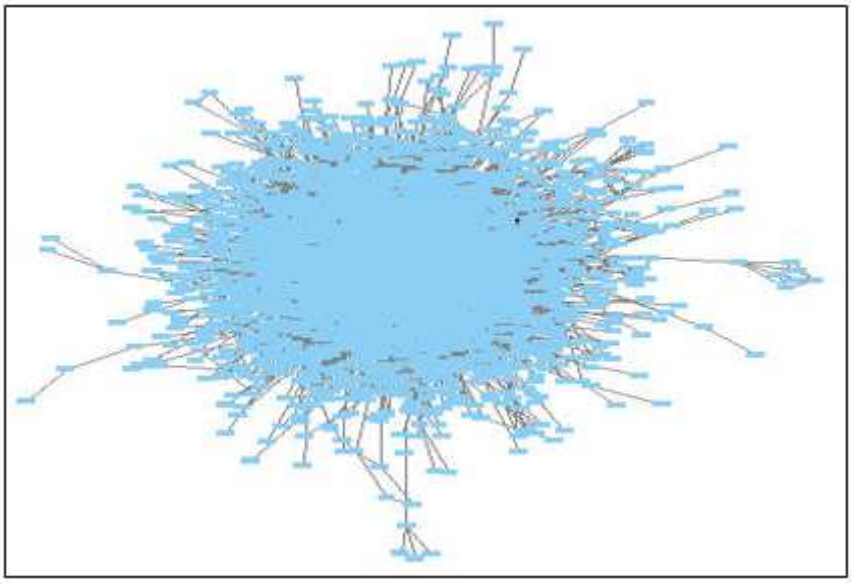

Figure 1

E. Coli network

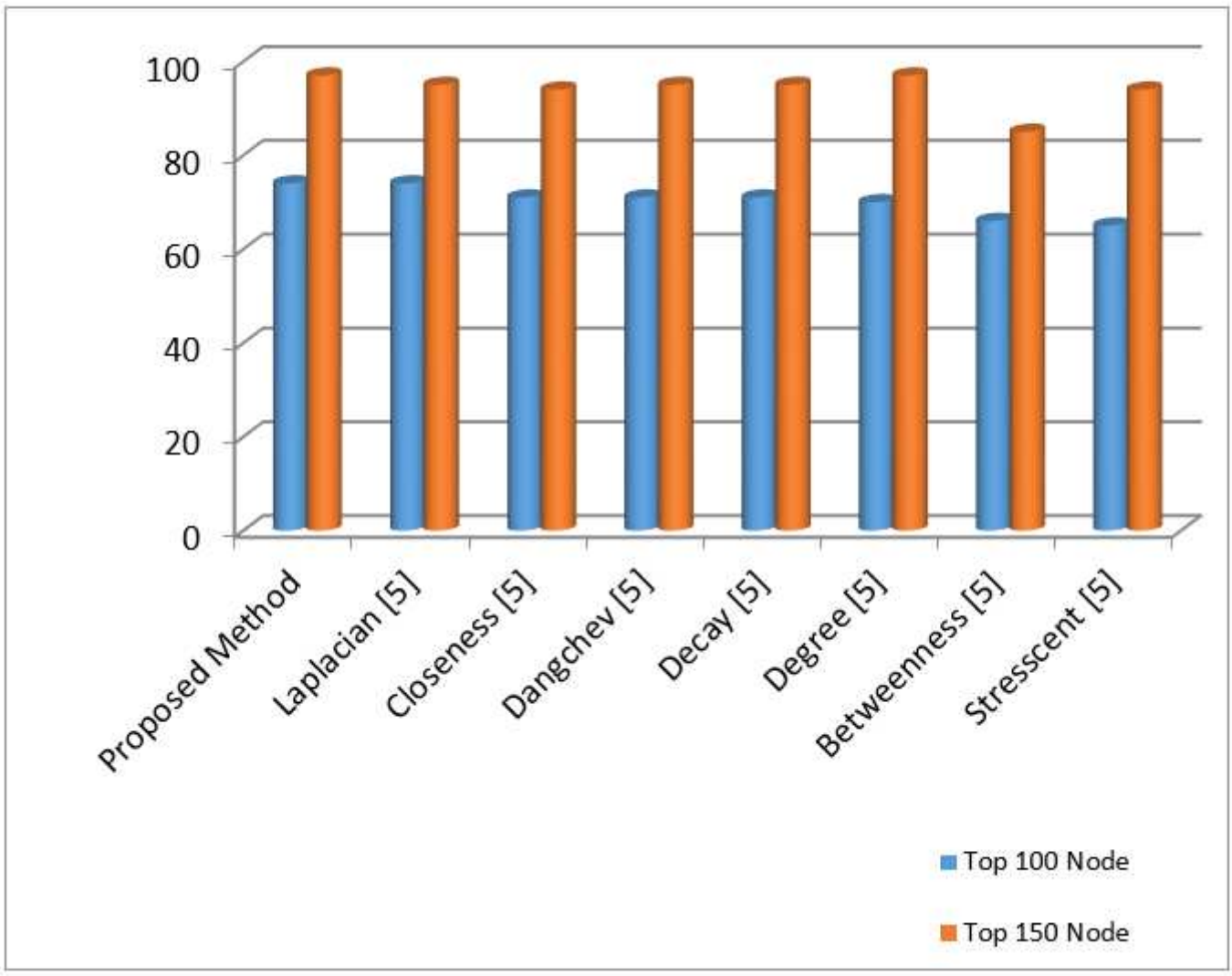

Figure 2

Illustrates these results