VOL. 18, NO. 3, FEBRUARY 2023 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

GLOBAL STRUCTURE MODEL MODIFICATION TO IMPROVE INFLUENTIAL NODE DETECTION

Mohd. Fariduddin Mukhtar[1,2], Zuraida Abal Abas[1], Amir Hamzah Abdul Rasib[2], Siti Haryanti Hairol Anuar, Nurul Hafizah Mohd. Zaki, Zaheera Zainal Abidin, Siti Azirah Asmai and Ahmad Fadzli Nizam Abdul Rahman
[1]Faculty of Communication and Information Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]Faculty of Mechanical and Manufacturing Engineering Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
E-Mail: fariduddin@utem.edu.my

ABSTRACT

Improving a network's robustness and information acceleration requires assessing the value of its nodes, which has been a central issue in network research. The concept of centrality is crucial since it allows for determining the most important nodes. It is possible to find prominent nodes with the help of centrality indices, but they have computational complexity and are limited by the singularity function. The global structure model (GSM) is one method that helps find these impactful nodes. One of the problems with using GSM is that it ignores these nodes' local information. To address this issue, we propose that considering the features of each index individually and then combining them can result in more accurate detection of influential nodes. An experiment incorporated four attributes: global and local impacts, random walk structure, and node position. In this research, we simulate a real-world network using the SIRIR model to derive its propagation process and then verify its efficacy with measures like the Jaccard similarity score and Kendall's correlation coefficient. According to the findings of the experiments, the Degree of Centrality of the local features has a substantial effect when combined with GSM.

Keywords: centrality indices, combine, SIR.

INTRODUCTION

Identifying the most crucial nodes for a network to remain stable and resilient is critical. These nodes can serve to accelerate information dissemination [1]. If nodes that do not satisfy the standard are employed, the network's structure and propagation may be affected. As a result, selecting the correct nodes might significantly impact the network's propagation.

The graph model includes four core analyses: community discovery, connectivity analysis, path analysis, and centrality. Centrality is a tool that offers an indicator to find the most important nodes. A centrality detects the most influential people in a social network [2]–[4], identifies central [5, 6] or urban networks, and identifies disease spreaders [7]. A network's centrality can be measured in various ways, but each has its limitations. Researchers have discovered that a node's ability to influence others should be comprised of the criteria listed to develop an efficient process for locating key nodes [8]–[10]:

Global Influence

Centrality metrics such as Katz Centrality (KC) Between's Centrality (BC), and Closeness Centrality (CC) evaluate a node's ability to influence global structural information. Although these measurements have good performance, they are frequently too complex for extensive networks, rendering them unsuitable.

Local Influence

In various centrality assessments, local information, such as degree, semi-local, and degree distance, can be utilized to determine a node's influence capacity. Local metrics, when employed alone, are simple yet unproductive because they only consider data from the local area. Some local metrics rely solely on information gleaned from the immediate surroundings to achieve higher rankings.

Random Walk Structure

It is possible to uncover influential nodes using random walk algorithms like Eigenvector Centrality (EC), Page Rank (PR), and Hypertext-induced topic search (HITS). Computational complexity is a significant issue because of how many iterative operations these methods require. For example, PR performs well in directed networks but degenerates to DC in undirected networks, which is ineffective.

Node's Position

K-shell (KS) decomposition says that a node's influence in the network is based on its location. KS can provide excellent performance while consuming minimal processing power in large networks. However, the value of KS for nodes influence consists of more nodes at one time.

Combining different metrics can further improve the discovery of the most significant node results. Considering more than one character of a node's topology is a good way to determine a node's influence. Wang *et al.* [11] presented a new method for calculating a node's importance, considering DC and its neighbours' degree. Yu *et al.* [12] considered a node's relevance in conjunction with the BC and KC. The global structure model (GSM) was also introduced, which considers the self and global influence of nodes composed of the core decomposition of the network [4].

However, current combination strategies have limitations since these methods are often complex to compute, making them hard to apply on large networks.

Second, the contributions of various qualities are treated as having the same value, making it hard to improve performance. The position attribute is one of the essential attributes used to rank nodes. The nodes' features around it significantly impact a node's potential to be influential.

This research builds on our work in [13], where we found that using multiple metrics together was better than using a single metric alone. The differences are that in this study, we extend the combinations procedure based on different metrics characteristics. We choose three primary indices; Degree Centrality (DC) based on local characteristics, PageRank, and HITS from global perspectives to be integrated with the global structural model (GSM). We are interested in determining whether the modification of GSM, which involves integrating each index separately, is capable of amplifying the detection of influential nodes. To evaluate the efficacy of the proposed method, the SIRIR model is used to analyze the nodes' capacity for transmitting the infection to other nodes.

PRELIMINARIES

This section will overview the overall study concept and explain the elements used in the current investigation. GSM argues that a node's influence includes not only its influence but also the influence of other nodes in the network. The model's impact is the K-shell value and neighbouring layer nodes combined with the path length to form a global influence. GSMs consists of two components: self-influence (SI) and global influence (GI). For each of these, GSM employs the KS. The amount of self-influence was calculated by dividing the natural logarithms of the KS by the total number of nodes, N. In the case of global, a node's influence includes the influence of any other nodes to whom it is connected. The formula for calculating GSM is in (1), where d_{ij} is the shortest distance between nodes.

$$GSM(v_i) = SI(v_i) \times GI(v_i) = e^{\frac{Ks(v_i)}{N}} \times \sum_{i \neq j} \frac{Ks(v_j)}{d_{ij}} \quad (1)$$

For the applied indices, we represent undirected and un weighted networks as graph $G = (V, E)$, $V = \{v_1, v_2, ..., v_n\}$ representing the set of nodes; $E = \{e_1, e_2, ..., e_n\}$ as the set of edges. Four indices involved in this study which is as follows:

Degree centrality (DC) is the most common and straightforward approach for measuring relationships. It tallies the number of connections between network nodes. Nodes with a greater degree have a more significant influence than nodes with a lesser degree. Due to its ease of computation, DC uses extremely little information compared to global resources. Nodes with low and high degrees are good candidates for influential nodes.

Page rank (PR) is a well-known web page ranking algorithm that Google utilizes. PR can accurately identify the network's influence node with minimal computational complexity. PR is defined for all sizes of directed graphs, although it suffices for small-scale graphs in undirected situations.

Hypertext-induced topic search (HITS) provides each node with two criteria that measure the influence of a node (hubs) and its ability to spread information in the network (authorities). HITS focuses on ranking network nodes.

K-shell decomposition (KS) is a popular approach that assesses centrality by splitting the network based on a node's existing degree. The procedure begins by deleting all nodes of degree 1. The technique is repeated for each layer of propagation until only nodes of a higher degree remain. According to research, KS is a good indicator for identifying group nodes with a strong influence but low individual rank.

A basic investigation was conducted to examine the convergence of node-finding measures. Figure-1 depicts a network of 13 nodes and 17 edges, while Table-1 shows the ranking analysis. The KS-value is represented in this figure by the letter k. Nodes with a higher k value are more influential.

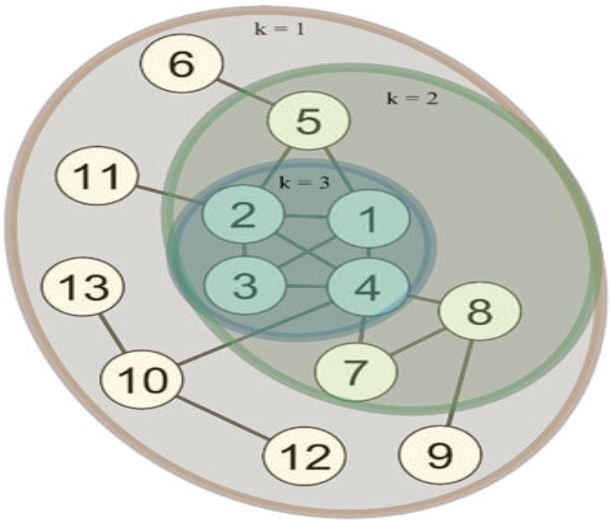

Figure-1. Sample network with KS partitions.

Table-1. Sample network analysis.

Rank	DC	PR	HITS	GSM
1	4	4	4	4
2	2	2	2	2
3	1	1	10	1
4	3, 5, 8, 10	3	1	3
5	7	5	8	5, 8
6	6, 9, 11, 12,13	8	5	7
7		7	3	10
8		10	7	11
9		11	12	9
10		6	13	6

DC divides all nodes into six tiers, and it isn't easy to distinguish the top ten rankings using DC due to connection value similarities. Although their ranks differ, PR, HITS, and GSM are all capable of separating nodes into ten tiers. PR and GSM identify equivalent node ranks for the first four rankings. However, GSM recognizes nodes 5 and 8 to be in rank 5. This could be since the two nodes are locally symmetric and contain the same information about themselves and their neighbours, resulting in the same influence. While for HITS, everything is fine until the second rank, where it diverges, indicating that node 10 is more necessary to be in the third rank than node 1. This would be a significant flaw in HITS because Figure-1 demonstrates that node 1 is substantially more critical and has far more connections than node 10. Our findings imply that to boost the detection of impacting nodes, we must continue our investigation by integrating DC, PR, and HITS with GSM.

MATERIALS AND METHODS

Datasets

This research is conducted to observe if the proposed strategies work on a small network. Three real networks will be employed, each undirected and unweighted. These data sets are publicly available on the internet. The datasets used in this study were as follows:

a) Adjectives and nouns network (Word): Charles Dickens's novel David Copperfield contains an adjacency network comprising frequent adjectives and nouns. There are 112 nodes and 425 edges in this graph.

b) Les miserables network (LesM): The network is about the relationships among actors involved in the novel Les Miserables with 74 nodes and 248 edges.

c) Zachary network (Zachary): The network is a friendship between members of a karate club at a US university. 78 edges and 34 nodes represent the friendship relationship.

SIRIR Models

The spreading influence of the top-ranking nodes was evaluated using the SIRIR model [14]. It is a revolutionary ranking mechanism based on the susceptible-infected-recovered with improved influence node ranking, which can assess information transmission. Node influence simulation in this model begins with an infected node as the starting point. Then the infected node spreads to its nearest neighbour with a specified probability, and the infected node recovers with the likelihood $\lambda = 1$ until the process is stable. In this study, we take on the average of 500 simulations based on three levels of probability propagation (β) for 0.3, 0.5, and 0.9. SIRIR asserts that it considers the self-local connectivities and the topological influences exerted by the nodes across the entire network.

Jaccard Similarity Score

The Jaccard similarity (JS) score compares two datasets by counting the number of components in each category. Mathematically, JS can be determined by dividing the intersection of sets by the union of sets and then multiplying that result by one. A higher value indicates a stronger relationship between the two datasets.

Kendall's τ Correlation Coefficient

The Kendall correlation coefficient, as defined in (2), where n_c and n_d indicate the number of concordant and discordant pairs, is used to compare the correlation consistency of two lists ranking the same set of things. A high value τ suggests that two ranking lists are comparable in ranking, while a low value shows that the ranking lists are divergent. In general $\tau \in [-1,1]$, where $\tau > 0$ indicates a positive correlation, and vice versa.

$$\tau(x,y) = \frac{n_c - n_d}{0.5n(n-1)} \quad (2)$$

RESULTS AND ANALYSIS

Nodes Detection using DC, PR, HITS, and GSM

We looked at all the metrics, which included DC, PR, HITS, and GSM, to witness the propagation nodes ranking using SIRIR. A ranking correlation value among two compared nodes ranked to each other will be assigned by Kendall. Figure-2 displays the analysis of the correlation with the SIR model. Compared with infections at various rates, PR and DC are strikingly comparable. This attitude demonstrates that the PR value for an undirected graph will eventually approach the DC value. HITS converge differently for each of the other networks and have a relatively moderate rate for both $\beta = 0.9$. On the other hand, nodes using GSM in the SIR model are adequate compared with other methods. In the subsequent stage, we will incorporate DC, PR, and HITS into GSM and then use the SIRIR model again to determine the propagating influence.

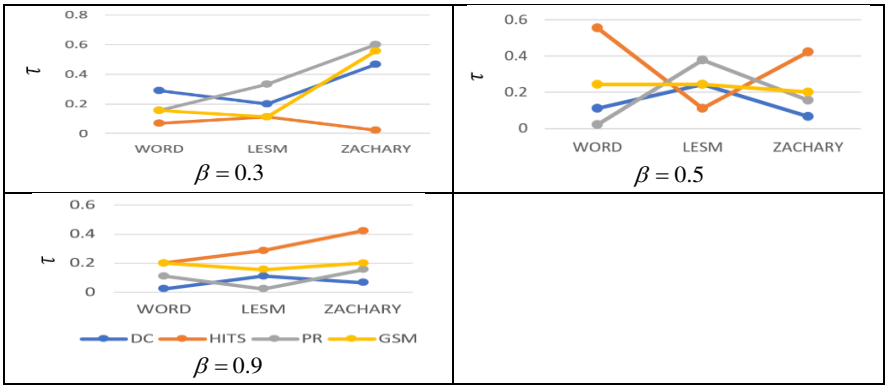

Figure-2. Kendall coefficient comparison of nodes ranking at different infection rates (0.3, 0.5 and 0.9) respectively.

Nodes Detection by Combined Indices

The symbolic regressions (SR) approach was applied to combine GSM and the necessary indices. It is feasible to use SR to identify the most accurate and sophisticated model for a given dataset by searching through all the different mathematical expressions that can be used. The SR method combines fundamental functions to arrive at simple equations that accurately forecast the target variables. Turing Bot, a machine learning program, is being used to approximate the global optimum of a given function to carry out the SR technique. To create a regression model, we'll use several different mathematical procedures, such as arithmetic, trigonometric, and exponential. An example of a potential combination of regression models for Word is listed in Table-2. DCGSM, PRGSM, and HITSGSM are GSM-based metrics combinations. Notations 1, 2, and 3 represent the infection rate increasing from 0.3 to 0.5 to 0.9, respectively.

The newly generated regression formulae were assessed and compared using the top ten most significant node outputs. Each network was compared using the Jaccard Similarity (JS) score and Kendall's τ -correlation analysis.

Table-2. Regressions model for Word network.

Regression model
$DCGSM1 = 0.471996+\left(0.0153482*\left(0.046145-tan\left(DC+\left(GSM-2.75983\right)\right)+\left(\begin{matrix}\left(17.2018-12.7528*cos\left(0.913322*DC\right)\right)*\\ \left(-0.44436+cos\left(cos\left(DC-tan\left(-8.77067*DC\right)\right)+\left(-6.71426*GSM\right)\right)\right)+DC\end{matrix}\right)\right)\right)$
$DCGSM2 = 0.0897559*\left(DC-\left(\begin{matrix}cos\left(\left(-0.49751\right)*\left(\left(\left(-2.43529\right)/DC\right)+\left(GSM*DC-tan\left(1.01737+\left(GSM-tan\left(DC\right)\right)\right)-0.629506\right)\right)\right)\\ /0.0695668\end{matrix}\right)\right)-1.69875$
$DCGSM3 = \left(\left(0.385107+cos\left(DC-GSM\right)\right)/\left(0.497407*DC\right)\right)+\left(\left(tan\left(1.65923-GSM\right)/cos\left(-0.0417621*GSM\right)\right)/GSM\right)+0.102925*DC$
$HITSGSM1 = \left(1.53117-cos\left(\left(-68.892\right)*\left(\left(-0.119007\right)/HITS\right)-\left(\begin{matrix}\left(-1.08005+sin\left(14.1822*\left(0.0659079-HITS\right)\right)\right)*\\ cos\left(\left(0.625266-HITS-0.008089-0.152077\right)*GSM+1.78071\right)\end{matrix}\right)\right)\right)*\left(HITS+0.239781\right)$
$HITSGSM2 = -1.6332+\left(\left(\tan\left(74.6889*\left(0.177093/\left(1.9997*HITS\right)\right)\right)-6.17071*\sin\left(-0.133728*GSM\right)+7.98067\right)*HITS\right)$
$HITSGSM3 = \left(tan\left(3.95785*\left(-0.00980377*GSM\right)\right)+tan\left(-0.422476*GSM\right)-tan\left(GSM/0.770675\right)+11.5821\right)*HITS$
$PRGSM1 = 0.313268-0.238885*cos\left(tan\left(0.54894-GSM\right)*PR-\left(0.098912*\left(\left(-68.0281+PR*tan\left(GSM+1.14902\right)\right)*GSM\right)\right)\right)+\left(PR/0.0588858\right)$
$PRGSM2 = -3.20382+\left(\left(\left(tan\left(\left(0.0456521+3*GSM\right)/9.6487\right)+\left(93.6435+\left(\begin{matrix}tan\left(GSM+0.532006\right)+\\ \left(PR*GSM-TAN\left(PR-\left(-1.00017+GSM\right)\right)\right)\end{matrix}\right)\right)\right)\right)*\left(0.0151578+PR\right)\right)$
$PRGSM3 = \left(\begin{matrix}tan\left(GSM-1.99183\right)+93.8752+tan\left(-1.99244+\left(-0.474512*GSM\right)\right)+tan\left(0.00545797*GSM\right)+tan\left(0.017907*GSM\right)+tan\left(0.2142*GSM\right)\\ +tan\left(GSM\right)\end{matrix}\right)*PR$

Table-3 demonstrates the analysis of JS-index value for Word, LesM, and Zachary networks. In Word, both DC and PR resemble striking. All the combined indices in JS contributed considerably to an increase in the ranking detection toward SIR. While compared to the original PR and SIR3 in JS, PRGSM3 demonstrated the most significant progress (0.5385 to 0.1111). InLesM, both DC and PR show an upward tendency in their respective JS indexes. The value of HITS has a smaller amount in common with the overall weight. Despite SIR demonstrating an increase in the JS value, DC presents a considerable rise in the combination of measures. The SIR similarity of DCGSM1, DCGSM2, and DCGSM3 increased to 0.5385 compared to the single DC. Similarly goes Zachary network, JS's value for DC presents a pretty high value as the infection rate increases compared to

HITS and PR in Table-5. The pattern continues for each DCGSM with SIR.

Figure-3 illustrates Kendall's correlation coefficient. In Word, PRGSM drops significantly when the infection rate increases from 0.5 to 0.9. Only DCGSM shows consistent dispersion in the τ-scores, which arrange between 0.3 to 0.4. Although HITS has a more substantial value at 0.5, this value drops abruptly when infection rates increase. As for LesM, results show that DCGSM has the highest position τ-value compared to other strains, which rises considerably whenever the infection rate increases. When examining the correlation of Zachary, it can be noticed that DCGSM has a consistent τ arranging between an infection rates of 0.4 to 0.9 compared to others despite PRGSM fast increasing at 0.9.

Table-3. Jaccard similarity score of combined indices with SIR for (a) Word, (b) LesM and (c) Zachary.

(a)

JACCARD SIMILARITY INDEX		DC	SIR1	SIR2	SIR3
	DC		0.1765	0.1765	0.1765
	DCGSM1	0.3333	▲ 0.3333		
	DCGSM2	0.4286		▲ 0.25	
	DCGSM3	0.5385			▭ 0.1765
		HITS	SIR1	SIR2	SIR3
	HITS		0.1765	0.1765	0.1765
	HITSGSM1	0.3333	▲ 0.25		
	HITSGSM2	0.3333		▲ 0.3333	
	HITSGSM3	0.4286			▲ 0.3333
		PR	SIR1	SIR2	SIR3
	PR		0.1765	0.1765	0.1111
	PRGSM1	0.4286	▲ 0.25		
	PRGSM2	0.5385		▭ 0.1765	
	PRGSM3	0.3333			▲ 0.5385

(b)

JACCARD SIMILARITY INDEX		DC	SIR1	SIR2	SIR3
	DC		0.25	0.3333	0.3333
	DCGSM1	0.4286	▲ 0.5385		
	DCGSM2	0.3333		▲ 0.5385	
	DCGSM3	0.5385			▲ 0.5385
		HITS	SIR1	SIR2	SIR3
	HITS		0.1765	0.1765	0.1111
	HITSGSM1	0.25	▲ 0.3333		
	HITSGSM2	0.1111		▲ 0.25	
	HITSGSM3	0.1765			▲ 0.4286
		PR	SIR1	SIR2	SIR3
	PR		0.25	0.4286	0.3333
	PRGSM1	0.3333	▼ 0.1111		
	PRGSM2	0.5385		▲ 0.8182	
	PRGSM3	0.4286			▲ 0.4286

(c)

JACCARD SIMILARITY INDEX		DC	SIR1	SIR2	SIR3
	DC		0.25	0.25	0.6667
	DCGSM1	0.25	▲ 0.6667		
	DCGSM2	0.1765		▲ 0.6667	
	DCGSM3	0.6667			▲ 0.6667
		HITS	SIR1	SIR2	SIR3
	HITS		0.3333	0.25	0.5385
	HITSGSM1	0.3333	▲ 0.6667		
	HITSGSM2	0.1765		▲ 0.6667	
	HITSGSM3	0.4286			▼ 0.4286
		PR	SIR1	SIR2	SIR3
	PR		0.25	0.25	0.6667
	PRGSM1	0.4286	▲ 0.4286		
	PRGSM2	0.1765		▲ 0.6667	
	PRGSM3	0.5385			▲ 0.6667

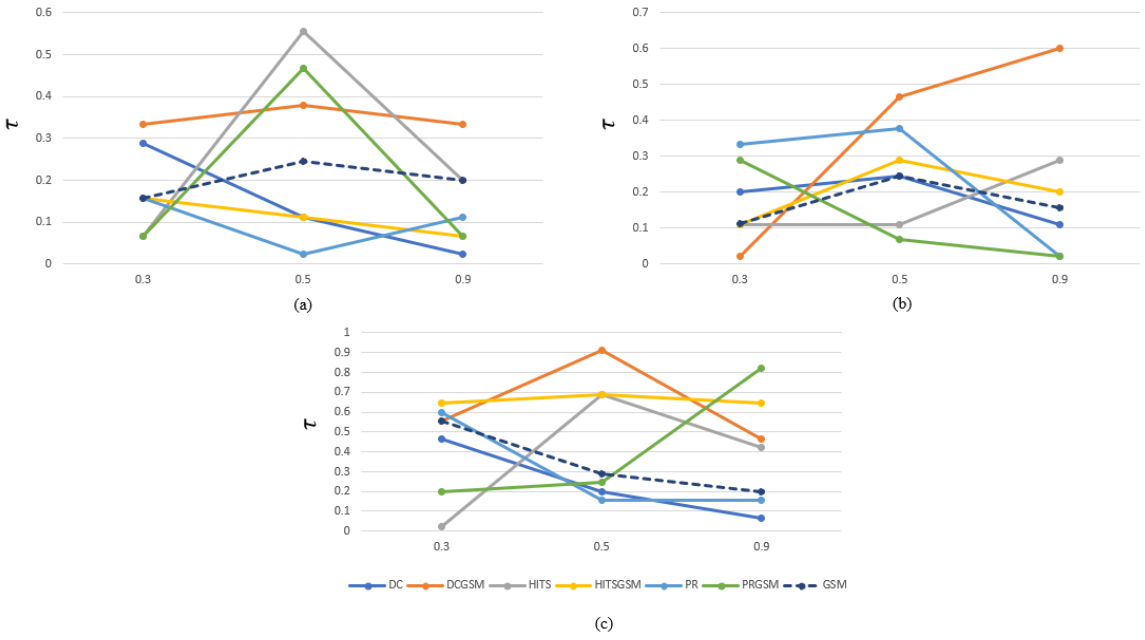

Figure-3. Kendall correlation coefficient of combined indices with SIR for (a) Word, (b) Les M, and (c) Zachary.

CONCLUSIONS

This paper focused on the impact of different combinations of centrality indices. We proposed various combinations of indices based on different characteristics that can boost the detection of the most prominent nodes in a network. Results reveal an increment in the Jaccard similarity score and Kendall's correlation coefficient when the proposed combinations are compared with DC regarding the influential nodes ranking in the SIR model, where DCGSM outperforms PRGSM and HITSGSM. This is because, comparing SIR with the original GSM, we observed that the detection of the most influential nodes was increased after the integration with DC. This would imply that combining DC and GSM would result in an increase in the detection and affect nodes' rank. Since GSM is composed of implications for both self and global influence, the integration of GSM with DC would increase GSM's level of self-influence.

Despite this, determining precise metric combinations for GSM can be a challenging task. The generated regression model is difficult to understand and will require simplification in subsequent research. Studying directed or weighted networks could add up to more value for the body of knowledge, and the scope of

the study can also be expanded to include more sophisticated and extensive networks.

ACKNOWLEDGEMENT

Appreciation to the Ministry of Education Malaysia for the research funding, Malaysia Research Assessment (MyRA), Universiti Teknikal Malaysia Melaka (UTeM), Faculty of Communication and Information Technology (FTMK), and all authors appreciate the valuable feedback from the proficient reviewers.

REFERENCES

[1] S. Gao, J. Ma, Z. Chen, G. Wang and C. Xing. 2014. Ranking the spreading ability of nodes in complex networks based on local structure. Physica A: Statistical Mechanics and its Applications, 403: 130-147, doi: 10.1016/j.physa.2014.02.032.

[2] J. S. More and C. Lingam. 2019. A SI model for social media influencer maximization. Applied Computing and Informatics, 15(2): 102-108, doi: 10.1016/j.aci.2017.11.001.

[3] J. Wu, J. Shen, B. Zhou, X. Zhang and B. Huang. 2019. General link prediction with influential node identification. Physica A: Statistical Mechanics and its Applications, 523: 996-1007, doi: 10.1016/j.physa.2019.04.205.

[4] A. Ullah, B. Wang, J. F. Sheng, J. Long, N. Khan and Z. J. Sun. 2021. Identification of nodes influence based on global structure model in complex networks. Scientific Reports, 11(1), doi: 10.1038/s41598-021-84684-x.

[5] P. Devi, A. Gupta and A. Dixit. 2014. Comparative Study of HITS and Page Rank Link based Ranking Algorithms. [Online]. Available: www.ijarcce.com

[6] G. Nomikos, P. Pantazopoulos, M. Karaliopoulos and I. Stavrakakis. 2014. Comparative assessment of centrality indices and implications on the vulnerability of ISP networks. 2014 26th International Teletraffic Congress, ITC 2014, no. 288021, doi: 10.1109/ITC.2014.6932932.

[7] S. M. Jenness, S. M. Goodreau and M. Morris. 2017. EpiModel: An R Package for Mathematical Modeling of Infectious Disease over Networks.

[8] S. P. Borgatti and M. G. Everett. 2006. A Graph-theoretic perspective on centrality. Social Networks, 28(4), doi: 10.1016/j.socnet.2005.11.005.

[9] J. Wang, C. Li and C. Xia. 2018. Improved centrality indicators to characterize the nodal spreading capability in complex networks. Applied Mathematics and Computation, 334: 388-400, doi: 10.1016/j.amc.2018.04.028.

[10] M. Simsek and H. Meyerhenke. 2020. Combined Centrality Measures for an Improved Characterization of Influence Spread in Social Networks. doi: 10.1093/comnet/cnz048.

[11] W. Jianwei, R. Lili and G. Tianzhu. 2008. A new measure of node importance in complex networks with tunable parameters. 2008 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2008, pp. 1-4, doi: 10.1109/WiCom.2008.1170.

[12] Y. Zhang, Y. Bao, S. Zhao, J. Chen and J. Tang. 2016. Identifying Node Importance by Combining Betweenness Centrality and Katz Centrality. In Proceedings - 2015 International Conference on Cloud Computing and Big Data, CCBD 2015, pp. 354-357. doi: 10.1109/CCBD.2015.19.

[13] M. F. Mukhtar *et al*. 2022. Identifying Influential Nodes with Centrality Indices Combinations using Symbolic Regressions. International Journal of Advanced Computer Science and Applications, 13(5), doi: 10.14569/IJACSA.2022.0130570.

[14] A. Salavaty, M. Ramialison and P. D. Currie. 2020. Integrated Value of Influence: An Integrative Method for the Identification of the Most Influential Nodes within Networks. Patterns, 1(5), doi: 10.1016/j.patter.2020.100052.