Hybrid Global Structure Model for Unraveling Influential Nodes in Complex Networks

Mohd Fariduddin Mukhtar¹, Zuraida Abal Abas², Amir Hamzah Abdul Rasib³, Siti Haryanti Hairol Anuar⁴, Nurul Hafizah Mohd Zaki⁵, Ahmad Fadzli Nizam Abdul Rahman⁶, Zaheera Zainal Abidin⁷, Abdul Samad Shibghatullah⁸ Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia^{1,2,4,5,6,7} Fakulti Teknologi Kejuruteraan Mekanikal dan Pembuatan, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia^{1,3} Institute of Computer Science & Digital Innovation, UCSI University, 56000 Cheras, Kuala Lumpur, Malaysia⁸

Abstract—In graph analytics, the identification of influential nodes in real-world networks plays a crucial role in understanding network dynamics and enabling various applications. However, traditional centrality metrics often fall short in capturing the interplay between local and global network information. To address this limitation, the Global Structure Model (GSM) and its improved version (IGSM) have been proposed. Nonetheless, these models still lack an adequate representation of path length. This research aims to enhance existing approaches by developing a hybrid model called H-GSM. The H-GSM algorithm integrates the GSM framework with local and global centrality measurements, specifically Degree Centrality (DC) and K-Shell Centrality (KS). By incorporating these additional measures, the H-GSM model strives to improve the accuracy of identifying influential nodes in complex networks. To evaluate the effectiveness of the H-GSM model, real-world datasets are employed, and comparative analyses are conducted against existing techniques. The results demonstrate that the H-GSM model outperforms these techniques, showcasing its enhanced performance in identifying influential nodes. As future research directions, it is proposed to explore different combinations of index styles and centrality measures within the H-GSM framework.

Keywords—Centrality indices; combination; hybrid; global structure model; influential nodes

I. INTRODUCTION

In the captivating world of graph analytics, identifying significant nodes is critical, providing invaluable insights into the structure and behavior of many real-world networks. Networks with considerable sways, such as social networks, biological networks, and information networks, are characterized by nodes that operate as hubs or influencers, shaping the behavior of the entire network. Understanding and locating these significant nodes improves our understanding of network dynamics and offers possibilities for applications such as targeted marketing, recommendation systems, and vulnerability analyses [1], [2].

Centrality measurements are the preferred metric in network analysis for evaluating the relevance and influence of individual nodes or edges. DC [3], betweenness centrality (BC)[4], closeness centrality (CC)[5], and PageRank (PR) [6] are metrics that have been created to identify nodes that play essential roles depending on a variety of parameters. These traditional metrics primarily focus on local or global network information [7], [8], frequently failing to capture the delicate interplay between the two. Local influence measurements, such as DC and CC, focus on a node's close connections and proximity to other nodes, elucidating its impact on information or resource flow within a narrow network section[9]-[11]. Global impact metrics, on the other hand, such as BC and PR, take into account the more extensive network structure and the importance of nodes to which a specific node is connected. These metrics excel at identifying nodes that serve as bridges between distinct network groups or enhance network connectivity across the board. These global measures are limited since they are computationally expensive and do not function well in the absence of a complete network structure [9], [12]–[14]. Evaluating local and global influence is critical to have a complete sense of node relevance. Nodes strongly influenced at both scales will likely hold critical positions within the network's complicated structure. They could impact immediate network behavior while altering its overall structure and dynamics.

Ullah's[15] Global Structure Model (GSM) provides a framework for ranking nodes in a network based on their local and global significance. This model uses the K-shell value to assess individual influence while considering neighboring node K-shell values and including path length to determine the global effect. It is worth noting, however, that the GSM falls short of adequately capturing the impact of path length, allowing the opportunity for further improvement. To remedy this issue, an improvement of GSM (IGSM) [16] has been made, which uses DC as its primary parameter rather than KS. Despite these advancements, precisely assessing the value of individual nodes within complex networks remains a substantial problem, emphasizing the need for ongoing study and developing novel ways to acquire more profound insights into network topologies.

This study continues our prior efforts in [17] and [18]. We successfully identified indices based on their similarity, demonstrating the improved performance obtained by combining indices from different network topologies, particularly when incorporating local and global centrality metrics. Next, we enhanced our findings by integrating the GSM with local and global centrality measurements by proposing a new hybrid model (H-GSM). Based on these findings, the current study intends to improve the algorithm using DC and KS framework.

The primary contribution of this study lies in the creation and evaluation of the H-GSM framework. The H-GSM framework combines the GSM with the comprehensive analysis provided by DC and KS, resulting in an improved ability to identify influential nodes. Notably, this integration enables the capturing of the intricate relationship between local and global network information, which is often overlooked by conventional centrality metrics. The findings of this study offer valuable insights for further exploration of different combinations of index styles and centrality measures, thereby advancing the understanding and application of network analysis methodologies.

II. PRELIMINARIES

GSM and IGSM consider the node's self-influence and global influence, except that GSM believes in KS-decomposition, while IGSM is the improvement that applied DC. The formula is expressed as follows:

$$GSM(i) = SI(i) \times GI(i) = e^{\frac{KS(i)}{n}} \times \sum_{i \neq j} \frac{KS(j)}{d_{ij}}$$
(1)

$$IGSM(i) = improved _SI(i) \times improved _GI(i)$$

$$= e^{\frac{DC(i)}{n}} \times \sum_{i \neq j} \frac{DC(j)}{d_{ij}^{ceil(\log_2(aver_deg ree))}}$$
(2)

where KS(i) refers to the K-shell decomposition value of node *i*, DC(*i*) refers to the degree value of a node, and d_{ij} refers to the path length between node *i* and node *j*.

These methods have grown in popularity because they are straightforward and enable researchers to quickly collect node values and use them in massive networks, broadening the scope of their potential applications. However, determining each node's significance within a network accurately presents a significant barrier for both the K-shell decomposition and the DC techniques. There need to be more levels in the K-shell decomposition method, which causes many nodes to be assigned to the same level[12], [19]. While DC only considers edge information, it ignores other essential elements like the structure of the entire network [20]. As a result, it might be challenging to accurately distinguish each node's relevance in large-scale networks where many nodes may have identical DC values. DC needs to recognize the significance of position data within the network. Due to their distinct placements or responsibilities in tying together various network regions, nodes with the same DC value may have differing degrees of influence. Alternative approaches that consider additional network properties and the impact of location information are needed to get over these constraints and develop a more thorough knowledge of node influence.

In this study, a hybrid approach of global structural model is proposed. It is suggested that the overall structure of the network is influenced by each node's capacity to impact itself. Because of this, the self-influence participant is essential to global influence. The following is how our suggested method is expressed:

$$H - GSM(i) = hSI(i) \times hGI(i)$$

= $e^{\frac{DC(i) + KS(i)}{n}} \times \sum_{i \neq j} \frac{e^{\frac{DC(i) + KS(i)}{n}}}{d_{ij}^{ceil(\log_2(aver_-hSI))}}$ (3)

Fig. 1 is a sample network of 7 nodes and 10 edges, showing each node's classification from the KS-Decomposition. Using node 3 as an example, the overall steps for calculating H-GSM is shown.



Fig. 1. Sample network.

Step 1: Determine KS and DC values.

DC(3) = 5 KS(3) = 3

Step 2: Calculate hybrid self-influence (hSI).

hSI(3) = $e^{\left[\frac{DC(3)+KS(3)}{n}\right]}$ = $e^{\left[\frac{5+3}{7}\right]}$ = 3.1357

Step 3: Calculate hybrid global influence (iGI).

$$hGI(3) = \sum_{i \neq j} \frac{hSI(j)}{d_{ij}^{cell(\log_2(ave_-hSI))}}$$

= $\frac{2.7183}{1^{cell(\log_2(2.1944))}} + \frac{2.3564}{1^{cell(\log_2(2.1944))}} + \frac{2.7182}{1^{cell(\log_2(2.1944))}}$
+ $\frac{1.3307}{2^{cell(\log_2(2.1944))}} + \frac{1.7708}{1^{cell(\log_2(2.1944))}} + \frac{1.3307}{1^{cell(\log_2(2.1944))}}$
= 10.9777

Step 4: Calculate node influence of H-GSM.

$$H$$
-GSM(3) = hSI(3) × hGI(3)
= 3.1357 × 10.9777
= 34.4228

Table I presents the rankings of centrality indices of different methods. Notably, the H-GSM metric consistently assigns rankings to nodes, distinguishing it from other centrality indices. The methodology is expanded through experimentation to conduct a comprehensive analysis of a more complex network.

Rank	DC	BC	CC	PR	GSM	IGSM	H-GSM
1	3	3	3	3	3	3	3
2	1, 2	1	1,2	0	1,2	2	2
3	0	2	0	2	0	1	0
4	5	0, 4, 5, 6	5	1	5	0	1
5	4,6		6	5	6	5	5
6			4	4	4	6	6
7				6		4	4

TABLE II. RANKING OF NODES OF THE SAMPLE NETWORK

III. METHODOLOGY

A. Datasets

This investigation utilized nine distinct real-world datasets, each featuring varying network sizes and unweighted attributes, to conduct further analyses. Table II presents information regarding the intricacies and classification of the network. The networks are available for download from KONECT (http://konect.cc/networks/) and NETWORK (http://networkrepository.com/). The variables n, m, k, and K_{max} are utilized to describe the characteristics of a network. Specifically, n represents the number of nodes, m represents the number of edges, k represents the average degree of the network, and K_{max} represents the highest degree present in the network.

TABLE III. DETAILS ON EXPERIMENTED NETWORK

Network	Туре	n	m	<k></k>	Kmax
Karate	Social	34	78	4.588	17
Netscience1	Co-authorship	379	914	4.82	34
Router	Networking	2113	6632	6.128	38

B. Experimental Environment

The experiment setup is performed on a system with configuration on Windows 11 platform 64-bit system; the machine hardware configuration is an Intel® Core i7-8550U CPU @ 2.4 Hz processor, 24 GB of RAM; and Python-Visual Studio Code 1.56.2 is used for programming.

Regarding the proposed model analysis, the model is subjected to testing and validation procedures to ensure its capacity to represent each node's relative significance accurately. The proposed model is assessed through the implementation of the following procedures:

C. SIR Model

The Susceptible-Infected-Recovered (SIR) model is wellknown for investigating each node's spreading dynamics. We will employ this section to quantify the performance of H-GSM and other benchmark centralities. All seed nodes are vulnerable for the first time. The seed node is likely to infect its nearest and next-nearest neighbor nodes (in the susceptible state) at each time step, and each node (the infected node) has a chance of recovering. This procedure was continued until no further infected nodes were discovered. Finally, all nodes gathered are used to simulate the actual node impact. S(t), I(t), and R(t) represent the number of nodes in the susceptible, infected, and recovered states, respectively. Each loop represents a time step, t, and F(t) returns the total number of infected and recovered nodes at time t, which can be used to assess the influence of the original infected node. The infected nodes will recover at step t with a probability of. When no infected nodes remain, the propagation process is complete. Identical operations are performed for each node in each network using 100 distinct SIR model iterations.

D. Comprehensive Cumulative Distribution Function

The comprehensive cumulative distribution function (CCDF) is a commonly utilized tool in network analysis to compare and analyze centrality measures. The CCDF facilitates examining the distribution of centrality values across network nodes. By comparing the CCDFs of various centrality measures, scholars can evaluate how these measures capture distinct facets of node significance and how they order nodes based on centrality. This comparative analysis offers valuable insights into the network's attributes and actions. The value of the CCDF at a given rank, r in a ranking list is obtained by summing the probabilities of all the ranks greater than r. The mathematical expression for the CCDF can be represented as:

$$CCDF(r) = 1 - \frac{\sum_{i=1}^{r} n_i}{n}$$
(5)

where n is the total number of network nodes and $\sum_{i=1}^{r} n_i$ refers to the number of numerical rankings less than or equal to r in the ranking list.

Through the graphical representation of the CCDF, it is possible to visually inspect the extreme values of the distribution, which correspond to nodes exhibiting elevated levels of centrality. This data can aid in identifying the most critical nodes within the network. A more pronounced slope of the centrality line in a CCDF plot indicates a higher degree of concentration of nodes with high centrality. In contrast, a less steep line suggests a more equitable distribution of centrality values among the nodes within the network.

E. Kendall's τ Correlation Coefficients

Kendall's τ -correlation coefficient is used to evaluate the consistency between two rankings or order of things, making it a valuable tool for comparing centrality indexes. Each centrality indices rank nodes based on their importance or centrality in a network. We may measure how well the ranks provided by different indices agree by comparing them using Kendall. Using Kendall, we may assess the degree of agreement or concordance between the levels of nodes obtained by various centrality measures. Suppose two centrality indices give similar rankings (i.e., nodes ranked highly by one index are also ranked highly by the other). The Kendall coefficient will be high, suggesting a strong positive association. If the ranks differ significantly, the Kendall tau coefficient will be low, indicating a weak connection or disagreement between the indices. We may use Kendall to statistically analyze the consistency or divergence of multiple centrality measures and acquire insights into how well they capture similar or dissimilar characteristics of node importance in a network. The formula is as follows:

$$\tau(X,Y) = \frac{2(C-D)}{n(n-1)} \tag{6}$$

where C and D are the numbers of concordant pairs discordant pairs, respectively.

IV. RESULTS AND DISCUSSIONS

Fig. 2 shows how the distribution of nodes on those three networks over time, changed for various centrality indices for the top-10 node rank. A distinct line represents each centrality index. The analysis of centrality indices reveals that H-GSM exhibits consistently higher F(t) values than other metrics. This implies they exhibit a higher degree of efficacy in disseminating the pathogen. They show unique modes of distribution. The H-GSM protocol demonstrates а comparatively gradual rise in its initial phase, a diminished apex, and a protracted duration of sustained levels. GSM-based approaches exhibit higher F(t) values than conventional centrality metrics like DC, BC, CC, and PR. According to the Karate dataset, the disease is less likely to spread effectively. Incorporating iterative and incremental elements within H-GSM enhances its capacity to propagate the pandemic. This illustrates the importance of integrating multiple components and iterative procedures in models of epidemic propagation.

The utilization of standard deviation (SD) values furnishes insights into the degree of variability exhibited by the performance of individual centrality measures. Successive display of the standard deviation values for DC, BC, CC, PR, GSM, IGSM, and H-GSM is observed. Smaller standard deviation values suggest more significant levels of consistency and stability in behavior, while larger standard deviation values indicate increased levels of unpredictability. Upon examination of the graph, it is evident that H-GSM networks exhibiting a greater quantity of nodes demonstrate a decreased standard deviation compared to networks with fewer nodes.

The CCDF plot in Fig. 3 compares various centrality indices when network nodes are eliminated. The findings indicate that the metrics above exhibit varied characteristics and levels of susceptibility toward removing nodes. By examining the diminishing patterns of the curves, one can gain insight into the unique features of each method. A linear curve devoid of inflection points implies that each node is categorized with a distinct value. At the same time, a more pronounced descent indicates a more significant number of nodes being allocated to the same rank. The swift reductions observed in DC and BC highlight their noteworthy susceptibility and impact on the broader network connectivity and dissemination of information. The data suggest that CC experiences a gradual decrease, implying a comparatively less significant influence on the overall structure of the network. The decline in PR is relatively slower, indicating a less severe impact on the network's connectivity. The slower declination of GSM, IGSM, and H-GSM results in moderate sensitivities when nodes are removed. The comparative analysis of the three approaches across the three networks reveals that H-GSM effectively discerns the impact of individual nodes.









Fig. 3. CCDF diagram of ranking results of each method.

Kendall's model with SIR was employed to assess the impact of nodes in various networks and verify the H-GSM's suitability and efficacy. Fig. 4 displays Kendall's values of the H-GSM algorithm and others under consideration. As evidenced by the data, H-GSM outperforms other methods regarding Kendall values. This indicates that H-GSM exhibits superior performance across diverse networks featuring varying node sizes.





Fig. 4. Kendall coefficients between different propagation probabilities with the SIR model.

To conduct a more comprehensive examination of the propagation phenomenon of H-GSM, we analyzed the dissemination influence of nodes ranked in the SIR model. Value of $\beta = 0.1$ was chosen, while α was varied between 0.01 and 0.1. This decision was made due to the potential for propagation across the entire network when larger alpha values are utilized. Tables III, IV, and V display the nodes ranked in the top ten for the Karate, Netscience1, and Router networks. It was observed that a significant proportion of the nodes that ranked within the top 10 of H-GSM were also present in other algorithms. Thus, the validity of the proposed H-GSM has been confirmed.

This study aimed to compare the efficacy of the proposed H-GSM with that of GSM in terms of node spreading. Consequently, in H-GSM and GSM, solely different nodes are considered seed nodes to analyze the propagation effect. A mean value of 100 rotations is calculated. In the illustrated instance presented in Fig. 5, it was observed that the impact of node 26, which is present in H-GSM, surpasses that of node 7 in GSM. The results indicate that our proposed H-GSM model exhibits a superior infection effect than the original GSM model. The findings are consistent across other networks, as depicted in Fig. 6 and 7.

(IJACSA) International	Journal of Ad	vanced Compu	ter Science	and App	licat	ions,
			Vo	l. 14, No	o. 6, 2	2023

Rank	DC	BC	CC	PR	GSM	IGSM	H-GSM
1	0	0	33	33	33	33	33
2	0	33	2	0	0	0	0
3	27	27	33	27	2	27	27
4	2	2	26	2	27	2	2
5	1	26	27	1	1	1	1
6	26	8	8	26	8	26	3
7	3	1	13	3	13	8	8
8	13	13	16	18	3	13	13
9	18	16	1	8	25	3	26
10	8	5	3	13	7	16	25

TABLE IV. TOP-10 RANKING NODES OF THE KARATE NETWORK



Fig. 5. Node propagation effect of Karate network.

TABLE V. **TOP-10 RANKING NODES OF THE NETSCIENCE1**

Rank	DC	BC	CC	PR	GSM	IGSM	H-GSM
1	3	58	58	58	4	58	3
2	4	106	119	3	3	4	4
3	58	189	106	4	5	3	58
4	5	119	44	119	13	106	5
5	72	72	187	72	14	119	119
6	219	4	107	5	28	44	13
7	119	44	4	106	29	107	44
8	13	187	6	142	16	5	106
9	142	6	130	219	15	187	107
10	106	178	135	53	119	189	219



Fig. 6. Node propagation effect of Netscience1 network.

TABLE VI. TOP-10 RANKING NODES OF THE ROUTER NETWORK

Rank	DC	BC	CC	PR	GSM	IGSM	H-GSM
1	100	2	2	100	89	100	100
2	139	0	100	139	384	139	139
3	350	100	89	62	350	2	350
4	62	139	139	0	356	89	89
5	48	159	0	99	369	0	384
6	242	508	242	159	279	242	0
7	113	99	384	350	381	99	135
8	135	350	426	2	185	62	48
9	0	62	99	242	367	384	2
10	89	179	216	310	100	350	356



Fig. 7. Node propagation effect of Router network.

V. CONCLUSIONS

In conclusion, this study addresses the challenge of identifying influential nodes in complex networks. Despite the existing methodologies, node identification remains a significant concern for researchers. To overcome this challenge, a new algorithm called H-GSM is proposed, which integrates degree and k-shell centrality measures. By incorporating both local and global centrality metrics, the H-GSM model improves upon the existing GSM model, effectively capturing the network's intricate influences. To evaluate the effectiveness of the H-GSM model, experiments are conducted on three different complex networks with varying sizes. The model's performance is assessed by examining its spreading ability using the SIR model and comparing various centrality metrics using Kendall's tau correlation coefficient. The experimental results demonstrate that the H-GSM algorithm outperforms established benchmarks in accurately identifying influential nodes. In future research, further enhancements of the algorithm's performance outcomes are planned by exploring different combinations of index styles and centrality measures. These investigations will contribute to advancing the understanding and application of network analysis techniques. Overall, the H-GSM algorithm presented in this study offers a promising approach for unraveling influential nodes in complex networks and holds potential for future advancements in the field.

ACKNOWLEDGMENT

The authors would like to thank Centre for Research and Innovation Management of Universiti Teknikal Malaysia Melaka (UTeM) for sponsoring this work under the Grant Tabung Penerbitan Fakulti dan Tabung Penerbitan CRIM UTeM.

REFERENCES

- [1] Z. A. Abas et al., "Analytics: a Review of Current Trends, Future," Compusoft, vol. 9, no. I, 2020.
- [2] J. S. More and C. Lingam, "A SI model for social media influencer maximization," Applied Computing and Informatics, vol. 15, no. 2, pp. 102–108, Jul. 2019.
- [3] L. C. Freeman, "Centrality in social networks conceptual clarification," Soc Networks, vol. 1, no. 3, 1978.
- [4] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," Sociometry, vol. 40, no. 1, 1977.
- [5] U. Brandes, S. P. Borgatti, and L. C. Freeman, "Maintaining the duality of closeness and betweenness centrality," Soc Networks, vol. 44, pp. 153–159, 2016.
- [6] P. Devi, A. Gupta, and A. Dixit, "Comparative Study of HITS and PageRank Link based Ranking Algorithms," 2014. [Online]. Available: www.ijarcce.com
- [7] Q. Shang, B. Zhang, H. Li, and Y. Deng, "Identifying influential nodes: A new method based on network efficiency of edge weight updating," Chaos, vol. 31, no. 3, Mar. 2021.
- [8] D. Chen, L. Lü, M. S. Shang, Y. C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," Physica A: Statistical Mechanics and its Applications, vol. 391, no. 4, pp. 1777–1787, Feb. 2012.
- [9] R. D. Shetty, S. Bhattacharjee, A. Dutta, and A. Namtirtha, "GSI: An Influential Node Detection Approach in Heterogeneous Network Using Covid-19 as Use Case," IEEE Trans Comput Soc Syst, 2022.

- [10] J. Zhang and Y. Luo, "Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network," 2017.
- [11] L. Candeloro, L. Savini, and A. Conte, "A new weighted degree centrality measure: The application in an animal disease epidemic," PLoS One, vol. 11, no. 11, Nov. 2016.
- [12] A. Namtirtha, A. Dutta, B. Dutta, A. Sundararajan, and Y. Simmhan, "Best influential spreaders identification using network global structural properties," Sci Rep, vol. 11, no. 1, Dec. 2021.
- [13] M. H. Ibrahim, R. Missaoui, and J. Vaillancourt, "Cross-Face Centrality: A New Measure for Identifying Key Nodes in Networks Based on Formal Concept Analysis," IEEE Access, vol. 8, pp. 206901–206913, 2020.
- [14] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method," Physica A: Statistical Mechanics and its Applications, vol. 468, pp. 278–288, 2017.
- [15] A. Ullah, B. Wang, J. F. Sheng, J. Long, N. Khan, and Z. J. Sun, "Identification of nodes influence based on global structure model in complex networks," Sci Rep, vol. 11, no. 1, Dec. 2021.
- [16] J. C. Zhu and L. W. Wang, "An extended improved global structure model for influential node identification in complex networks," Chinese Physics B, vol. 31, no. 6, Jun. 2022.
- [17] M. F. Mukhtar et al., "Identifying Influential Nodes with Centrality Indices Combinations using Symbolic Regressions," International Journal of Advanced Computer Science and Applications, vol. 13, no. 5, 2022.
- [18] M. Fariduddin Mukhtar et al., "GLOBAL STRUCTURE MODEL MODIFICATION TO IMPROVE INFLUENTIAL NODE DETECTION," vol. 18, no. 3, 2023, [Online]. Available: www.arpnjournals.com
- [19] A. Namtirtha, A. Dutta, and B. Dutta, "Identifying influential spreaders in complex networks based on kshell hybrid method," Physica A: Statistical Mechanics and its Applications, vol. 499, pp. 310–324, Jun. 2018.
- [20] T. Bian and Y. Deng, "A new evidential methodology of identifying influential nodes in complex networks," Chaos Solitons Fractals, vol. 103, pp. 101–110, Oct. 2017.