Journal of AI and Data Mining
Vol 8, No 4, 2020, 545-555. ***Applied Paper*** DOI: 10.22044/JADM.2020.9025.2040

DINGA: A Genetic-algorithm-based Method for Finding Important Nodes in Social Networks

H. Kamali[1], H. Rahmani[2*] and H. Shah-Hosseini[2]

1. Department of Computer Engineering, Faculty of Mechanic, Electrical and Computer, Science and Research Branch, Islamic Azad University, Tehran, Iran.
2. Department of Computer Engineering, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

Received 21 October 2019; Revised 29 December 2019; Accepted 03 April 2020
**Corresponding author: h_rahmani@iust.ac.ir (H. Rahmani).*

Abstract
Nowadays, a significant amount of studies are devoted to the discovery of important nodes in graph data. Social networks, as graph data, have attracted much attention. There are various purposes for discovering the important nodes in social networks such as finding the leaders in them, i.e. the users who play an important role in promoting advertising, etc. Different criteria have been proposed in discovering the important nodes in graph data. Measuring a node's importance by a single criterion may be inefficient due to the variety in the graph structures. Recently, a combination of criteria has been used in the discovery of the important nodes. In this paper, we propose a system for the Discovery of Important Nodes in social networks using Genetic Algorithms (DINGA). In our proposed system, the important nodes in social networks are discovered by employing a combination of eight informative criteria and their intelligent weighting. We compare our results with a manually weighted method that uses random weightings for each criterion in four real networks. Our proposed method shows an average of 22% improvement in the accuracy of discovery of important nodes.

Keywords: *Social Networks, Important Nodes, Genetic Algorithm, Graph Mining, Graph Data.*

1. Introduction

Nowadays, discovering the important nodes in graph data is of great interest. The relationship among the entities of many domains has been modeled as graph data [1, 2]. Discovering the important nodes in graph domains has been employed for various purposes [3-8]. For example, in the protein-protein interaction (PPI) networks, in which each protein represents a node and each physical interaction is an edge, the cancer-related proteins are considered as the important nodes. The discovery of cancer-related proteins can be effective in cancer treatment, in preventing its progression, and in improving the patient's overall condition [5, 9]. In the terroristic network, in which each node represents a person and the communications between them are modeled as edges, the leaders have important roles. Finding the main leaders in these networks can be useful in predicting, recognizing, and analyzing the occurrence of terrorist attacks [10, 11].

Among graph-represented domains, social networks with millions of users can provide a lot of useful information about human interactions [12]. Each user in these networks is considered as a node and each relationship among the users is considered as an edge [13-15]. There are various relationships in social networks such as kin relationships, workplace relationships, and friendships. The discovery of important nodes in these domains is very useful. For example, in the social networks that are used for advertisement purposes, we are looking for the most influential people. The information transfer speed increases significantly among the influential people. Therefore, by considering these people as the starting nodes to advertise, we expect that the message propagation is done in a wide range in the network. Hence, we select a small fraction of

the users as targets for advertisement. Then they will inform a large portion of users in the network about the advertised products [16, 17].

So far, the researchers have proposed many methods for discovering the important nodes in the graph-related domains [18]. The previous methods have been categorized into two main categories.

The first category of methods apply each criterion of centrality such as betweenness [9, 19] and entropy [20-23], individually. Due to the diversity of the graph structures, selecting the best criteria in advance is a challenging task. The second category of methods, in contrast to the first one, proposes that a weighted combination of criteria can be used for discovering the important nodes in graph data. The weighted vector is determined manually [24-28]. Thus the result of using a combination of criteria strongly depends on the criterion weight vector determined manually by the domain expert.

In this work, we address the challenges and drawbacks of the previous methods and propose a new method called DINGA. It can discover automatically the best weight vector of criteria for input graphs with varied structures, and accordingly, resolving the challenges of the first and second categories simultaneously. DINGA employs a weighted combination of eight criteria: Degree Centrality (DC) [9], Sub-graph Centrality (SC) [29], Eigenvector Centrality (EC) [30], Network Centrality (NC) [31], Information Centrality (IC) [32], Local Average Connectivity(LAC) [32], Betweenness Centrality (BC) [9,19] and Closeness Centrality (CC) [9]. The combination of these criteria is weighted automatically using the genetic algorithm. The main contributions of this paper are as follows:

- DINGA combines the varied ranking procedures automatically and proposes a better one.

- DINGA is a very general framework (with respect to centrality measure) for discovering the important nodes. Whenever the new centrality measure is discovered, DINGA could consider that as a new gene in the chromosome. If calculating one centrality measure is hard or impossible in some cases, DINGA could simply drop the corresponding gene in the GA process. As a conclusion, DINGA is very general, and the prediction power of DINGA is not dependent on the 8 mentioned criteria.

Section 2 discusses the previous works in the field of important nodes of social networks. In Section 3, we describe the proposed method thoroughly. Section 4 discusses the result of the experiments. In Section 4, we present the results using DINGA and we compare ours with the results of randomized weighting as well. Section 5 discusses the conclusion and future work.

2. Related Work

We categorized the previous work on the discovery of important nodes in social networks in two main categories. The first category uses varied centrality criteria individually. The second category combines various criteria for discovering the important nodes.

The first category of methods has proposed a variety of criteria for discovering the important nodes in graph data since the 1950s [33]. Degree Centrality (DC) [9] is a simple and efficient criterion but it neglects the global structure of the network. Wang *et al*. [34] have proposed the degree of nodes and the degree of their neighborhoods as a new criterion for discovering the important nodes in graph data. Chen *et al*. [35] have proposed a semi-local centrality criterion as a new criterion. Newman [36] has applied the criteria such as closeness and betweenness for discovering the important nodes in graph data. Closeness Centrality (CC) [9] and Betweenness Centrality (BC) [9, 19] result in a low computational complexity, although both of these criteria are not efficient in large-scale networks. Getoor *et al*. [37] have utilized the node clustering approach for discovering the important nodes. Kaur *et al*. [30] have applied an eigenvector for discovering the important nodes. In 2005, Shetty *et al*. [21] employed the entropy criterion of each node in the Enron e-mail dataset for this purpose. In 2017, Bashiri *et al*. [23] tried to improve the results of the Shetty's article, using the measurement of the entropy of each node.

The second category of methods was started in 2010. In this year, Hu *et al*. [24] proposed a multi-criteria ranking system for discovering the important nodes using five intuition-based rules. In 2013, Yajing *et al*. [25] used the Analytic Hierarchical Process (AHP) method to calculate the weights of a combination of metrics. The AHP method was presented by Saaty in 1970 [38]. Also in 2017, Bian *et al*. [26] employed a combination of metrics using the AHP method. In 2013, Yu *et al*. [27] used a combination of metrics that focused on the structural information of the network. They used the Multi-Attribute Decision-Making method (MADM) to find the weights of a

combination of criteria. In 2014, Dave *et al*. [28] utilized the TOPSIS method. The TOPSIS method is a MADM. This method was proposed by Wang and Yun in 1981. In order to improve the approach presented by Yu *et al*. [27], in 2016, Yang *et al*. [29] considered a combination of five metrics to discover the importance of nodes in the network. They used TOPSIS with AHP to find the combination of metrics' weight. HyperBall accesses the graph in a semi-streaming fashion and computes the distance distribution and approximates all geometric (i.e. distance-based) centralities. One of its main characteristics is the small amount of core memory usage [39, 40]. De Meo *et al*. [41, 42] have proposed a novel centrality metric called the potential gain, which quantifies the easiness at which a target node can be reached. The potential gain tries to combine the popularity of a node in G with its similarity to all other nodes.

Weskia *et al*. [43] and Roy *et al*. [44] have used evolutionary algorithms to discover the most influential nodes in social networks. They started with the initial important seed nodes, and then they applied a propagation model to expand the influences of seed nodes on the whole network.

The main drawbacks of the first and second categories of the previous methods are selecting and applying the best criteria and determining the weight vector manually, respectively. The manual building of the weight vector is time-consuming and error-prone. In the next section, we propose a new method to overcome these challenges.

3. Proposed Method

Considering the main drawbacks and challenges of the previous methods, we propose the DINGA system. This system takes the graph of a social network as the input, and then returns the important nodes in graph data as the output. An overview of the proposed method is presented in figure 1.

We discuss each main block of DINGA shown in figure 1 in the following sub-sections.

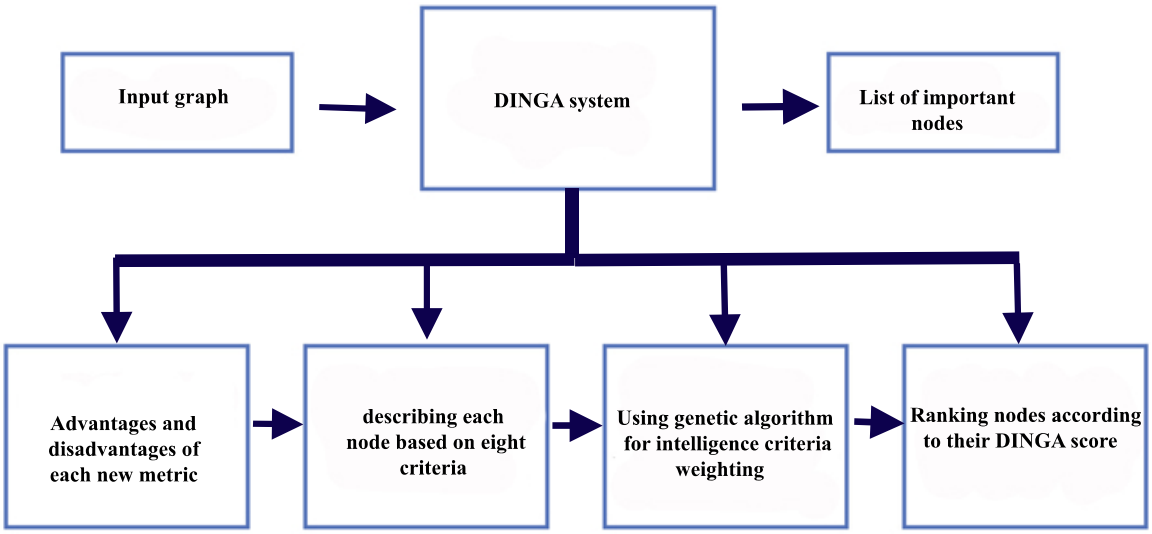

Figure 1. The proposed DINGA method to discover the important nodes in social networks. The DINGA system returns the list of important people as the output by receiving the social network graph as the input.

3.1. Input Graph

We represent the social network as a graph $G<V,E>$, in which each member $v \in V$ is considered as a node and each edge $e_{u,v} \in E$ indicates a direct communication between the two nodes $u \in U$ and $v \in V$.

3.2. DINGA

In this section, we discuss the four main steps of the DINGA system.

3.2.1. Advantages and Disadvantages of Each Criterion

In this section, we discuss the four main steps of the DINGA system. Graph criteria can be divided into two categories of local and global. After reviewing the advantages and disadvantages of each criterion, at the end, we consider the following eight criteria for the graph nodes: Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC), Sub-graph Centrality (SC), Network Centrality (NC), Eigenvector Centrality (EC), Local Average Connectivity (LAC), and Information Centrality

(IC). Finally, we compute the values of these criteria for all the nodes in the graph.

3.2.2. Description of Each Node based on 8 Criteria

In this section, we describe each node of the graph based on the 8 selected criteria presented in 3.2.1. Matrix CM is built for this purpose. Each cell $CM[i, j]$ indicates the value of criterion $j(j \leq 8)$ for node $i(i \leq |v|)$.

3.2.3. Weighting Criteria in Combination

According to the input graph, different criteria can have a varied influence on discovering the important nodes. In this section, we use the genetic algorithm to determine the weight of each criterion automatically. Figure 2 shows the main steps of the genetic algorithm.

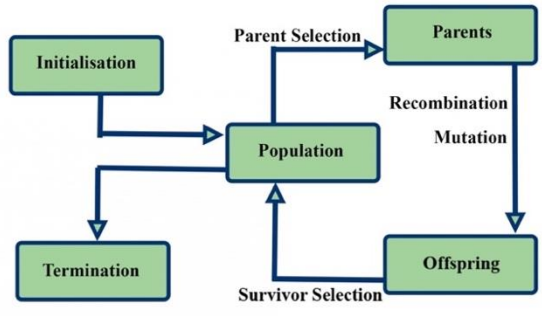

Figure 2. Genetic algorithm.

Each member of the population P_i is represented by 8-dimensional chromosome C_i, in which $C_{(i,j)}$ shows the weight of criterion j [45, 46]. Each chromosome is defined as:

$$C_i = \{p(i,1), p(i,2), p(i,3), .., P(i,m)\} \quad (1)$$

$$\sum_{m=1}^{8} p(i,m) = 1$$

We use accuracy for the fitness function of each chromosome.

In a genetic algorithm, a set of chromosomes is produced randomly. These chromosomes are known as the initial population [47]. In order to generate the new population, first, we use a tournament selection algorithm [46] to select two parents. Secondly, we use blend cross-over [46] and uniform mutation [46] to generate new offspring. At the end, the chromosomes of the individual with the best fitness value are picked up from the population to determine the weight of the criteria.

3.2.4. Ranking Nodes According to Their DINGA's Score

After training the weights of the 8 criteria, (2) is used to calculate the importance of each node in the graph.

$$important\text{-}score(i) = \sum_{j=1}^{8} cm(i,j) w_j \quad (2)$$

w_j is the result of a genetic algorithm and indicates the weight of criterion j in the discovery of the important nodes in graph data. The nodes are sorted in a descending order according to their important-scores.

4. Empirical Result

Several experiments were conducted for evaluating the performance of the proposed method.

4. Dataset

Several Experiments were conducted for evaluating the performance of the proposed method. In this section, we discuss, in detail, the four main real-world networks that are used to evaluate our proposed DINGA system. The basic statistical information of each network is presented in table 1.

Table 1. Basic statistical information of 4 real-world datasets.

Dataset	Number of nodes in the largest connected component	Number of important nodes in the graph	Clustering coefficient	Average degree	Label
Enron [21]	146	33	0.539	23.685	Labeled
Karate [48, 49]	34	6	0.088	1.941	Not labeled
AIDS (HIV) [50]	40	8	0	1.85	Not labeled
Protein-protein interaction networks [51]	8755	124	0.153	10.632	Labeled

4.1. Enron

The Enron E-mail dataset contains 150 users and also contains 517.431 internal E-mails. The E-mail represents the interactions between the employees of the company. In this dataset, the employees are considered as the nodes and are sent E-mails between them as edges [22]. The number of nodes in the largest connected component of the graph is 146 nodes. In order to implement the proposed method in the training phase of the genetic algorithm, 33 members of the company are considered to be the important nodes with positions of the president, vice president, chief operating officer, CEO, and government relative executive [52,53].

4.2. Zachary's Karate Club

In this network, 34 members of a karate club are considered as the nodes, and their friendships are considered as the edges [49]. Figure 3 shows the graph of the karate club.

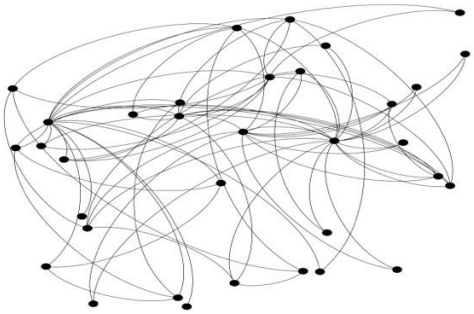

Figure 3. Zachary's Karate Club.

4.3. AIDS (HIV)

The dataset includes 40 AIDS patients who are in relation to each other. In the study of this dataset, the nodes represent the patients and the edges indicate the relationships between the patients. Figure 4 shows the graph of the AIDS dataset.

In the performance evaluation of the DINGA system, generally, two approaches are considered. In one approach, the dataset includes the importance label of each node. For this purpose, we use the Enron dataset, where the employee position is known for the training phase of the genetic algorithm. In the second approach, the nodes in graph data are not labeled (Karate [48], AIDS [50] datasets). For the latter case, the graph nodes are labeled according to Yang *et al*. [29].

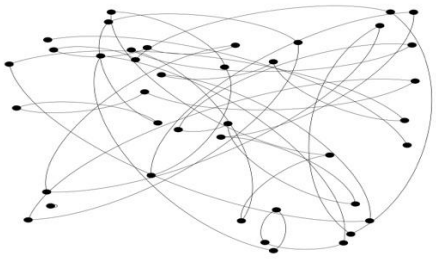

Figure 4. AIDS(HIV) network. Evaluation.

4.4. Essential Proteins in Protein-Protein Interaction Networks

We used the same PPI dataset as Rahmani *et al*. [51], which contains 45, 353 interactions among 9, 591 proteins. This dataset was generated by the union of three human PPI datasets HPRD [54], BIOGRID [55], and the dataset used by Radivojac *et al*. [56]. Additionally, we used the OGEE [57] database to evaluate the important proteins labeled by EMDIP. We collected a list of the essential genes from the OGEE database, in which all genes were grouped into three categories: essential, non-essential, and conditionally essential. According to OGEE, there are 124 essential, 4430 conditional, and 4910 non-essential proteins in our PPI network. We considered all the essential and conditional proteins as the important proteins.

5.1. DINGA System

We generated one hundred 8-dimensional vectors randomly as the initial population. The cross-over probability was 0.9 and the mutation probability was 0.14. We generated offspring until the deviation in the fitness function of the population became zero. The accuracies of the final population in Enron, Karate, and AIDS were 81%, 80%, 80%, respectively.

Figure 5 shows the average weighting of each criterion in the calculation of the important-score criterion by performing a 10-fold cross-validation on the DINGA system in the Enron dataset.

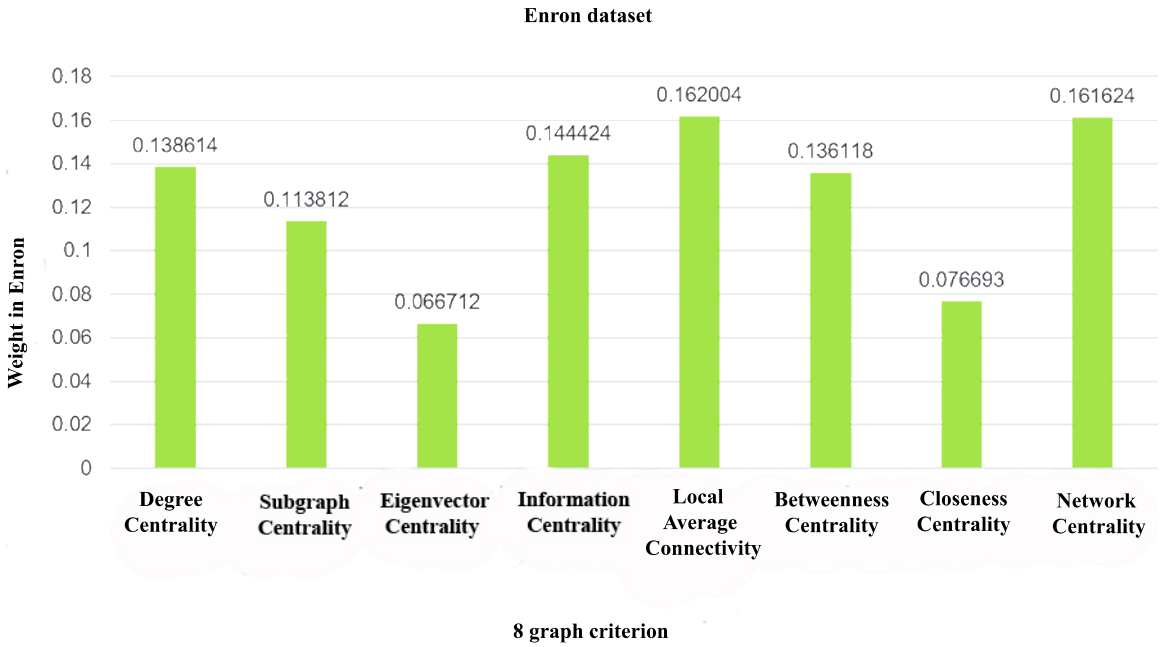

Figure 5. The average weighting of each criterion in the calculation of the important-score criterion by performing a 10-fold cross-validation on the DINGA system in the Enron dataset.

According to the Enron e-mail network structure, the LAC and network centralities are more important than the other criteria. Also the eigenvector and closeness centralities are less important than the other criteria for discovering the important nodes in the Enron network.

Figure 6 shows the average weighting of each criterion in the calculation of the important-score criterion by performing a 10-fold cross-validation on the DINGA system in the Karate dataset.

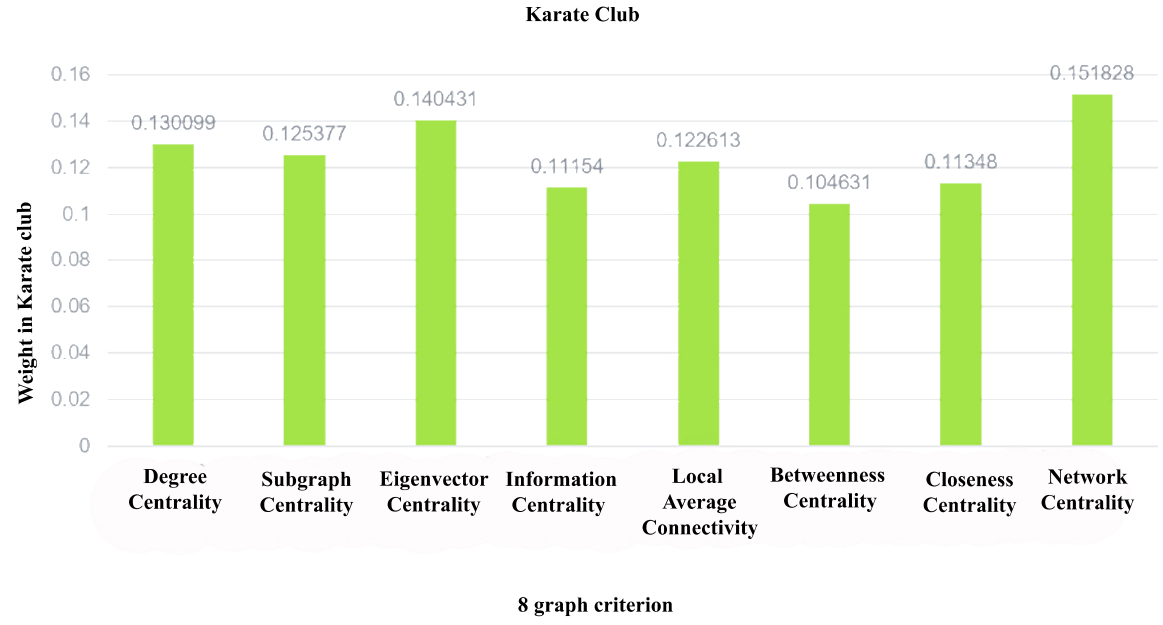

Figure 6. The average weighting of each criterion in the calculation of the important-score criterion by performing a 10-fold cross-validation on the DINGA system in the karate dataset.

According to the structure of the karate network, as shown in figure 6, the criteria of the network and eigenvector are more important than the other criteria. Also the betweenness and information centralities are less important in comparison with the other criteria for discovering the importance of graph nodes.

Figure 7 shows the average weighting of each criterion in the calculation of the important-score criterion by performing a 10-fold cross-validation on the DINGA system in the AIDS network.

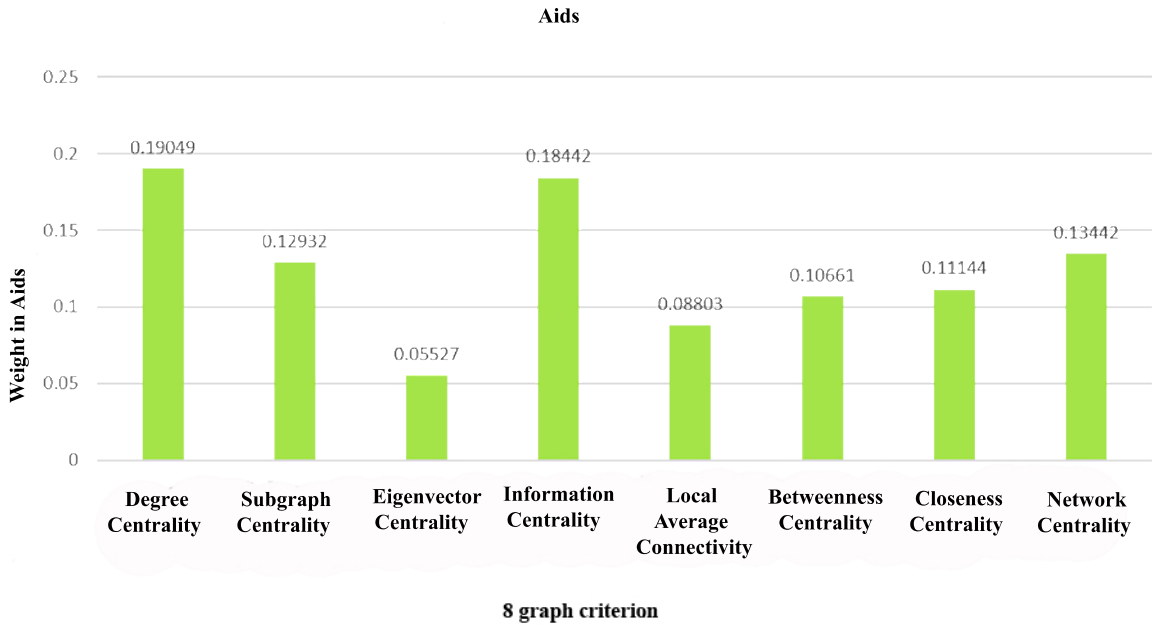

Figure 7. The average weighting of each criterion in the calculation of the important-score criterion by performing a 10-fold cross-validation on the DINGA system in the AIDS dataset.

According to the structure of the AIDS network, as shown in figure 7, the degree centrality and information centrality are more important than the other criteria. Also the eigenvector centrality and LAC have the lowest importance in discovering the important graph nodes in relation to the other criteria.

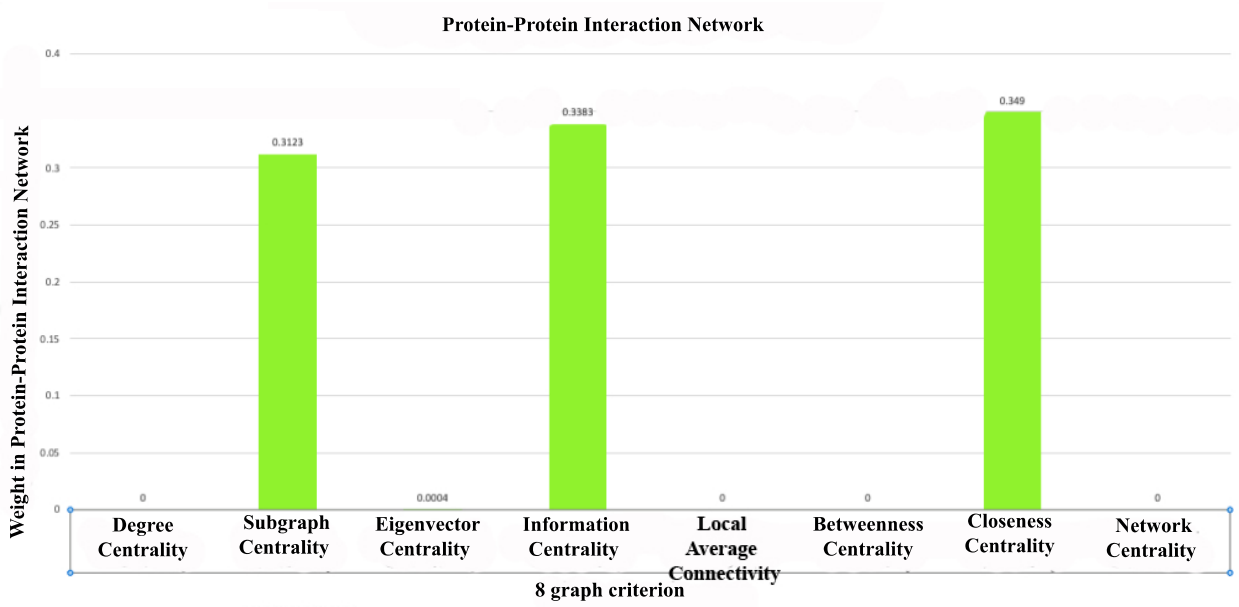

Figure 8. The average weighting of each criterion in the calculation of the important-score criterion by performing a 10-fold cross-validation on the DINGA system in the protein-protein interaction networks dataset.

According to the structure of the PPI network, as shown in figure 8, the closeness centrality and information centrality and sub-graph centrality are the more important features compared to the rest of our features.

Table 2 shows the most effective criterion for each of the considered datasets to discover the important nodes while using the DINGA system. As shown in table 2, the most discriminative criterion in each graph depends strongly on the graph structure.

Table 2. Discovery of the most efficient graph criteria in each graph by the DINGA system.

Dataset	Most important criterion discovered by DINGA
Enron [21]	Local Average Connectivity (LAC)
Karate [48,49]	Network Centrality (NC)
AIDS (HIV) [50]	Degree Centrality (DC)
Protein-protein interaction networks [51]	Closeness Centrality (CC)

The results of table 2 approve our initial hypothesis that most discriminative network centrality criteria is varied in different domains and networks and are strongly related to each specific network's structure.

5.2. Randomized Weighting

Instead of applying the genetic algorithm, we might consider determining the weight of each criterion completely at random [58]. For this purpose, we weighted the criteria for one hundred times arbitrarily. Then we computed the average accuracy of all for 100 times using randomized weighting. The accuracies of the average in Enron, Karate, AIDS, and protein-protein interaction networks are 60%, 51%, 62%, and 43%, respectively. Comparing the DINGA system's results with randomized weighting indicates 21%, 30%, 18%, and 17% accuracy improvement in Enron, Karate, AIDS, and protein-protein interaction networks datasets, respectively.

5.3. Conclusions and Future Work

Recently, discovering the important nodes in graph data has attracted much attention. In the previous research work, important nodes have been discovered using a single criterion. As a result of the inefficiency caused by the use of individual criteria in some cases, the researchers tend to employ a combination of criteria. In all the research work that exploit a combination of criteria, the weighting is done manually.

In this paper, we proposed a novel system called DINGA, which is a genetic-based algorithm that is capable of automatically discovering the weight of each criterion in the input graph. We evaluated DINGA in 4 real-world datasets, and we found that Local Average Connectivity, Network Centrality, Degree Centrality, and Closeness Centrality are discovered to be discriminative in the 4 datasets Enron, Karate, AIDS, and protein-protein interaction network, respectively. Additionally, our proposed method outperforms a randomized weighting method 22% with respect to accuracy. As it has been indicated in Section 4.4, our proposed method is easily able to discover the important nodes in the PPI network with 45000 edges among around 10000 nodes. In the case of a very large network, our method could become more scalable by either of the following ways:

1- Sampling: Instead of applying DINGA to the whole graph, we applied it to different samples of the data.
2- Graph Similarity: We indicated the most informative network features in 4 datasets. For a new large network, we could find the most similar network with smaller and reasonable sizes, and then we used the same network criteria as a most similar ones for discovering the important nodes in a new large network.

As a future work, we could extend DINGA in the following directions:

- In addition to the genetic algorithm (GA), we could use other machine learning methods (such as linear regression) to discover the best combination of weights plus some regularization term to discard the irrelevant metrics.

- We could extend the current GA method by considering more centrality criteria and applying varied fitness functions.

References

[1] Zhang, X., & Dong, D. (2008). Ways of identifying the opinion leaders in virtual communities. International Journal of Business and Management, nol. 3, no.7, pp. 21-27.

[2] Karimi Zandian, Z., & Keyvanpour, M. R. (2019). MEFUASN: a helpful method to extract features using

analyzing social network for fraud detection. Journal of AI and Data Mining, vol. 7, no. 2, pp. 213-224.

[3] Yada, K., Motoda, H., Washio, T., & Miyawaki, A. (2006). Consumer behavior analysis by graph mining technique. New Mathematics and Natural Computation, vol. 2, no. 01, pp. 59-68.

[4] Parthasarathy, S., Tatikonda, S., & Ucar, D. (2010). A survey of graph mining techniques for biological datasets. In Managing and mining graph data, Springer, Boston, MA, pp. 547-580.

[5] Rahmani, H., Blockeel, H., & Bender, A. (2016). Using a human drug network for generating novel hypotheses about drugs. Intelligent Data Analysis, vol. 20, no. 1, pp. 183-197.

[6] Klopman, G. (1984). Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. Journal of the American Chemical Society, vol. 106, no. 24, pp. 7315-7321.

[7] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv: 1009.6119.

[8] Xiaojun, L., Song, H., & Weikun, X. (2017). The Analysis of Logistics Influence of the Important Node Cities of Beijing-Tianjin-Hebei. International Journal of Business and Economics Research, vol. 6, no.5, pp. 88.

[9] Rahmani H, Blockeel H, Bender A. (2009). Predicting the functions of proteins in protein-protein interaction networks from global information. In: Machine Learning in Systems Biology, Ljubljana, Slovenia. pp. 82-97.

[10] Cook, D. J., Holder, L. B., & Ketkar, N. (2006). Unsupervised and supervised pattern learning in graph data. Mining Graph Data, pp. 159-180.

[11] Liu, J. G., Ren, Z. M., & Guo, Q. (2013). Ranking the spreading influence in complex networks. Physica A: Statistical Mechanics and its Applications, vol. 392, no. 18, pp. 4154-4159.

[12] Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. EPJ data science, vol. 4, no. 1, pp. 10.

[13] Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. Journal of computer-mediated communication, vol. 3, no. 1, JCMC313.

[14] Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods. University of California, Riverside. CA (Online book).

[15] Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications, Cambridge university press, vol. 8.

[16] Borgatti, S. P. (2006). Identifying sets of key players in a social network. Computational & Mathematical Organization Theory, vol. 12, no. 1, pp. 21-34.

[17] Yang WS, Dia JB, Cheng HC, Lin HT. (2006). Mining social networks for targeted advertising. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), Kauia, HI, USA, pp. 137a-137a.

[18] Kazienko, P., & Musial, K. (2007). On utilising social networks to discover representatives of human communities. IJIIDS, vol. 1, no. 3/4, pp. 293-310.

[19] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology, vol. 25, no. 2, pp. 163-177.

[20] Noble CC, Cook DJ. (2003). Graph-based anomaly detection. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (ACM), New York, NY, USA, pp. 631-636.

[21] Shetty, J., & Adibi, J. (2005). Discovering important nodes through graph entropy the case of enron email database. In Proceedings of the 3rd international workshop on Link discovery, ACM, pp. 74-81.

[22] Kajdanowicz, T., & Morzy, M. (2016). Using graph and vertex entropy to compare empirical graphs with theoretical graph models. Entropy, vol. 18, no. 9, pp. 320.

[23] Bashiri v, Rahmani H, Bashiri H. (2017). Finding Important Nodes in Social Networks. In: Third International Conference on Web Research (ICWR), Tehran, Iran.

[24] Hu, J., Han, Y., & Hu, J. (2010). Topological potential: modeling node importance with activity and local effect in complex networks. In 2010 Second International Conference on Computer Modeling and Simulation, IEEE, vol. 2, pp. 411-415.

[25] Xu, Y., Gao, Z., Xiao, B., Meng, F., & Lin, Z. (2013). Key nodes evaluation with multi-criteria in complex networks based on AHP analysis. In 2013 5th IEEE International Conference on Broadband Network & Multimedia Technology, IEEE, pp. 105-109.

[26] Bian T, Hu J, Deng Y. (2017). Identifying influential nodes in complex networks based on AHP. Physica A: Statistical Mechanics and its Applications, vol. 479, pp. 422-36.

[27] Yu, H., Liu, Z., & Li, Y. J. (2013). Key nodes in complex networks identified by multi-attribute decision-making method.

[28] Du, Y., Gao, C., Hu, Y., Mahadevan, S., & Deng, Y. (2014). A new method of identifying influential nodes in complex networks based on TOPSIS. Physica A: Statistical Mechanics and its Applications, vol. 399, pp. 57-69.

[29] Yang, Y., & Xie, G. (2016). Efficient identification of node importance in social

networks. Information Processing & Management, vol. 52, no. 5, pp. 911-922.

[30] Kaur, M., & Singh, S. (2016). Analyzing negative ties in social networks: A survey. Egyptian Informatics Journal, vol. 17, no. 1, pp. 21-43.

[31] Li, M., Wang, J., Wang, H., & Pan, Y. (2012). Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, pp. 1070-1080.

[32] Li, M., Wang, J., Chen, X., Wang, H., & Pan, Y. (2011). A local average connectivity-based method for identifying essential proteins from the network level. Computational biology and chemistry, vol. 35, no. 3, pp. 143-150.

[33] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 160-168.

[34] Wang, S., Du, Y., & Deng, Y. (2017). A new measure of identifying influential nodes: Efficiency centrality. Communications in Nonlinear Science and Numerical Simulation, vol. 47, pp. 151-163.

[35] Chen, D., Lu, L., Shang, M. S., Zhang, Y. C., & Zhou, T. (2012). Identifying influential nodes in complex networks. Physica a: Statistical mechanics and its applications, vol. 391, no. 4, pp. 1777-1787.

[36] Newman, M. E. (2000). Who is the best connected scientist? A study of scientific coauthorship networks, arXiv preprint cond-mat/0011144.

[37] Getoor, L., Segal, E., Taskar, B., & Koller, D. (2001). Probabilistic models of text and link structure for hypertext classification. In IJCAI workshop on text learning: beyond supervision, pp. 24-29.

[38] Emshoff, J. R., & Saaty, T. L. (1982). Applications of the analytic hierarchy process to long range planning processes. European Journal of Operational Research, vol. 10, no. 2, pp. 131-143.

[39] Boldi, P., Luongo, A., & Vigna, S. (2017). Rank monotonicity in centrality measures. Network Science, vol. 5, no. 4, 529-550.

[40] Boldi, P. (2015). Large-scale Network Analytics: Diffusion-based Computation of Distances and Geometric Centralities. In Proceedings of the 24th International Conference on World Wide Web, pp. 1313-1313.

[41] De Meo, P., Levene, M., & Provetti, A. (2019). Potential gain as a centrality measure. In IEEE/WIC/ACM International Conference on Web Intelligence, pp. 418-422.

[42] De Meo, P., Levene, M., Messina, F., & Provetti, A. (2019). A general centrality framework based on node navigability. IEEE Transactions on Knowledge and Data Engineering.

[43] Roy, M., & Pan, I. (2018). Most Influential Node Selection in Social Network using Genetic Algorithm. In 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE, pp. 214-220.

[44] Weskida, M., & Michalski, R. (2019). Finding influentials in social networks using evolutionary algorithm. Journal of Computational Science, vol. 31, pp. 77-85.

[45] Gen M, Cheng R. (2000). Genetic algorithms and engineering optimization, Hoboken, NJ, USA, John Wiley \& Sons.

[46] Eiben, A. E., & Smith, J. E. (2003). Introduction to evolutionary computing, Berlin, springer, vol. 53, pp. 18.

[47] Vertan, C., Vertan, C. I., & Buzuloiu, V. (1997, July). Reduced computation genetic algorithm for noise removal, In: 1997 Sixth International Conference on Image Processing and Its Applications, IET, vol. 1, pp. 313-316.

[48] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. Journal of anthropological research, vol. 33, no. 4, pp. 452-473.

[49] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, vol. 99, no. 12, pp. 7821-7826.

[50] Auerbach, D. M., Darrow, W. W., Jaffe, H. W., & Curran, J. W. (1984). Cluster of cases of the acquired immune deficiency syndrome: Patients linked by sexual contact. The American journal of medicine, vol. 76, no. 3, pp. 487-492.

[51] Rahmani, H., Weiss, G., Mendez-Lucio, O., & Bender, A. (2016). ARWAR: A network approach for predicting Adverse Drug Reactions. Computers in biology and medicine, vol. 68, 101-108.

[52] Creamer, G., Rowe, R., Hershkop, S., & Stolfo, S. J. (2007). Segmentation and automated social hierarchy detection through email network analysis. In International Workshop on Social Network Mining and Analysis, Springer, Berlin, Heidelberg, pp. 40-58.

[53] Gilbert, E. (2012). Phrases that signal workplace hierarchy. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM, pp. 1037-1046.

[54] Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., ... & Rashmi, B. P. (2004). Human protein reference database as a discovery resource for proteomics. Nucleic acids research, vol. 32, D497-D501.

[55] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a

general repository for interaction datasets. Nucleic acids research, vol. 34, D535-D539.

[56] Radivojac, P., Peng, K., Clark, W. T., Peters, B. J., Mohan, A., Boyle, S. M., & Mooney, S. D. (2008). An integrated approach to inferring gene–disease associations in humans. Proteins: Structure, Function, and Bioinformatics, vol. 72 no. 3, 1030-1037.

[57] Chen, W. H., Lu, G., Chen, X., Zhao, X. M., & Bork, P. (2016). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic acids research, gkw1013.

[58] Lu, Z. M., & Feng, Y. P. (2015). Critical Nodes and Links Evaluation with Multi-Criteria Based on Entropy-Weighted Method.

نشریه هوش مصنوعی و داده کاوی

DINGA: روشی مبتنی بر الگوریتم ژنتیک جهت شناسایی گره‌های مهم در شبکه‌های اجتماعی

هستی کمالی[1]، حسین رحمانی[2,*] و حامد شاه‌حسینی[1]

[1] دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، دانشکده مکانیک، برق و کامپیوتر، تهران، تهران، ایران.

[2] دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر، تهران، تهران، ایران.

ارسال ۲۰۱۹/۱۰/۲۱؛ بازنگری ۲۰۱۹/۱۲/۲۹؛ پذیرش ۲۰۱۹/۰۴/۰۳

چکیده:

امروزه با افزایش روزافزون کاربرد شبکه‌های اجتماعی، مسئله شناسایی گره‌های مهم در این شبکه‌ها، از اهمیت بالایی برخوردار است. اهداف متنوعی جهت شناسایی گره‌های مهم در شبکه‌های اجتماعی وجود دارد. از جمله این اهداف می‌توان به شناسایی رهبران در گروه‌ها و افراد تأثیرگذار در شبکه‌های اجتماعی، به منظور انتشار سریع اخبار، تبلیغات و غیره اشاره کرد. با توجه به وجود ساختار گرافی در این شبکه‌ها، تاکنون معیارهای متفاوتی با بهره‌گیری از علم گراف‌کاوی، جهت شناسایی گره‌های مهم معرفی شده است. بدیهی است که با توجه به ساختارهای متنوع گرافی، شناسایی گره‌های مهم در داده‌های گرافی، با به کارگیری هر معیار به تنهایی و بدون در نظر گرفتن ساختار گراف، در بسیاری از موارد، ناکافی و ناکارآمد است. در راستای حل این مسئله، در تحقیقات اخیر به کارگیری چند معیار در شناسایی گره‌های مهم گرافی طرح شده است. در این مقاله سیستمی هوشمند با عنوان DINGA به منظور شناسایی گره‌های مهم در داده‌های گرافی مربوط به شبکه‌های اجتماعی پیشنهاد شده است. این سیستم با به کارگیری الگوریتم ژنیتیک و علوم گراف‌کاوی در گراف با ساختاری نامعلوم، به وزن‌دهی هوشمند ۸ معیار گرافی و شناسایی گره‌های مهم متناسب با ساختار گراف می‌پردازد. نتایج حاصل از به کارگیری سیستم DINGA در چهار شبکه واقعی، در مقایسه با نتایج حاصل از به کارگیری وزن‌دهی تصادفی، به طور میانگین، بیانگر ۲۲ درصد بهبود دقت در شناسایی گره‌های مهم در شبکه‌های اجتماعی است.

کلمات کلیدی: شبکه‌های اجتماعی، گره‌های مهم، الگوریتم ژنتیک، گراف‌کاوی، داده‌های گرافی.