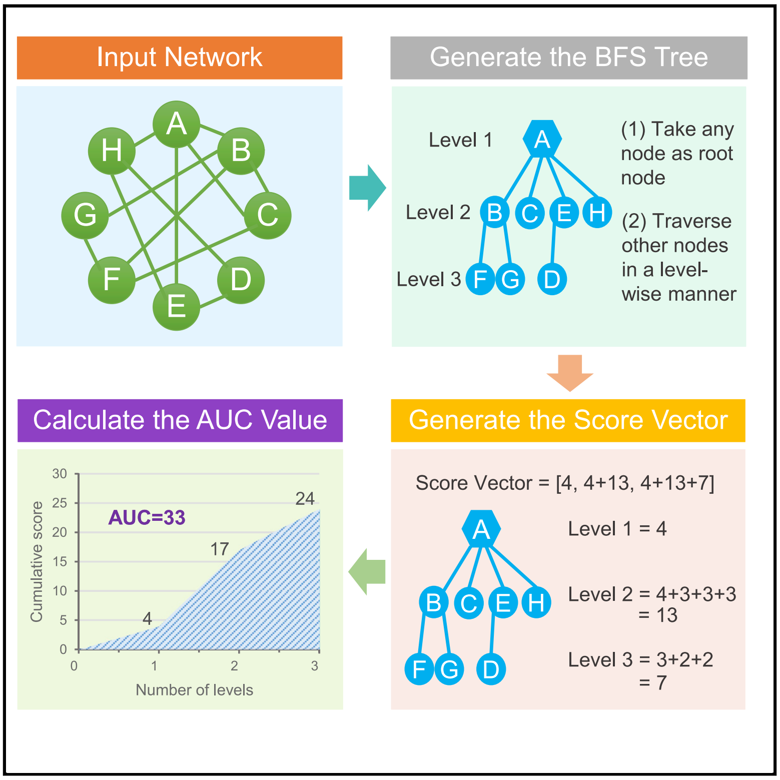
Patterns

Article

A graph-traversal approach to identify influential nodes in a network

Graphical abstract



Highlights

- We propose an influential node detection method, TARank, in a graph-traversal framework
- We evaluate the influence of each node by constructing a breadth-first search tree
- TARank is capable of enhancing existing centrality measures
- TARank can yield new, yet effective, centrality measures as well

Authors

Yan Liu, Xiaoqi Wei, Wenfang Chen, Lianyu Hu, Zengyou He

Correspondence

zyhe@dlut.edu.cn

In brief

To identify influential nodes, many methods have been proposed during the past decades. However, no single method can always achieve the best performance across different networks, and it is still necessary to develop new methods based on some principles that remain unexplored. Here we present a novel method, TARank, which tackles the influential node identification issue from a graph-traversal perspective. TARank is a general framework that can both unify some existing methods and create many new variants.

Liu et al., 2021, Patterns *2*, 100321
September 10, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.patter.2021.100321

CellPress

Article

A graph-traversal approach to identify influential nodes in a network

Yan Liu,[1] Xiaoqi Wei,[1] Wenfang Chen,[1] Lianyu Hu,[1] and Zengyou He[1,2,3,*]
[1]School of Software, Dalian University of Technology, Dalian 116024, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116024, China
[3]Lead contact
*Correspondence: zyhe@dlut.edu.cn

https://doi.org/10.1016/j.patter.2021.100321

THE BIGGER PICTURE The discovery of influential nodes is a fundamental research issue in network science. To quantify the influence of each node in a network, various methods have been presented in the literature. To the best of our knowledge, no previous research efforts address the influential node identification problem from a graph-traversal perspective. To fulfill this void, we propose the TARank method that integrates the information collected from the breadth-first search tree to identify influential nodes. The formulation under the graph-traversal framework opens the door to a fundamentally new type of method of influential node identification. In the future, more effective recognition methods can be expected to be constructed based on this general framework. Since empirical studies have validated the effectiveness of TARank, it would be plausible to employ this method in different applications to reveal new findings.

1 2 3 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Influential node identification plays a significant role in understanding network structure and functions. Here we propose a general method for detecting influential nodes in a graph-traversal framework. We evaluate the influence of each node by constructing a breadth-first search (BFS) tree in which the target node is the root node. From the BFS tree, we generate a curve in which the x axis is the level number and the y axis is the cumulative scores of all nodes visited so far. We use the area under the curve value as the final influence score of the target node. Experimental results on various networks across different domains demonstrate that our method can be significantly superior to widely used centrality measures on the task of influential node detection.

INTRODUCTION

The network (graph) is a common type of data structure that offers a holistic and top-down view to make sense of various interactive systems, including social systems, biological systems, traffic systems, communication systems, and so on,[1–3] that are highly affected by a small portion of influential nodes, also called influential spreaders.[4] Such nodes play a critical role and can significantly enrich our understanding of the above systems. For example, being able to effectively and properly detect influential nodes allows us to control the spread of epidemics,[5,6] design a valid marketing plan,[7,8] prevent the power grid from failing,[9,10] predict future traffic flow,[11,12] and identify essential proteins.[13,14]

Since finding the influential node is a general network analysis issue, many metrics have been proposed to evaluate the importance of each node in a network from different perspectives. For example, degree centrality (DC)[15] and the H-index[16] are typical neighborhood-based methods for centrality evaluation. Path-based methods such as closeness centrality (CC),[17] load centrality (LC),[18] betweenness centrality (BC),[19] and information centrality (IC) take global topological features into consideration.[20] Eigenvector centrality (EC)[21] and PageRank (PR)[22] are typical methods that evaluate the node centrality in an iterative refinement manner. Existing centrality measures and their applications in different fields have been summarized in several reviews.[23–27]

Despite decades of research on developing centrality measures for identifying influential nodes, there is still no consensus on the best centrality measure across different types of networks in various domains. This is mainly because each type of centrality measure has its own advantages and drawbacks,[28] making it difficult even to offer a universal solution to displace the most simple DC measure. Therefore, one new centrality evaluation method

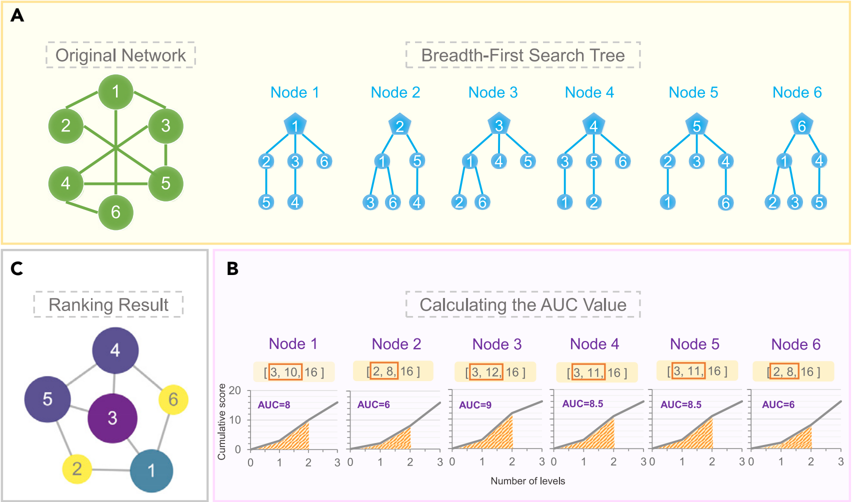


Figure 1. The main workflow of TARank

(A) Starting with each node as the root node, the graph is traversed in a level-order to construct a breadth-first search (BFS) tree. Each node in the tree is associated with an initial score that is generated from an arbitrary centrality measure.

(B) At each level of the tree, a cumulative score is obtained by adding up the scores of all nodes traversed by the BFS procedure so far. As a result, a score vector for each BFS tree can be derived. The first k values in the score vector ($k = 2$ in this figure) are used to plot a curve, and the AUC value is used as the centrality score in TARank for quantifying the influence of each node.

(C) All nodes in the network can be sorted according to the AUC score. In this figure, the size of each node is proportional to its corresponding AUC score.

with the following desired characteristics should be developed: (1) it is a general formulation such that different types of existing centrality measures can be incorporated, (2) it solves the node centrality issue based on some principles that remain unexplored, and (3) it has good performance on different types of networks.

In this paper, we formulate the problem of influential node recognition in a graph-traversal framework. For each node in the network, we build a breadth-first search (BFS) tree by traversing the graph in a level-wise manner, in which the target node is the root node (Figure 1A). Intuitively, the BFS tree for an influential node will have the following distinct features: (1) There will be many nodes on the top levels of the tree. The nodes that appear on the top levels of the BFS tree fall into the local neighborhood of the root node. The number of nodes in the local neighborhood has been widely used as a criterion for influential node recognition in many classical evaluation methods, such as DC. (2) Each node on the top levels of the tree is expected to be an influential one as well. The assumption that one influential node is expected to have many highly influential neighbors has been adopted in many popular methods, such as HITs,[29] PR,[22] and TwitterRank.[30] (3) The height of the tree should be low. The height of a BFS tree on an undirected network corresponds to the largest "shortest path length" between the root node and the remaining nodes. Hence, the lower the height is, the more centric the root node is. Based on the above observations, we derive a score vector from each BFS tree in which each element is the sum of the influence scores of all nodes above a certain level. The influence score of each node in the tree can be obtained from any existing centrality measure. According to the first k entries in the score vector, we can plot a curve in which the x axis is the level number and the y axis is the corresponding score (Figure 1B). The area under the curve (AUC) can be used as the overall centrality score for quantifying the node importance. Such a general method for influential node identification is named TARank (tree- and AUC-value-based rank).

The proposed method addresses the influential node identification issue from a graph-traversal perspective, which is totally different from existing methods in the literature. Meanwhile, it provides a general framework for influential node quantification, in which any existing centrality measures can be utilized for generating the final importance score. We evaluate our method on 53 real networks, and experimental results demonstrate that our method significantly outperforms widely used centrality evaluation methods.

RESULTS

An overview of TARank

The TARank method consists of the following steps (see Figure 1). First, each node in the input network is regarded as the root node, and a corresponding BFS tree is generated by traversing the graph in a level-order. Each node in the tree has an initial centrality score, which can be obtained from any centrality measure, such as DC (Figure 1A). Second, a cumulative score vector of length h is constructed from each BFS tree, where h is the largest level number (the level number of the root node is 1). The i-th element in the score vector is calculated as the sum of scores of all nodes whose levels are no larger than i. Based on a user-specified parameter, $k(1 \leq k \leq h)$, the first k cumulative scores in the score vector are used to generate a curve in which the x coordinate ranges from 0 to k and the y coordinate is the corresponding cumulative score. The AUC value is used as the final centrality score for assessing the node influence (Figure 1B). Finally, TARank detects influential nodes according to the AUC scores (Figure 1C).

Performance evaluation

We conducted experiments on 53 real networks, including transportation networks, technological networks, social networks, informational networks, economic networks, and biological networks. The basic features of these networks can be found in Table S1. To evaluate the performance of TARank, we employ the Kendall's tau correlation coefficient τ[31] as the performance indicator to measure the statistical relationship between the AUC score in TARank and the node influence obtained by the well-known susceptible-infected-recovered (SIR) spreading model.[32] Under the SIR spreading model, an infected node can transmit the disease with an infection probability (transmission probability) (see details under experimental procedures). The influence of each node is quantified by averaging the number of recovered nodes after 1,000 independent runs of the SIR spreading model.

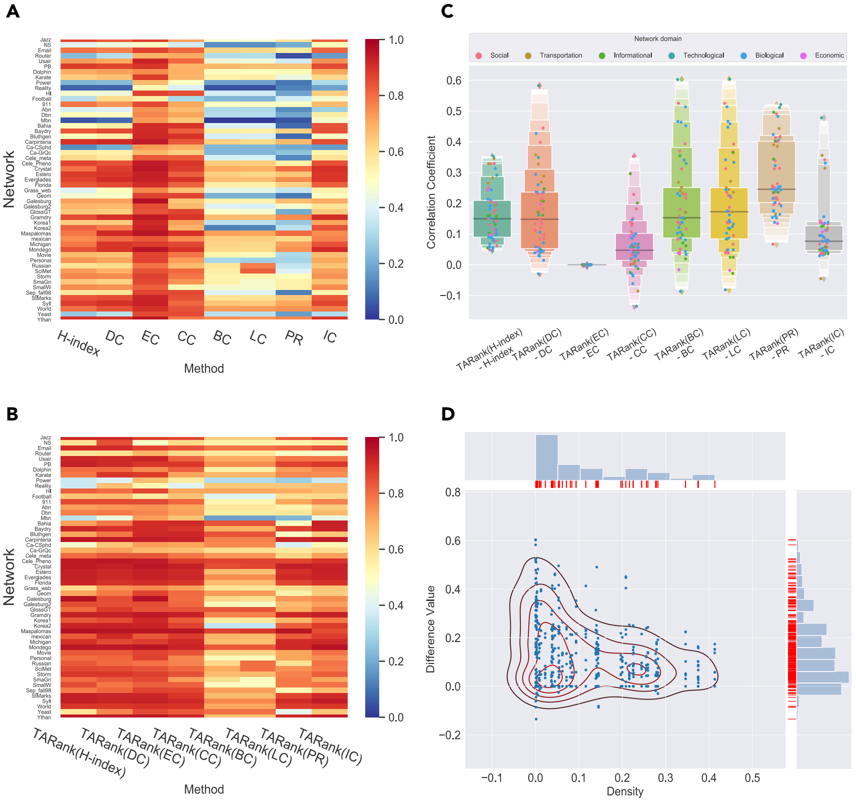


Figure 2. Comparisons of the Kendall's tau correlation coefficient between SIR node influences and centrality scores on real networks

(A) The correlation coefficients between eight centrality measures and node influences are shown in different colors, where the y axis represents the network name and the x axis shows the centrality evaluation method.

(B) The correlation coefficients between TARank variants and node influences are presented, where TARank (*x*) denotes that the centrality evaluation method *x* is used for generating the initial node importance score in TARank. In both (A) and (B), larger correlation coefficients are represented by dark red squares and smaller correlation coefficients are colored with dark blue squares.

(C) The difference values of correlation coefficients between each TARank variant and its counterpart are exhibited. There are eight boxes and 53 points in each box, where the points with different colors emphasize different kinds of networks.

(D) The joint plot composed of a scatterplot and a kernel density estimation plot shows the correlation between the correlation difference value and the network density.

TARank is compared with eight state-of-the-art methods: H-index, DC, EC, CC, BC, LC, PR, and IC. Since TARank is a generic framework in which any centrality measure can be utilized to generate an initial score for each node, we chose the above eight centrality measures for calculating the final AUC value as well. As a result, there were eight variants of TARank, which are denoted as TARank (H-index), TARank (DC), TARank (EC), TARank (CC), TARank (BC), TARank (LC), TARank (PR), and TARank (IC), respectively.

Figure 2 shows the comparison between the variants of TARank and the eight centrality evaluation methods in terms of the correlation coefficient. In Figures 2A and 2B, different correlation coefficients are labeled in different colors. Compared with its counterpart in Figure 2A, each variant of TARank in Figure 2B exhibits higher correlation with the node influence derived from the SIR model. That is, the overall performance of TARank is better than those competing centrality evaluation methods. To quantitatively describe the performance improvement of TARank against each competitor, the difference value of the correlation coefficient on each network is calculated. The distribution of correlation difference values on all 53 networks for each centrality evaluation method is plotted in Figure 2C. Obviously, it can be observed that most difference values are larger than 0, which means that TARank can yield a performance gain in most cases. There are a small portion of networks on which TARank does not work well. On one hand, some biological networks are noisy, in that many spurious edges are present but some true edges are missing.[33,34] On the other hand, some sparse networks may hinder spreading, which even leads to hostile influence prediction results, as shown in Figure 2D. The suspicion that the network density can affect the performance of influential node recognition has been raised in a previous study.[35] However, in our context, the association between the network density and the correlation difference is not statistically significant. This is because the correlation difference measures only if TARank can yield performance improvement over each competing centrality evaluation method. For instance, when EC is utilized for generating the initial node score, the correlation difference is almost zero on most networks (see Figure 2C). Although the correlation difference is negligible, the final performance of TARank (EC) is quite good, as shown in Figures 2B and 3.

Then, we compare the overall performance of eight TARank variants using a boxplot of correlation coefficients on the 53 networks. As shown in Figure 3, TARank (H-index), TARank (DC), and TARank (EC) can generally achieve better performance than the other TARank variants with respect to the average correlation coefficient. Among TARank (H-index), TARank (DC), and TARank (EC), the TARank (DC) method is more preferable than the other two methods in regard to the time complexity (see Table S2). Hence, based on the trade-off between effectiveness and efficiency, it seems that TARank (DC) is a good candidate to be employed in practice for influential node identification.

To further check whether the effectiveness of TARank can be significantly affected by the infection probability β, we ran the spreading process with different β values on 53 real networks. In Figure 4, we compare the correlation coefficients of TARank and eight centrality measures under different infection probabilities. We can see from Figure 4 that the performance of TARank is consistently superior to the eight centrality measures when the infection probability is varied from $1.0\beta_c$ to

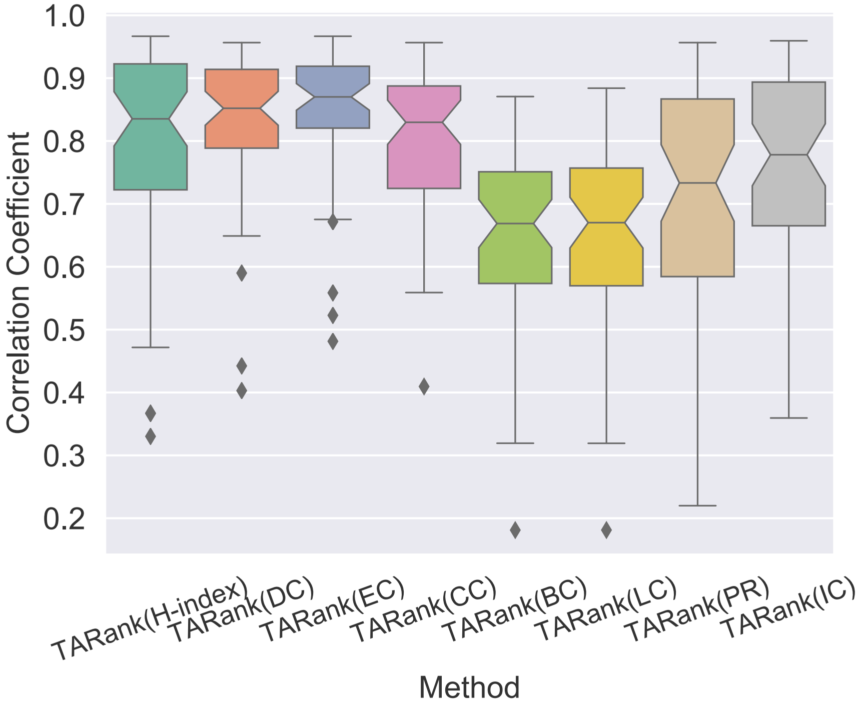


Figure 3. The comparison of eight TARank variants in terms of the distribution of correlation coefficients over 53 networks

Here a boxplot is used to depict the distributions: minimum (lower line), lower quartile (lower edge of the box), median (center line), upper quartile (upper edge of the box), and maximum (upper line). Meanwhile, outliers are also plotted as isolated points.

$2.0\beta_c$, in which β_c is the epidemic threshold. In addition, the t test for paired samples is adopted to test if the difference in correlation coefficient between TARank and each centrality measure is statistically significant. We conducted eight significance tests and the results are shown in Figure 4. In each test, one competing centrality measure is compared with the corresponding TARank variant. We can see that the p values are less than the significance level 0.05 in most cases. To empower better generalizations, we tested the difference between two groups in which one group consisted of all correlation coefficients generated by TARank and another group was derived from eight centrality measures. The result is shown in the last graph in Figure 4, in which the p values under the different transmission probabilities are far less than the significance level 0.05. The hypothesis testing results indicate that the performance of TARank is significantly better than that of contrastive centrality evaluation algorithms under the different infection probabilities.

In addition, we tested the performance of TARank under another acclaimed spreading model: the susceptible-infected-susceptible (SIS) spreading model.[36] The influence score of each node can be defined as the probability that a node will be infected. The results (Figure S1) suggest that TARank is still effective under the SIS spreading model. To investigate the performance of TARank when the “ground-truth” influential nodes are available, we followed the strategy proposed by Salavaty et al.[28] to compare our method with those competing methods based on the SIRIR model in Figure S2. It indicates that most TARank variants can achieve better performance than their counterparts and are comparable to advanced modes such as integrated value of influence (IVI).[28]

In previous experiments, the parameter *k* was fixed to be 2. To test if TARank is sensitive to this parameter, we varied *k* from 2 to 4 to check its performance fluctuations on 53 networks in Figures S3 and S4. The experimental results show that TARank can achieve better performance than existing methods in most cases under different *k* values.

DISCUSSION

In this paper, we did not distinguish the concepts among the influential node, the spreader node, and the hub node. Kitsak et al.[4] have discussed the difference between the hub node and the spreader node, where the hub node will directly link to more nodes but the spreader node tends to have more interpersonal influence on others. In addition, it is believed that the influential node should have both many direct neighbors and high spreading potential.[28]

In the experiments, eight existing centrality measures are adopted for generating the initial node score. Thanks to the versatility of our method, other existing centrality measures (e.g., IVI,[28] Local H-index[37]) can be employed for this purpose as well. We have empirically tested the feasibility of using these recently developed centrality measures in Figure S2. Meanwhile, we will investigate the potential of our framework for identifying influential nodes from weighted networks[38] and time-varying networks.[39]

When the initial node score is generated by existing centrality metrics, the final node score derived from TARank can be regarded as the refinement of current centrality measures. From this perspective, our method is related to those ensemble methods such as IVI.[28] Both TARank and IVI can utilize the existing centrality methods to identify the influential nodes. But in contrast, IVI focuses on integrating several centrality methods, while TARank can provide the refinement for an existing measure.

Conclusion

The scale-free property[40,41] suggests that different nodes may play totally different roles in a network. Hence, many research efforts have been devoted to finding influential nodes in networks. TARank provides a general framework for identifying influential nodes by use of graph traversal. As indicated by the empirical studies, TARank can incorporate different centrality measures and offer significant performance improvements on various types of networks.

We believe that the framework introduced here will serve as a foundational graph theoretic tool for identifying influential nodes in network science. Meanwhile, it can be applied to real

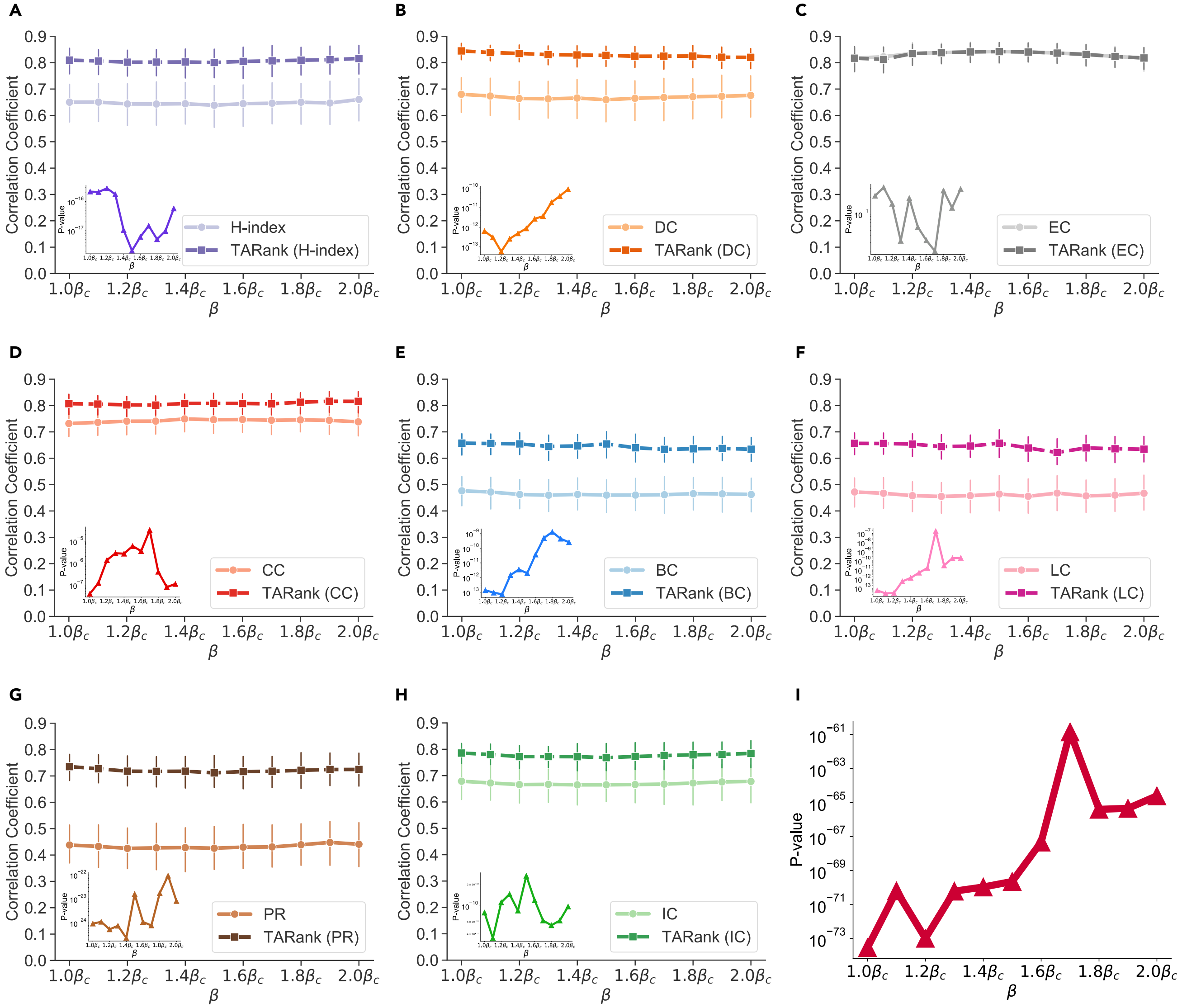


Figure 4. The effects of the infection probability on TARank and eight centrality evaluation methods on 53 networks

In the first eight graphs, the light solid line represents one existing centrality evaluation method and the dark dashed line denotes the corresponding TARank variant, where each bar that represents the 99% confidence interval is the graphical representation of the variability on different networks. The p values were calculated for quantifying the significance of paired difference between each centrality evaluation method and its TARank variant. The last graph shows p values under the different infection probability β, where the paired samples are collected from all eight tested centrality measures.

applications in different fields, such as essential protein identification in proteomics and critical node recognition in a communication network.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this work is Zengyou He at zyhe@dlut.edu.cn.

Materials availability

This study did not generate new unique reagents.

Data and code availability

All the networks analyzed in the experiment and the source code of TARank are publicly available at https://github.com/LiuYan-Peggy/TARank.

Method details

Data preparation

We used 53 real networks to test the effectiveness of our method. These networks can be divided into six categories: social networks, transportation networks, informational networks, technological networks, biological networks, and economic networks. All networks can be found at Pajek Datasets (http://vlado.fmf.uni-lj.si/pub/networks/data/) and Network Data Repository.[42] A brief description of each network is provided in Table S1.

The TARank framework

Given a network with a set of N nodes $V = \{v_1, v_2, \cdots, v_N\}$ and a set of M edges $E = \{e_1, e_2, \cdots, e_M\}$, we can generate a set of initial centrality scores $CS = \{cs_1, cs_2, \cdots, cs_N\}$. In CS, each $cs_i (i = 1, \cdots, N)$ is the centrality score of node v_i, which can be obtained from an arbitrary centrality evaluation method. For each node, we can specify it as the root node and traverse the network in a breadth-first manner to construct a BFS tree T. Then, the cumulative centrality score at the k-th level of the BFS tree can be calculated as:

$$cum\_score(k) = \sum\nolimits_{q=1}^{k} \sum_{v_j \in T(q)} cs_j, \tag{Equation 1}$$

where $T(q)$ is the set of nodes that are at the q-th level of the BFS tree. Thereafter, we can create a line chart for each node, where the horizontal (x) axis is labeled by the level number of the BFS tree and the vertical (y) axis shows the cumulative centrality score. Finally, the final centrality score for each node can be obtained by calculating the AUC value[43] of the corresponding line chart when the number of levels is no larger than k, as follows:

$$\begin{aligned} AUC(k) \quad &= cum\_score(1)/2 + (cum\_score(1) + cum\_score(2))/2 \\ &\quad + \cdots + (cum\_score(k-1) + cum\_score(k))/2 \\ &= \sum_{q=1}^{k} \left(\left(k - q + \frac{1}{2} \right) \sum_{v_j \in T(q)} cs_j \right). \end{aligned} \tag{Equation 2}$$

As shown in Equation (2), the AUC score is essentially a weighted linear combination of initial node scores in which top nodes are associated with large weight coefficients. Note that the initial node score can be either generated from an existing centrality measure or assigned in an independent way. To demonstrate the fact that TARank is a general purpose method for influential node identification without the reliance on existing centrality measures, we derive the following interesting variants under our framework by manipulating the initial node score:

- If we let $cs_i = 1 (i = 1, \cdots, N)$ and $k = 2$, then the final AUC score will be $\frac{3}{2} + \frac{|T(2)|}{2}$, where $|T(2)|$ is the number of nodes at the second level of the BFS tree. Obviously, the final AUC score of each node is the linear transformation of its degree. That is, this variant under our framework is equivalent to DC. Obviously, we can also obtain a final AUC score that is exactly the same as DC by setting $k = 2$ and $cs_i = \frac{2|T(2)|}{3+|T(2)|}$. In a nutshell, we have derived the classical centrality measure DC under our framework without using any existing evaluation methods.
- When $k = h$ (h is the height of the BFS tree), if the initial score of the root node is 1 and the scores of all other nodes are 0, then the AUC score will be $h - \frac{1}{2}$. Since the eccentricity centrality (ECC)[44] is equivalent to the height of the BFS tree, ECC is obtained as a variant of our framework under the above settings.
- CC is defined as the inverse of the average (shortest-path) distance from the root node to all other nodes. From the perspective of the BFS tree, CC can be calculated as the inverse of $\frac{\sum_{q=1}^{h}(q-1)|T(q)|}{N-1}$, where $T(q)$ is the set of nodes at the q-th level of the BFS tree. If we set $cs_i = 1$ and $k = h$, then the AUC score is $\sum_{q=1}^{h}\left(h - q + \frac{1}{2}\right)|T(q)|$. It is clear that this TARank variant can be regarded as the inverse of the weighted CC. When we define $cs_i = \frac{2}{N-1} \times \frac{\sum_{q=1}^{h}(q-1)|T(q)|}{\sum_{q=1}^{h}(2h-2q+1)|T(q)|}$ and $k = h$, this TARank variant is the same as the inverse of CC.
- So far, we have shown that it is feasible to derive some popular centrality measures under our framework. In point of fact, it is also possible to create some new, yet effective, centrality measures by using level (distance)-sensitive functions to specify the initial node score. For instance, we may utilize $g(d_i) = exp\left\{-\frac{d_i^2}{2\sigma^2}\right\}$ to set the initial score for node v_i, where d_i is the shortest-path distance from the root node to node v_i. In Figure S5, we use such a Gaussian similarity function to generate the initial score for each node. The experimental results show that the TARank variants based on this score initialization function have performance comparable to that of TARank (DC). This means that TARank can be a good candidate for identifying influential nodes without the reliance on existing centrality metrics.

The spreading models

The SIR model and SIS model are two widely used spreading models, which can be performed using the EoN[45,46] package for Python and are usually utilized to analyze the spread of diseases,[5] the diffusion of microfinance,[47] and the propagation of news.[48]

The SIR spreading model consists of three compartments: S stands for the number of susceptible nodes, I is the number of infectious nodes, and R denotes the number of recovered or removed nodes. Initially, one seed node is infected and all other nodes are in the susceptible state. Each infected node tries to infect its susceptible neighbors with the transmission probability β. Then, each infected node changes to enter the recovered state with a probability λ. The spreading process stops until there is no infected node in the network. Finally, we average the number of recovered nodes after 1,000 independent runs of the SIR spreading model. All nodes in the network will be treated as the seed node in turn to run the SIR spreading model. That is, we need to simulate the SIR model 1,000N times independently, where N is the number of nodes. For simplicity, we set $\lambda = 1$ and $\beta = \alpha\beta_c$, where α varies from 1.0 to 2.0 and β_c denotes the epidemic threshold. On the basis of the heterogeneous mean-field theory,[49–51] the epidemic threshold is roughly equivalent to $\beta_c \approx \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}$, where $\langle k\rangle$ represents the average degree and $\langle k^2\rangle$ denotes the second-order average degree.

In the SIS spreading model, there is a group of initially infected nodes and all other nodes are susceptible. Infected nodes try to infect their susceptible neighbors with the transmission probability β and then become susceptible with a probability λ_c. Then, the number of infected nodes will enter the dynamic equilibrium "endemic" state in which both the number of susceptible nodes and the number of infected nodes are constant proportions of all nodes in the network.[52] Finally, we calculate the probability that the initial infected node is still at the state of infected after simulating the SIS model 1,000 times.[4,36] For convenience, we set $\beta = \alpha\beta_c \approx \alpha\frac{\langle k\rangle}{\langle k^2\rangle}$[53] and $\lambda_c = 0.1$, where α is varied from 1.0 to 2.0.

Kendall's tau correlation coefficient

Kendall's tau correlation coefficient, also called the Kendall rank correlation coefficient, is commonly applied to measure the ordinal association between two rank lists. Consider a set of observations $(x_1, y_1), \cdots, (x_N, y_N)$ of joint ranks from X and Y. Any pair of two observations (x_i, y_i) and (x_j, y_j) $(i<j)$ is called concordant if it satisfies either both $(x_i > x_j)$ and $(y_i > y_j)$ or both $(x_i < x_j)$ and $(y_i < y_j)$; otherwise it is discordant. It should be noted that a pair of observations is neither concordant nor discordant if $(x_i = x_j)$ or $(y_i = y_j)$. Then, the Kendall's tau correlation coefficient τ is defined as follows:

$$\tau = \frac{N_c - N_d}{\binom{N}{2}} = \frac{2(N_c - N_d)}{N(N-1)},$$

where N_c is the number of concordant pairs, N_d is the number of discordant pairs, and the denominator $\binom{N}{2}$ represents the total number of pair combinations. In this paper, $\tau = 1$ means that the rank list derived from a centrality evaluation method is identical to the rank list derived from the spreading process, while $\tau = -1$ indicates that one rank list is the reverse of the other.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100321.

ACKNOWLEDGMENTS

This work has been supported by the Natural Science Foundation of China under grants 61972066 and 61572094 and the Fundamental Research Funds for the Central Universities (DUT20YG106).

AUTHOR CONTRIBUTIONS

Y.L. and Z.H. conceptualized the algorithm and wrote the manuscript. X.W. implemented the algorithm. W.C. conducted the experiments. L.H. collected the datasets and conceived the experiments. All authors discussed the results and reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 29, 2021
Revised: May 16, 2021
Accepted: July 7, 2021
Published: August 4, 2021

REFERENCES

1. Frainay, C., and Jourdan, F. (2016). Computational methods to identify metabolic sub-networks based on metabolomic profiles. Brief. Bioinform. *18*, 43–56.
2. Lordan, O., Sallan, J.M., and Simo, P. (2014). Study of the topology and robustness of airline route networks from the complex network approach: a survey and research agenda. J. Transp. Geogr. *37*, 112–120.
3. Dorogovtsev, S.N., Goltsev, A.V., and Mendes, J.F.F. (2008). Critical phenomena in complex networks. Rev. Mod. Phys. *80*, 1275.
4. Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., and Makse, H.A. (2010). Identification of influential spreaders in complex networks. Nat. Phys. *6*, 888–893.
5. Keeling, M.J., and Rohani, P. (2011). Modeling Infectious Diseases in Humans and Animals (Princeton University Press).
6. Pastor-Satorras, R., and Vespignani, A. (2002). Immunization of complex networks. Phys. Rev. E *65*, 036104.
7. Sheikhahmadi, A., and Nematbakhsh, M.A. (2017). Identification of multi-spreader users in social networks for viral marketing. J. Inf. Sci. *43*, 412–423.
8. Richardson, M., and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery)), p. 6170.
9. Albert, R., Albert, I., and Nakarado, G.L. (2004). Structural vulnerability of the north american power grid. Phys. Rev. E *69*, 025103.
10. Motter, A.E., and Lai, Y.C. (2002). Cascade-based attacks on complex networks. Phys. Rev. E *66*, 065102.
11. Jia, T., Qin, K., and Shan, J. (2014). An exploratory analysis on the evolution of the us airport network. Phys. A Stat. Mech. Appl. *413*, 266–279.
12. Daganzo, C.F. (1994). The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp. Res. B Methodol. *28*, 269–287.
13. Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.
14. Liu, Y., Liang, H., Zou, Q., and He, Z. (2020). Significance-based essential protein discovery. IEEE/ACM Trans. Comput. Biol. Bioinform. https://doi.org/10.1109/TCBB.2020.3004364.
15. Freeman, L.C. (1978). Centrality in social networks conceptual clarification. Soc. Netw. *1*, 215–239.
16. Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U S A *102*, 16569–16572.
17. Sabidussi, G. (1966). The centrality index of a graph. Psychometrika *31*, 581–603.
18. Goh, K.I., Kahng, B., and Kim, D. (2001). Universal behavior of load distribution in scale-free networks. Phys. Rev. Lett. *87*, 278701.
19. Bavelas, A. (1948). A mathematical model for group structures. Hum. Organ. *7*, 16–30.
20. Stephenson, K., and Zelen, M. (1989). Rethinking centrality: methods and examples. Soc. Netw. *11*, 1–37.
21. Bonacich, P. (1987). Power and centrality: a family of measures. Am. J. Sociol. *92*, 1170–1182.
22. Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems *30*, 107–117.
23. Lü, L., Chen, D., Ren, X., Zhang, Q., Zhang, Y., and Zhou, T. (2016). Vital nodes identification in complex networks. Phys. Rep. *650*, 1–63.
24. Das, K., Samanta, S., and Pal, M. (2018). Study on centrality measures in social networks: a survey. Soc. Netw. Anal. Min. *8*, 13.
25. Bian, R., Koh, Y.S., Dobbie, G., and Divoli, A. (2019). Identifying top-k nodes in social networks: a survey. ACM Comput. Surv. *52*, 1–33.
26. Li, X., Li, W., Zeng, M., Zheng, R., and Li, M. (2020). Network-based methods for predicting essential genes or proteins: a survey. Brief. Bioinform. *21*, 566–583.
27. Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., Zou, Q., and Zeng, X. (2020). Computational methods for identifying the critical nodes in biological networks. Brief. Bioinform. *21*, 486–497.
28. Salavaty, A., Ramialison, M., and Currie, P.D. (2020). Integrated value of influence: an integrative method for the identification of the most influential nodes within networks. Patterns *1*, 100052.
29. Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. J. ACM *46*, 604–632.
30. Weng, J., Lim, E.P., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining (Association for Computing Machinery)), pp. 261–270.
31. Kendall, M.G. (1938). A new measure of rank correlation. Biometrika *30*, 81–93.
32. Anderson, R.M., and May, R.M. (1992). Infectious Diseases of Humans: Dynamics and Control (Oxford university press).
33. Teng, B., Zhao, C., Liu, X., and He, Z. (2015). Network inference from AP-MS data: computational challenges and solutions. Brief. Bioinform. *16*, 658–674.
34. Newman, M.E.J. (2018). Network structure from rich but noisy data. Nat. Phys. *14*, 542–545.
35. Lü, L., Zhou, T., Zhang, Q., and Stanley, E.H. (2016). The h-index of a network node and its relation to degree and coreness. Nat. Commun. *7*, 10168.
36. Pastor-Satorras, R., and Vespignani, A. (2001). Epidemic spreading in scale-free networks. Phys. Rev. Lett. *86*, 3200–3203.
37. Liu, Q., Zhu, Y., Jia, Y., Deng, L., Zhou, B., Zhu, J., and Zou, P. (2018). Leveraging local h-index to identify and rank influential spreaders in networks. Phys. A Stat. Mech. Appl. *512*, 379–391.
38. Wei, D., Deng, X., Zhang, X., Deng, Y., and Mahadevan, S. (2013). Identifying influential nodes in weighted networks based on evidence theory. Phys. A Stat. Mech. Appl. *392*, 2564–2575.
39. Qu, C., Zhan, X., Wang, G., Wu, J., and Zhang, Z. (2019). Temporal information gathering process for node ranking in time-varying networks. Chaos *29*, 033116.
40. Barabási, A.L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509–512.
41. Barabási, A.L. (2009). Scale-free networks: a decade and beyond. Science *325*, 412–413.
42. Rossi, R.A., and Ahmed, N.K. (2016). An interactive data repository with visual analytics. ACM SIGKDD Explor. Newsl. *17*, 3741.
43. Purves, R.D. (1992). Optimum numerical integration methods for estimation of area-under-the-curve (AUC) and area-under-the-moment-curve (AUMC). J. Pharmacokinet. Biopharmaceut. *20*, 211–226.
44. Hage, P., and Harary, F. (1995). Eccentricity and centrality in networks. Soc. Netw. *17*, 57–63.
45. Miller, J.C., and Ting, T. (2019). EoN (epidemics on networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. J. Open Source Softw. *4*, 1731.
46. Kiss, I.Z., Miller, J.C., and Simon, P.L. (2017). Mathematics of Epidemics on Networks (Springer International Publishing).

47. Banerjee, A., Chandrasekhar, A.G., Duflo, E., and Jackson, M.O. (2013). The diffusion of microfinance. Science *341*, 1236498.

48. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. Rev. Mod. Phys. *87*, 925–979.

49. Newman, M.E.J. (2002). Spread of epidemic disease on networks. Phys. Rev. E *66*, 016128.

50. Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. (2000). Resilience of the internet to random breakdowns. Phys. Rev. Lett. *85*, 4626–4628.

51. Castellano, C., and Pastor-Satorras, R. (2010). Thresholds for epidemic spreading in networks. Phys. Rev. Lett. *105*, 218701.

52. Tassier, T. (2013). The Economics of Epidemiology (Springer Berlin Heidelberg), pp. 9–16.

53. Ferreira, S.C., Castellano, C., and Pastor-Satorras, R. (2012). Epidemic thresholds of the susceptible-infected-susceptible model on networks: a comparison of numerical and theoretical results. Phys. Rev. E *86*, 041125.