RESEARCH ARTICLE

Identifying influential nodes based on network representation learning in complex networks

Hao Wei, Zhisong Pan, Guyu Hu*, Liangliang Zhang, Haimin Yang, Xin Li, Xingyu Zhou

College of Command Information System, Army Engineering University of PLA, Nanjing, China

* huguyu@189.cn

Abstract

Identifying influential nodes is an important topic in many diverse applications, such as accelerating information propagation, controlling rumors and diseases. Many methods have been put forward to identify influential nodes in complex networks, ranging from node centrality to diffusion-based processes. However, most of the previous studies do not take into account overlapping communities in networks. In this paper, we propose an effective method based on network representation learning. The method considers not only the overlapping communities in networks, but also the network structure. Experiments on real-world networks show that the proposed method outperforms many benchmark algorithms and can be used in large-scale networks.

OPEN ACCESS

Citation: Wei H, Pan Z, Hu G, Zhang L, Yang H, Li X, et al. (2018) Identifying influential nodes based on network representation learning in complex networks. PLoS ONE 13(7): e0200091. https://doi.org/10.1371/journal.pone.0200091

Editor: Irene Sendiña-Nadal, Universidad Rey Juan Carlos, SPAIN

Received: January 7, 2018

Accepted: June 19, 2018

Published: July 9, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information file.

Funding: This work was supported by The National Key Research Development Program of China, number: 2017YFB0802800.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

Identifying influential nodes in complex networks has gained great attention in the research community [1–7]. In recent years, many methods have been put forward to find influential nodes in complex networks. The knowledge of node's spreading ability shows new insight for application such as controlling propagation of messages and rumors in social networks[8], ranking reputation of scientists[9] and finding social leaders [10],etc.

The early measure of identifying influential nodes proposed by Shimbel is Stress Centrality [11] in 1950s. He suggested that the centrality of a node should be the total number of shortest paths that go through it. Degree Centrality[12] is a direct and effective method to measure the importance of nodes, but it neglects the global structure of the network. Eigenvector Centrality [13] considers the importance of node's neighbors. Betweenness Centrality[14] and Closeness Centrality[15] need to know all topology information of networks in advance and cannot be applied to large-scale networks. Comin et al.[16] combined degree and betweenness, but it is a time-consuming measure. Chen et al.[2] proposed a semi-local centrality measure, which is a tradeoff between the low-relevant degree centrality and other time-consuming measures. Additionally, Chen et al.[17] proposed ClusterRank, a local ranking method that considers the clustering coefficient of a node. Kitsak et al.[1] suggested that the influence of a node is mainly dependent on its position in the network and proposed K-shell to measure the importance of a node. However, K-shell considers only the links between the residual nodes, whereas the links that connect to the exhausted nodes are entirely ignored. Johanhyun Bae et al.[18] extended the K-shell and proposed C_{nc} and C_{nc+}. Zeng et al.[6] proposed a mixture decomposition method called Mixed Degree Decomposition (MDD), which considers both the residual degree and the exhausted degree. FD Malliaros et al.[19] proposed K-truss decomposition and suggested that the topological properties of the nodes play a crucial role. Lü et al.[20] showed the relationship between degree, H-index and coreness by constructing an operator and proved that the convergence to coreness can be guaranteed even under an asynchronous updating process. Numerical analyses in real networks suggested that the H-index is a good tradeoff that can better quantify node influence than either degree or coreness. In the same year, Lü et al.[21] reviewed the vital nodes identification methods and experimented on real-world networks to compare the mainstream algorithms. The methods of identifying influential nodes based on random walk are mainly used in web page sorting. The typical methods are Kleinberg's HITS algorithm[22], Google's PageRank algorithm[23] and Lv's LeaderRank algorithm[8].

Most of the previous methods only consider the node's topology information. In fact, real-world networks often have a strong community structure[24]. In social as well as other types of networks, nodes often belong to multiple communities simultaneously[25]. Influential nodes always act as "bridging" between the communities and exist in community overlaps. In this paper, we propose a new local central method to identify the influential nodes. The method assumes that the more communities a node belongs to, the greater influence of the node. To identify the influential nodes, we use the network representation learning to detect overlapping communities, and then combine with the topology information of the nodes. Experiments show state of the art performance in terms of the quality of identified influential nodes.

2. Method

2.1 Network representation learning model

Network representation learning aims at learning distributed vector representation for each vertex in a network. It is also increasingly recognized as an important aspect for network analysis. Network representation learning tasks can be broadly abstracted into the following four categories: (a) node classification[26], (b) link prediction[27], (c) clustering[28], and (d) visualization[29].

J. Yang et al.[25] proposed the BIGCLAM model for network representation learning, which also covers the overlapping community detection. The model assumes that the overlaps of communities tend to be more densely connected than the non-overlapping parts. We briefly introduce this model with a bipartite graph in Fig 1. In Fig 1, the circles on the top represent communities, the squares at the bottom represent the nodes of the graph, and the edges indicate node community affiliations. Each affiliation edge in the bipartite affiliation network has a

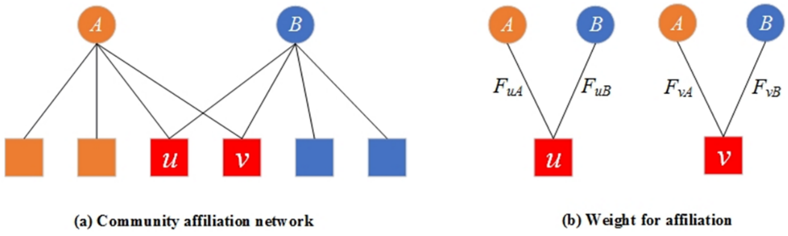

Fig 1. Bipartite community affiliation graph.

https://doi.org/10.1371/journal.pone.0200091.g001

nonnegative weight. The higher the node's weight to the community the more likely is the node to be connected to other members in the same community. Each community c creates an edge between nodes u and node v with probability $1 - exp(-F_{uc} \cdot F_{vc})$. Where F_{uc} is the nonnegative weight of node u to community c. The higher the value of F_{uc}, the more likely is the node u has an edge with the nodes in community c. Furthermore, the model assumes that each community creates edges independently. For example, in Fig 1(A), node u and node v belong to community A and community B simultaneously. In Fig 1(B), F_{uA} and F_{uB} indicate the node u's weight of the affiliation to the community A and community B respectively. In community A, the probability of existing an edge between node u and node v is $1 - exp(-F_{uA} \cdot F_{vA})$. Similarly, the probability that there is an edge between node u and node v is $1 - exp(-F_{uB} \cdot F_{vB})$. Note that since node u and node v belong to community A and community B simultaneously, node u and v receive two chances to create a link. As each community creates edges independently, the probability of an edge existing between node u and node v is $1 - exp(-\sum_{c\in A,B} F_{uc} \cdot F_{vc})$.

Given a network $G(V,E)$, where V is the node set and E is the edge set. Let $F \in \mathbb{R}^{N\times K}$ be a nonnegative matrix, where N is the number of nodes and K is the number of communities. F_{uc} is the weight between node $u \in V$ and community $c \in K$. Given F, BIGCLAM generates $G(V, E)$ by creating edge $(u,v) \in E$ between a pair of nodes $u,v \in V$ with probability $p(u,v)$:

$$p(u,v) = 1 - exp(-F_u \cdot F_v^{\,T}), \tag{1}$$

where F_u is a weight vector for node u. Each element in F_u is the weight of node u to the corresponding community. The model aims to finding the most likely affiliation factor matrix $\hat{F} \in \mathbb{R}^{N\times K}$ of the underlying network G by maximizing the likelihood:

$$\hat{F} = argmax\ P(G|F), \tag{2}$$

where

$$P(G|F) = \prod_{(u,v)\in E} p(u,v) \prod_{(u,v)\notin E} (1 - p(u,v)). \tag{3}$$

Many times, we take the logarithm of the likelihood and call it log- likelihood:

$$\hat{F} = argmax \log P(G|F), \tag{4}$$

where

$$logP(G|F) = \sum_{(u,v)\in E} log(1 - exp(-F_u F_v^T)) - \sum_{(u,v)\notin E} F_u F_v^T. \tag{5}$$

BIGCLAM learns a K-dimensional non-negative vector for each node in the network by optimizing the problem of Eq 4. Each dimension in the vector represents the probability that the node belongs to the corresponding community. After learning $\hat{F}$, the model need to determine whether node u belongs to community c or not from the value of F_{uc}. It ignores the membership of node u to community c if F_{uc} is below some threshold δ. Otherwise ($F_{uc} > \delta$), it regards u as belonging to c. One node can belong to more than one community simultaneously. Based on the BIGCLAM model, we assume that the nodes in the community overlaps play the 'bridging' role between the communities. As these nodes belong to multiple communities, information through these nodes can be easily spread to other communities. It is reasonable to assume that nodes in community overlaps have greater influence.

2.2 Network constraint coefficient

Structural holes is a concept from social network research, which is originally developed by Burt[30][31]. A structural hole is understood as a gap between two individuals who have

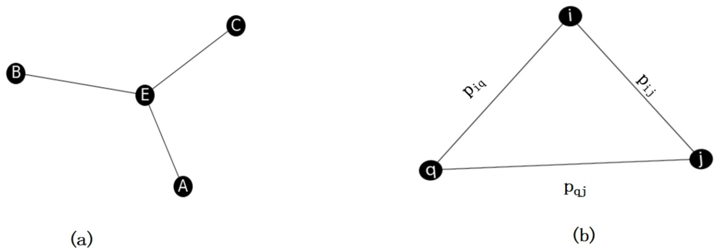

Fig 2. The concept of structural holes.

https://doi.org/10.1371/journal.pone.0200091.g002

complementary sources to information. Fig 2(A) is a structural hole of node E. The position of node E makes it serve as a bridge or a 'broker' between three different nodes. Thus, node E is likely to receive some non-redundant information from its contacts. The term 'structural holes' is used for the separation between non-redundant contacts. Because of the hole between two contacts, they provide network benefits to the third party (to node E).

Burt used the network constraint coefficient C to measure the constraints imposed by forming a structural hole:

$$C_i = \sum\nolimits_{j\in\Gamma(i)} \left(p_{ij} + \sum\nolimits_{q\epsilon(\Gamma(j)\cap\Gamma(i))} (p_{iq}p_{qj})\right)^2, \tag{6}$$

where $\Gamma(i)$ is the neighbor set of node i. As shown in Fig 2(B), p_{ij} is proportion of i's energy invested in relationship with j and $p_{ij} = \frac{1}{N(i)}$, where $N(i)$ represents the degree of node i. $p_{iq} = \frac{1}{N(i)}$ and $p_{qj} = \frac{1}{N(q)}$ represent node i's and node j's energy respectively invested in relationship with the common neighbor q. From Eq 6 we can see that a node with a small constraint coefficient indicates that the degree of the node is large and the connections among neighbors are sparse. Thus, the node with a small constraint coefficient would have more chances to spread the information to a large portion of the network. The smaller the constraint coefficient of a node is, the faster the node can spread information.

2.3 Ranking method

Nodes in community overlaps play the 'bridging' role between the communities. Information can be spread to multiple communities through these nodes. The number of communities a node be longs to can be regarded as its propagation capacity. The more communities a node belongs to, the more communities the node can influence. Network constraint coefficient of a node can be regarded as the propagation speed in community. The smaller the constraint coefficient of a node is, the faster the node can spread information. We consider both propagation capacity and propagation speed of a node to evaluate its influence, which denoted by OC. The OC of node i is defined as follows:

$$OC_i = \frac{\sum_{j\in\Gamma(i)}\sum_{k\epsilon\Gamma(j)} 10^{-C_k} * Nb(k)}{maxOC}, \quad i = 1, 2, 3 \ldots N, \tag{7}$$

where $\Gamma(i)$ is the neighbor set of node i, C_k is the network constraint coefficient of node k, $maxOC$ is the normalization factor, $Nb(k)$ represents the total number of communities that node k's neighbors belong to. For example, assuming that node i has 3 neighbors a, b, c, node a has 2 communities named 1, 4, node b has 3 communities named 1, 2, 3, and node c has 4 communities named 2, 4, 5, 6, then the communities that the neighbors of node i have are 1, 2, 3, 4, 5, 6. Thus, $Nb(i)$ = 6.

To identify influential nodes in the network, we need to know the total number of communities that each node belongs to and their constraint coefficients. First, we use the BIGCLAM model to detect overlapping communities, and calculate the total number of communities that each node belongs to. Second, we calculate the network constraint coefficient for each node based on Eq 6. Third, according to Eq 7, we calculate the OC value of each node. Note that if there are no overlapping communities in the network, OC degrades to the network constraint coefficient.

3. Experiments

3.1 Evaluation method

To evaluate the performance of the proposed method, we use the SIR model[32] to examine the influence of nodes. The model is used to simulate the spread of the virus or information process. The SIR model divides the network nodes into three types: (1) Susceptible nodes, healthy but not immune and can be infected; (2) Infected nodes, already infected and can infect susceptible nodes; (3) Recovered nodes, which have been cured and cannot be infected again. In the beginning, the node to be tested is in the Infected state whereas the rest of the nodes of the network are in the Susceptible state. This node triggers a spreading process where every infected node can infect its neighbors at each timestep t with probability β. Each infected node is cured with probability γ. In this paper, we set $\gamma = 1$, which means that each node has only one chance to infect its neighbors in every round. The sum ($F(t)$) of recovered nodes at time t when there is no infected node exiting in the network are defined as the influence of the node. In order to ensure the spreading process, we set β to be slightly larger than the epidemic threshold ($\beta_{th} = \frac{\langle k\rangle}{[\langle k^2\rangle - \langle k\rangle]}$) in the network[33], where $\langle \mathrm{k}\rangle$ and $\langle \mathrm{k}^2\rangle$ denote the average degree and the second order average degree, respectively. In this paper, we set $\beta_{th} = \frac{\langle k\rangle}{\langle k^2\rangle}$.

To quantify the correctness of the ranking methods, we adopt Kendall's tau[34] as a rank correlation coefficient, which is defined as follows:

$$\tau(R_1, R_2) = \frac{n_c - n_d}{\sqrt{(n_t - n_{t1})(n_t - n_{t2})}}, \tag{8}$$

where R_1 and R_2 are two different rank lists, $n_t = n(n - 1)/2$, $n_{t1} = \sum_i t_i(t_i - 1)/2$, $n_{t2} = \sum_j t_j(t_j - 1)/2$, t_i and t_j are the number of tied values in the ith and jth groups of ties, respectively. n_c and n_d are the numbers of concordant and discordant pairs, respectively. For example, let X and Y be two ranking lists. (x_1, y_1), (x_2, y_2),. . .,(x_i, y_j) are a set of joint ranks from X and Y, respectively. Any pair of ranks (x_i, y_i) and (x_j, y_j) is said to be concordant if $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$. If $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$, the pair is said to be discordant. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. This metric quantifies the similarity between the orderings of the measures and the real ranking.

3.2 Experimental data

Nine real-world networks are used to evaluate the performance of the proposed method: (1) all meeting articles that appeared in 1994-2000(**GDciting**). The data can be obtained on "*https://www.aminer.cn/citation*"; (2) US airport flights(**USAir97**). The data can be downloaded on "*http://vlado.fmf.uni-lj.si/pub/networks/data/*"; (3) collaboration network of scientists(**Netscience**)[35]; (4) communication network of Blogs(**Blogs**)[36]; (5) an e-mail communication network(**Email**)[37]; (6) C.elegans networks(**C.elegans**)[38]; (7) Lusseau's Bottlenose Dolphins(**Dolphins**)[39]; (8) Arxiv COND-MAT collaboration network(CA-CondMat)[40]; (9) Amazon network data(**Amazon**). The data can be downloaded on "*https://snap.stanford.edu/*". Table 1 shows the information of each network, where n is the number of

Table 1. The statistical properties of the networks, where *n* is the number of nodes, *m* is the number of edges, C is the number of communities divided by BIGCLAM model, MLC is the maximum number of communities the nodes have in the network.

Network	*n*	*m*	*C*	*MLC*
GDciting	311	647	13	4
USAir97	332	2126	17	10
Netscience	1461	2742	96	5
Blogs	112	425	8	3
Email	1133	5451	39	10
C.elegans	248	468	13	4
Dolphins	62	159	7	2
CA-CondMat	23133	93497	100	34
Amazon	334863	925872	73854	139

https://doi.org/10.1371/journal.pone.0200091.t001

nodes, *m* is the number of edges, *C* is the number of communities divided by the BIGCLAM model, *MLC* is the maximum number of communities owned by the node in the network.

3.3 Experimental results

In this section, we compare the proposed method OC with Degree Centrality(DC), Betweenness Centrality(BC), Closeness Centrality(CC), Eigenvector Centrality(EC), C_{nc}, $\mathrm{C}_{\mathrm{nc+}}$, Network Constraint Coefficient(NC) and K-shell(KS). In each implementation, one node is selected to be infected, and then infects its neighbors according to the SIR model. The influence of the node ($F(t)$) is the sum of recovered nodes when the spreading process fade out. This value represents the average over multiple executions of the model (we performed 1000 simulations for large and 100 simulations for small datasets). Without special explanation, in this paper, the value of β is shown in Table 2, $\gamma = 1$, and the threshold for the dataset is 1000 nodes.

In Table 2, we compare the Kendall correlation coefficient τ of different ranking methods. The results in Table 2 manifest that our method outperforms the other methods in most cases.

Based on the above results, we plot the correlation of the influence measures in GDciting, Dolphins and CA-CondMat. The results are shown in Fig 3, Fig 4 and Fig 5 respectively. Due to the large number of nodes in CA-CondMat, we only show the result of the top 500 nodes. We can witness that there is a clear correlation between $F(t)$ and OC, while the traditional measures, i.e., the BC and the CC, have little relationship with the influence capability of the spreaders in an epidemic process.

Table 2. The ranking results of each network. Here β_{th} is the epidemic threshold for networks; β is the infection probability in SIR simulation; $\tau(\cdot)$ represents the Kendall correlation coefficient of corresponding methods for given β. "-" means the method is still no result when running time exceeds 48 hours.

Network	β_{th}	β	τ_{DC}	τ_{BC}	τ_{CC}	τ_{NC}	τ_{KS}	$\tau_{C_{nc}}$	$\tau_{C_{nc+}}$	τ_{EC}	τ_{OC}
GDciting	0.102	0.176	0.71	0.52	0.82	0.71	0.70	0.84	0.85	0.89	**0.95**
USAir97	0.021	0.033	0.85	0.60	0.84	0.88	0.84	0.94	0.94	0.93	**0.96**
Netscience	0.115	0.264	0.63	0.20	0.81	0.75	0.57	0.72	0.73	0.88	**0.89**
Blogs	0.067	0.106	0.86	0.69	0.88	0.82	0.78	0.91	0.92	**0.94**	0.92
Email	0.05	0.079	0.74	0.61	0.77	0.75	0.75	0.80	0.81	0.84	**0.87**
C.elegans	0.143	0.204	0.72	0.57	0.84	0.70	0.70	0.83	0.83	0.85	**0.93**
Dolphins	0.147	0.231	0.72	0.52	0.70	0.76	0.56	0.79	0.80	0.75	**0.97**
CA-CondMat	0.045	0.054	0.32	0.25	0.35	0.32	0.12	0.48	0.26	0.64	**0.66**
Amazon	0.095	0.114	0.43	-	-	0.46	0.23	0.52	0.53	0.56	**0.61**

https://doi.org/10.1371/journal.pone.0200091.t002

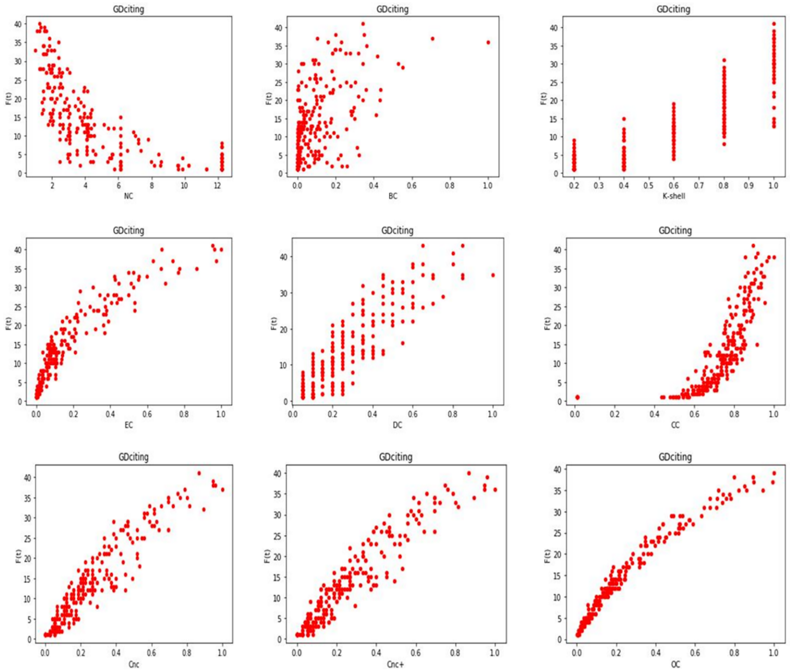

Fig 3. The relation between node's influence and the ranking methods in GDciting.

https://doi.org/10.1371/journal.pone.0200091.g003

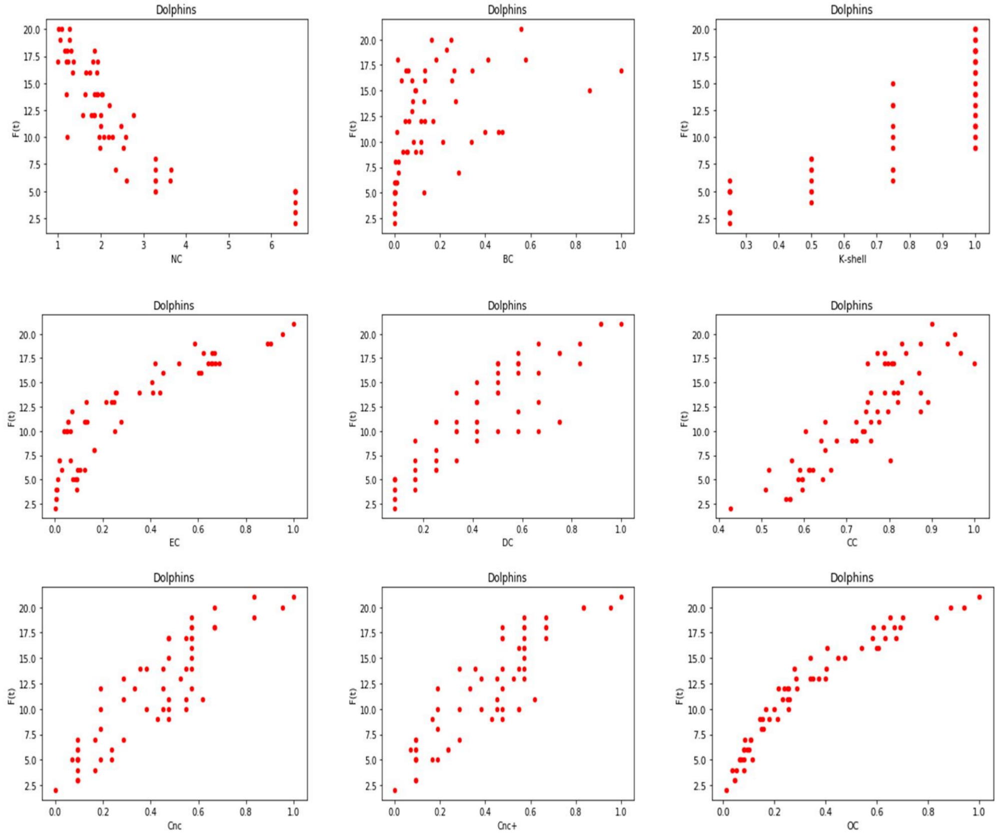

Fig 4. The relation between node's influence and the ranking methods in Dolphins.

https://doi.org/10.1371/journal.pone.0200091.g004

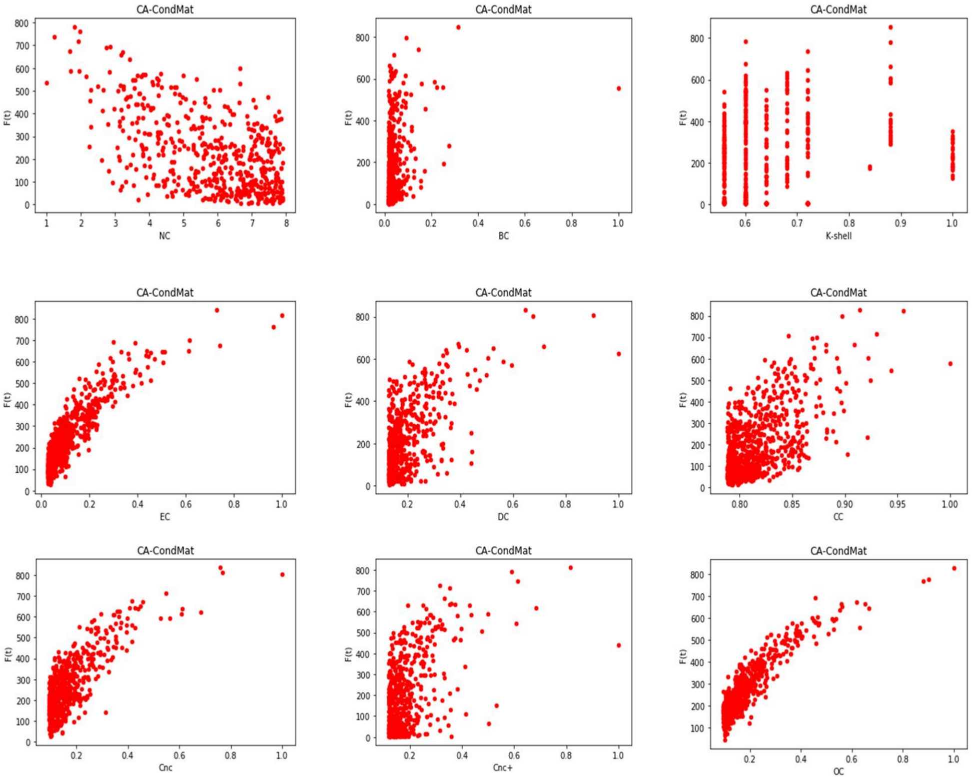

Fig 5. The relation between node's influence and the ranking methods in CA-CondMat.

https://doi.org/10.1371/journal.pone.0200091.g005

As the proposed method OC contains the network constraint coefficient and the node degree information, we compare OC with DC and NC. The results are shown in Fig 6. The color of each point represents the influence of the node. We can observe that OC has strong correlation with DC and NC, but there are still many influential nodes with small values of DC and many little influential nodes with small values of NC. It indicates that NC or DC alone is

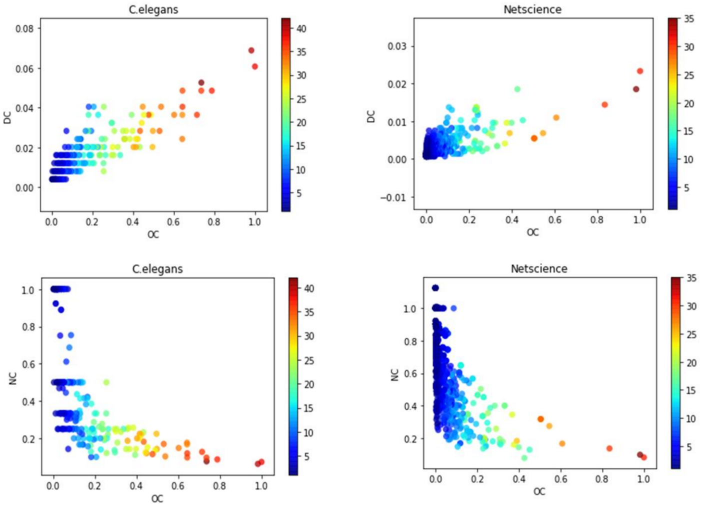

Fig 6. The relations between OC and NC, OC and DC in Netscience and C.elegans.

https://doi.org/10.1371/journal.pone.0200091.g006

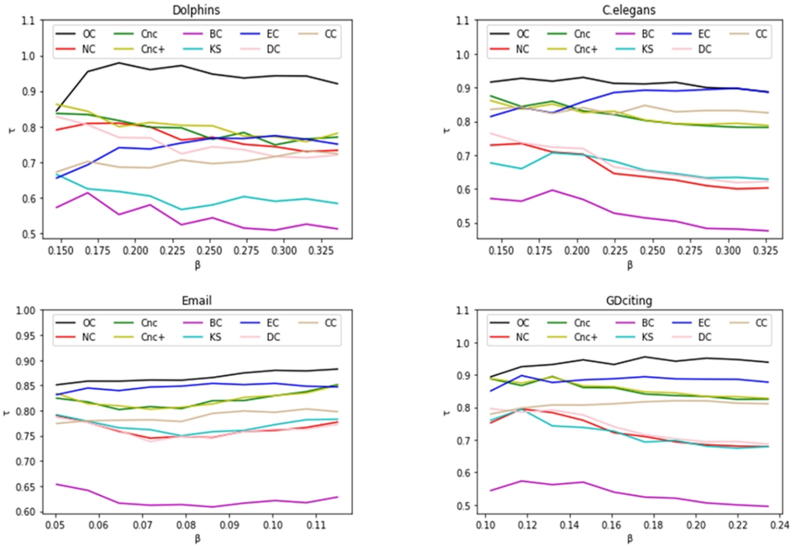

Fig 7. The rank correlation coefficient, Kendall's tau τ, is plotted by varying the infection probability β in four networks: Dolphins, C.elegans, Email and GDciting.

https://doi.org/10.1371/journal.pone.0200091.g007

not sufficient to identify influential nodes. The proposed method OC contains both constraint coefficient and node degree information. It can identify influential nodes better.

To further estimate how the infection probability β affects the effectiveness of different methods, the Kendall correlation coefficient τ as a function of β for different methods is shown in Fig 7. The infection probability β varies from β_{th} to $2\beta_{th}$. As described in Fig 7, on a wide range of probabilities β, OC is better than other measures in the four networks. In Fig 8, we conduct the same experiments for different values of γ, which varies from 0.5 to 1. As shown in Fig 8, the proposed OC presents better results than the other measures in the four networks.

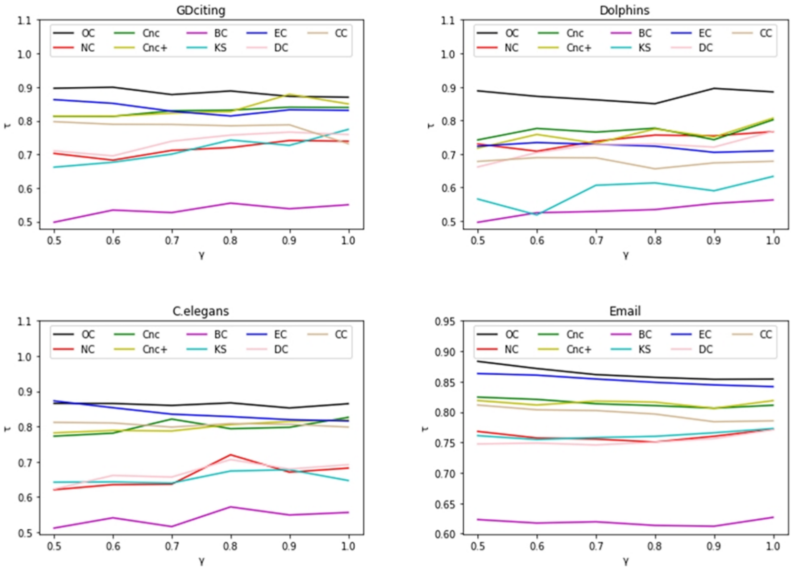

Fig 8. The rank correlation coefficient, Kendall's tau τ, is plotted by varying the recovery probability γ in four networks: GDciting, Dolphins, C. elegans, Email.

https://doi.org/10.1371/journal.pone.0200091.g008

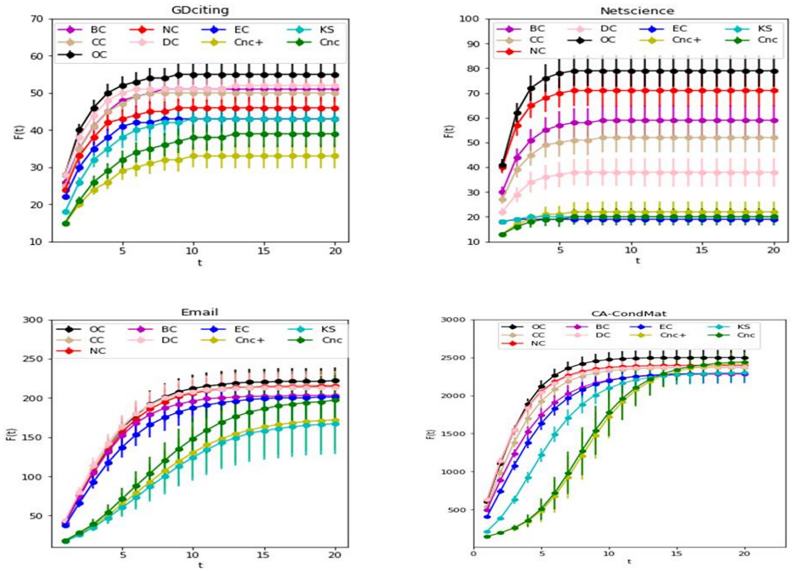

Fig 9. The cumulative number of infected nodes as a function of time *t* in four networks: GDciting, Netscience, Email and CA-CondMat.

https://doi.org/10.1371/journal.pone.0200091.g009

Furthermore, we investigate the number of nodes being at the infected and recovered state for various timesteps of the SIR model. We focus on the nodes that appear in the top-*k* lists of each method. We initially set these nodes to be infected. Here we set $k = 10$ for the small networks (GDciting, Netscience and Email) and $k = 100$ for the large network (CA-CondMat). The cumulative number of infected nodes ($F(t)$) as a function of time t in the four networks are shown in Fig 9. Due to the randomness of transmission, the experimental result of SIR model is different in each experiment. We use the error bar graph to present the results. As shown in Fig 9, the number of cumulative infected nodes increases with time and ultimately reach the steady value. For all these four networks, OC outperforms the other methods for both spreading rate and the number of infected nodes.

In many cases, people are more interested in a small fraction of the most influential nodes in the network. Here, we use L to represent the fraction of the most influential nodes measured by each method. We let L vary from 0.1 to 1.0 and do the influence comparison experiment between the top nodes ranked by different methods. As shown in Fig 10, our method outperforms the other methods on almost the entire range of L in the four networks.

Determine the number of communities c is a challenging task in community detection. As our method need to use the result of community detection, it is necessary to evaluate the impact of the number of communities c on the result. We divide the network into different numbers of communities and then identify the influential nodes. Fig 11 shows the results in the three networks. The horizontal axis represents the number of communities divided and the vertical axis is the network correlation coefficient τ. As can be seen from Fig 11, the number of communities divided has a limited effect on the results. The fluctuation of τ does not exceed 0.1 with c varying.

4. Conclusion

Identifying influential nodes in complex networks is very important in theoretical and practical applications. In this paper, we proposed an efficient method based on BIGCLAM model. The method suggests that the community overlaps play the "bridging" role between the

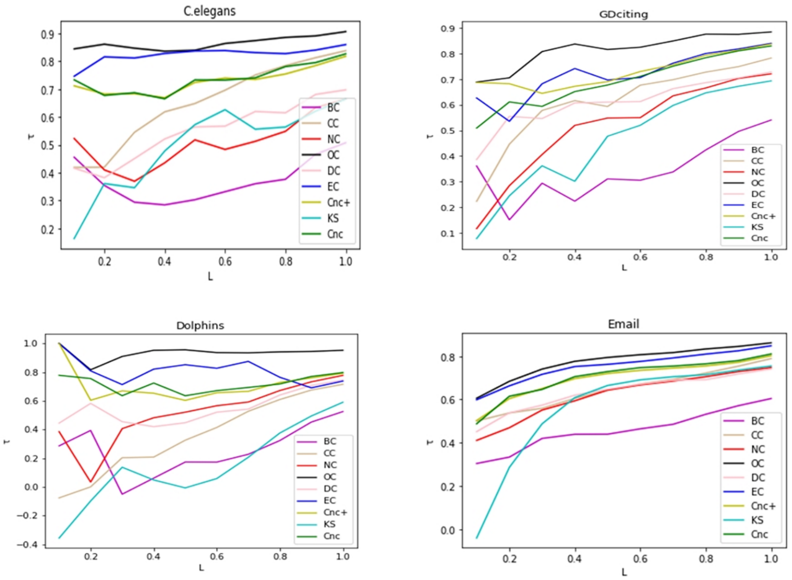

Fig 10. The values of τ for all methods in four networks when *L* varies from 0.1 to 1.0. Here *L* represents the percentage of nodes with the largest spreading ability.

https://doi.org/10.1371/journal.pone.0200091.g010

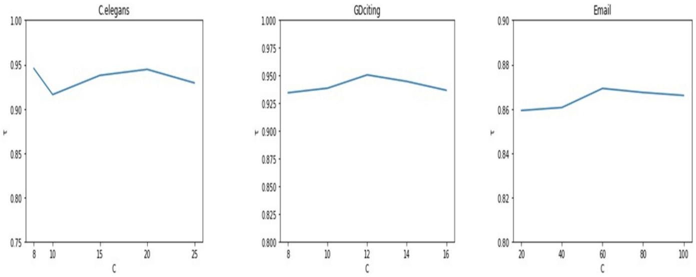

Fig 11. The rank correlation coefficient, Kendall's tau *τ*, is plotted by varying the number of communities divided in three networks: C.elegans, GDciting and Email.

https://doi.org/10.1371/journal.pone.0200091.g011

communities. The number of communities that a node belongs to represents its propagation capacity. In addition, we consider the network constraint coefficient of the node, which represents its propagation speed in community. The comparison results between the proposed method and the benchmark algorithms demonstrated that proposed method can obtain the best results. Our results could shed some light on how to utilize network representation learning and overlapping community detection to identify influential nodes.

Supporting information

S1 Dataset. Nine real-world networks used in this paper.
(ZIP)

Acknowledgments

Our work is supported by the National Key Research Development Program of China (No.2017YFB0802800). The authors would like to thank the Editor-in-Chief and anonymous reviewers for their insightful and constructive commendations that have led to an improved version of this paper.

Author Contributions

Conceptualization: Guyu Hu.

Data curation: Zhisong Pan.

Formal analysis: Zhisong Pan.

Funding acquisition: Zhisong Pan, Guyu Hu.

Investigation: Hao Wei.

Methodology: Hao Wei, Liangliang Zhang.

Project administration: Hao Wei, Liangliang Zhang.

Resources: Guyu Hu, Xin Li.

Supervision: Hao Wei, Liangliang Zhang.

Validation: Xin Li.

Visualization: Hao Wei.

Writing – original draft: Hao Wei.

Writing – review & editing: Hao Wei, Haimin Yang, Xingyu Zhou.

References

1. Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H. E, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2011, 6(11):888–893.
2. Chen D, Lü L, Shang M S, Zhang Y C, Zhou T. Identifying influential nodes in complex networks[J]. Physica A Statistical Mechanics & Its Applications, 2012, 391(4):1777–1787.
3. Zhang X, Zhu J, Wang Q, Zhou H. Identifying influential nodes in complex networks with community structure[J]. Knowledge-Based Systems, 2013, 42(2):74–84.
4. Hou B, Yao Y, Liao D. Identifying all-around nodes for spreading dynamics in complex networks[J]. Physica A Statistical Mechanics & Its Applications, 2012, 391(15):4012–4017.
5. Basaras P, Katsaros D, Tassiulas L. Detecting Influential Spreaders in Complex, Dynamic Networks[J]. Computer, 2013, 46(4):24–29.
6. Zeng A, Zhang C J. Ranking spreaders by decomposing complex networks[J]. Physics Letters A, 2012, 377(14):1031–1035.
7. Liu J G, Ren Z M, Guo Q. Ranking the spreading influence in complex networks[J]. Physica A Statistical Mechanics & Its Applications, 2014, 392(18):4154–4159.
8. Lü L, Zhang Y C, Chi H Y, Tao Z. Leaders in Social Networks, the Delicious Case[J]. Plos One, 2011, 6 (6):e21202. https://doi.org/10.1371/journal.pone.0021202 PMID: 21738620
9. Zhou Y B, Lü L, Li M. Quantifying the influence of scientists and their publications: Distinguish prestige from popularity[J]. New Journal of Physics, 2011, 14(3):33033–33049(17).
10. Lü L, Chen D B, Zhou T. Small world yields the most effective information spreading[J]. New Journal of Physics, 2011, abs/1107.0429(12):825–834.
11. Shimbel A. Structural parameters of communication networks[J]. Bulletin of Mathematical Biology, 1953, 15(4):501–507.
12. Gao S, Ma J, Chen Z, Wang G, Xing C. Ranking the spreading ability of nodes in complex networks based on local structure[J]. Physica A Statistical Mechanics & Its Applications, 2014, 403(6):130–147.
13. Stephenson K, Zelen M. Rethinking centrality: Methods and examples[J]. Social Networks, 1989, 11 (1):1–37.
14. Freeman L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1979, 1 (3):215–239.

15. Sabidussi G. The centrality index of a graph[J]. Psychometrika, 1966, 31(4):581–603. PMID: 5232444

16. Comin C H, Costa L F. Identifying the starting point of a spreading process in complex networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2011, 84(5 Pt 2):056105.

17. Duanbing C, Hui G, Linyuan L. & Tao Z. Identifying influential nodes in large-scale directed networks: The role of clustering. PLoS ONE 8(10), e77455 (2013). https://doi.org/10.1371/journal.pone.0077455 PMID: 24204833

18. Bae J, Kim S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness[J]. Physica A Statistical Mechanics & Its Applications, 2014, 395(4):549–559.

19. Malliaros F D, Rossi M E, Vazirgiannis M. Locating influential nodes in complex networks[J]. Scientific Reports, 2016, 6:19307. https://doi.org/10.1038/srep19307 PMID: 26776455

20. Lü L, Zhou T, Zhang Q M, Stanley H E. The H-index of a network node and its relation to degree and coreness[J]. Nature Communications, 2016, 7:10168. https://doi.org/10.1038/ncomms10168 PMID: 26754161

21. Lü L, Chen D, Ren X L, Zhang Y C, Zhou T. Vital nodes identification in complex networks[J]. Physics Reports, 2016, 650:1–63.

22. Kleinberg J M. Authoritative sources in a hyperlinked environment[M]. ACM, 1999.

23. Bryan K, Leise T. The Eigenvector: The Linear Algebra behind Google[J]. Siam Review, 2006, 48 (3):569–581.

24. Fu J, Wu J, Liu C, Xu J. Leaders in communities of real-world networks [J]. Physica A Statistical Mechanics & Its Applications, 2015, 444:428–441.

25. Yang J, Leskovec J. Overlapping community detection at scale: a nonnegative matrix factorization approach[C]// ACM International Conference on Web Search and Data Mining. ACM, 2013:587–596.

26. Bhagat S, Cormode G, Muthukrishnan S. Node Classification in Social Networks[M]// Social Network Data Analytics. Springer US, 2011:115–148.

27. Wang H, Wang J, Wang J, Zhao M, Zhang W, Zhang F, et al. GraphGAN: Graph Representation Learning with Generative Adversarial Nets[J]. 2017.

28. Ding C H Q, He X, Zha H, Gu M, Simon H D. A Min-max Cut Algorithm for Graph Partitioning and Data Clustering[C]// IEEE International Conference on Data Mining. IEEE Computer Society, 2001:107–114.

29. Maaten L V D, Hinton G. Visualizing Data using t-SNE[J]. Journal of Machine Learning Research, 2017, 9(2605):2579–2605.

30. Krackhardt D. Structural Holes: The Social Structure of Competition[J]. Administrative Science Quarterly, 1995.

31. Burt R S, Kilduff M, Tasselli S. Social Network Analysis: Foundations and Frontiers on Advantage[J]. Annual Review of Psychology, 2013, 64(1):527–547.

32. Pastorsatorras R. Epidemic Spreading in Scale-Free Networks[J]. Physical Review Letters, 2001, 86 (14):3200–3. https://doi.org/10.1103/PhysRevLett.86.3200 PMID: 11290142

33. Castellano C, Pastorsatorras R. Thresholds for epidemic spreading in networks. [J]. Physical Review Letters, 2010, 105(21):218701. https://doi.org/10.1103/PhysRevLett.105.218701 PMID: 21231361

34. Knight W R. A Computer Method for Calculating Kendall's Tau with Ungrouped Data[J]. Journal of the American Statistical Association, 1966, 61(314):436–439.

35. Newman M E J. Finding community structure in networks using the eigenvectors of matrices. [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2006, 74(3 Pt 2):036104.

36. N. Xie, Social network analysis of blogs, MSc Dissertation. University of Bristol, 2006.

37. Bassett D.S., Porter M.A., Wymbs N.F., Grafton S.T., Carlson J.M., Mucha P.J., Robust detection of dynamic community structure in networks, Chaos: An Interdisciplinary Journal of Nonlinear Science, 23 (2013) 013142.

38. Duch J, Arenas A. Community detection in complex networks using extremal optimization. [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2005, 72(2 Pt 2):027104.

39. Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E, Dawson S M. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology & Sociobiology, 2003, 54(4):396–405.

40. Leskovec J, Kleinberg J, Faloutsos C. Graph evolution:Densification and shrinking diameters[J]. Acm Transactions on Knowledge Discovery from Data, 2006, 1(1):2.