Identifying influential nodes based on graph signal processing in complex networks*

Zhao Jia(赵 佳), Yu Li(喻 莉)†, Li Jing-Ru(李静茹), and Zhou Peng(周 鹏)

Department of Electronics and Information Engineering, Huazhong University of Science and technology, Wuhan 430074, *China*

(Received 28 July 2014; revised manuscript received 9 December 2014; published online 27 March 2015)

Identifying influential nodes in complex networks is of both theoretical and practical importance. Existing methods identify influential nodes based on their positions in the network and assume that the nodes are homogeneous. However, node heterogeneity (i.e., different attributes such as interest, energy, age, and so on) ubiquitously exists and needs to be taken into consideration. In this paper, we conduct an investigation into node attributes and propose a graph signal processing based centrality (GSPC) method to identify influential nodes considering both the node attributes and the network topology. We first evaluate our GSPC method using two real-world datasets. The results show that our GSPC method effectively identifies influential nodes, which correspond well with the underlying ground truth. This is compatible to the previous eigenvector centrality and principal component centrality methods under circumstances where the nodes are homogeneous. In addition, spreading analysis shows that the GSPC method has a positive effect on the spreading dynamics.

Keywords: complex networks, graph signal processing, influential node identification

PACS: 89.75.–k, 02.30.Nw, 02.70.Hm **DOI:** 10.1088/1674-1056/24/5/058904

1. Introduction

Complex networks have attracted much attention in recent years since they meet the urgent need to understand the structures and dynamics of many real systems.[1] In this research field, identifying influential nodes is an essential task.[2] A small fraction of influential nodes can greatly affect the dynamics of networks, such as disease spreading,[3] information propagation,[4,5] election,[6] and cascading failures.[7] Moreover, the identification of the influential nodes plays an important role in controlling rumor spreading, defining new marketing strategies, and even predicting the total sale.[8,9] All of these practical applications contribute to the importance of identifying influential nodes in complex networks.

In previous work, nodes occupying important positions in the network topology are regarded as important in complex networks. Known methods of ranking node centrality include degree centrality, closeness centrality,[10] betweenness centrality,[11] cluster centrality,[12] and eigenvector centrality.[13] The degree centrality is straightforward and identifies the most connected nodes. The closeness centrality identifies those near to other nodes most. The betweenness centrality identifies those located on the most traveled paths. The cluster centrality identifies those with high clustering coefficient and influential neighbors. The eigenvector centrality (i.e., EVC, including PageRank[14] and subsequent LeaderRank[15]) considers nodes connected to other high degree nodes as highly central. Since influential nodes identified by the EVC tend to locate in a restricted region of the network topology, the principal component centrality (PCC) method[16] using several principal eigenvalues is proposed as a global view.

The prior methods mainly focus on the network topology assuming that the nodes are homogeneous. However, in reality, node heterogeneity exists.[17–19] For example, in wireless sensor networks, some nodes may be of different batteries to prolong their lifetime and reliability. Even homogeneous sensors have different levels of initial energy, depletion rate, etc. As a matter of fact, node heterogeneity greatly affects routing, epidemic spread, and safety deployment.[17–19] Besides, the topological equivalence of homogeneous networks (e.g., regular networks, small-world networks[20]) makes the network topology based node centrality approximately equal. Therefore, it is necessary to adopt node attributes as a supplement for ranking node centrality. Moreover, owing to the accumulated prior knowledge of complex networks, node attributes entitle us to rank the node centrality more comprehensively. In other words, node heterogeneity needs to be taken into account in the influential node identification.

In this paper, we propose a graph signal processing based centrality (GSPC) method to identify the influential nodes considering both the node heterogeneity and the network topology. The first intuitive step is to derive an overall evaluation of node heterogeneity (i.e., different attributes such as interest, energy, age, and so on), which is called the node signal. We categorise node attributes into three types and propose a modulated 2-norm based model to obtain the node signal. In order to integrate the node signals with the network topology, we ap-

*Project supported by the National Natural Science Foundation of China (Grant No. 61231010) and the Fundamental Research Funds for the Central Universities, China (Grant No. HUST No. 2012QN076).

†Corresponding author. E-mail: hustlyu@hust.edu.cn

© 2015 Chinese Physical Society and IOP Publishing Ltd

ply the graph signal processing method. In particular, the node signals are processed by the graph Fourier transform. Signal processing techniques and methodologies are used to pick up the most principal eigenvectors. Then, the node centrality is calculated by the combination of the node signal and the selected eigenvectors. By this means, we identify the influential nodes with both node attributes and the network topology.

We first evaluate our proposed method in two real-world datasets with node heterogeneity information and underlying ground truth. We have three major findings. First, identified nodes based on our GSPC method corresponds well with the underlying ground truth. Second, GSPC degenerates to the previous EVC and PCC methods under the circumstance where the nodes are homogeneous. Furthermore, the results under varying parameters and null models prove the stability of our GSPC method. In addition, spreading in a synthetic network and a real social network is investigated, which indicates that the GSPC method has a positive effect on the spreading dynamic.

The rest of this paper is organized as follows. Section 2 presents the modulated 2-norm based evaluation model of node heterogeneity. In Section 3, we propose a GSPC method which integrates both node attributes and the network topology. The data description and centrality measures of two networks are, respectively, given in Subsections 4.1 and 4.2, compatibility analysis and stability analysis are, respectively, discussed in Subsections 4.3 and 4.4. Then, spreading is analyzed in Subsection 4.5. Section 5 ends this paper with some conclusions.

2. Evaluation of node heterogeneity

Node behaviors are significantly affected by node heterogeneity (i.e., different node attributes) in various networks. For example, individuals in social networks determine their interactions concerning age, gender, profession, opinions, and so on, while nodes in wireless networks forward packets due to trust, delay, battery, and so on. Therefore, in this section, node attributes are investigated and a modulated 2-norm based model is proposed to achieve an overall evaluation of each node.

We divide node attributes into three types. Firstly, since dynamics like rumor control, immunization, and opinion diffusion involve opposite sides, parts of the node attributes are used to decide the sign, which is denoted as g. The sign type refers to whether or not you could support, ignore, or oppose an event, which can be modeled as a triple $g \in (+1, 0, -1)$, where $+1$, 0, -1 stand for endorsement, disinterest, and opposition, respectively. For example, for opinions diffusion in social networks, if node i stands for one opinion, g is positive. And if the node is opposed to the opinion, it should be negative. Otherwise, if it pays no attention, $g = 0$, that is, it is insignificant in the network dynamics.

Besides, analogous to the trust evaluation of nodes,[21] each node has two other types of node attributes: confidence and ability. The confidence type corresponds to the probability at which nodes function like online time, activity, and frequency of tweets, denoted as t. For a simplified example, the online time in social networks could be modeled as a probabilistic model. For a specific hour in a day, if a node gets online α times in β days, then the probability that this node gets online in this hour is $t = \alpha/\beta$ according to the minimum unbiased estimation.

The ability type corresponds to the potential service provided by the nodes, denoted as a. A high ability means that the node has more potential to respond to dynamics on the networks fast and wide. Owing to the development of computer science, empirical analysis[9,22] of various social networks have sprung up. Empirical data of a tweet indicates that a user generally follows experts on various topics of her/his interest.[23] Elite users like media, celebrities, bloggers, and organizations roughly are 0.05% of the population and account of about 50% attention.[24,25] Other personal information like age, gender, location, and so on is less mentioned. Thus identity (e.g., celebrities, actors, writers), role (e.g., connectors, experts, and salesman) and interest (e.g., favorite book, music, and so on) are regarded as some of the more important node attributes in social networks. In this way, we roughly classify the ability attributes into two categories in Table 1.

Table 1. The empirical classification of the ability attributes.

Important	identity (e.g., celebrities, actors, writers)
	role (e.g., connector, experts, salesman)
	interest (e.g., favorite book and music)
Less important	age, gender, location

Because ability attributes have different importances, coefficients η_i are introduced to modulate the significance of each attribute $\boldsymbol{a}_i$ based on two rules. First, make sure that $\eta_1 > \eta_2$ if $\boldsymbol{a}_1$ belongs to the important category while $\boldsymbol{a}_2$ belongs to the less important category. Attributes that belong to the same category are assigned the same coefficients. Second, $\sum_i \eta_i = 1$. Examples and details of η will be discussed in the stability analysis in Subsection 4.4.

Then, node ability is codified as the distance to the zero point and is calculated by a modulated 2-norm model. That is, $a = ||\eta \boldsymbol{a}||_2 = |\eta_1 \boldsymbol{a}_1 + \cdots + \eta_N \boldsymbol{a}_N|$ ($\sum_i \eta_i = 1$, N is the number of the ability attributes). Without loss of generality, we just take 3-dimension as a simple example, as shown in Fig. 1.

In conclusion, the overall evaluation of node i is calculated as

$$f(i) = g \times t \times a. \tag{1}$$

It could be either positive or negative, the higher the absolute value is, the more important role the node plays in networks.

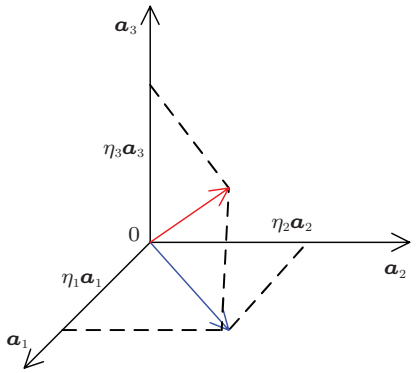


Fig. 1. (color online) A simple example of nodes with three ability attributes. It is codified as a distance to the zero point. And each attribute is modulated according to its importance.

With an overall evaluation of node attributes at each node, denoted as the node signal, a complex network can be intuitively represented as a graph with node signals on it, as in the example of the Zachary club network shown in Fig. 2. Each node has a value $f(i)$. The blue line means positive, and the red line means negative. The height represents the absolute value of each node. Nodes 10, 17, and 19 have zero signal because they do not support either node 1 or node 34 in the Zachary club network. By this means, we can integrate node heterogeneity with the network topology to identify influential nodes in the next section.

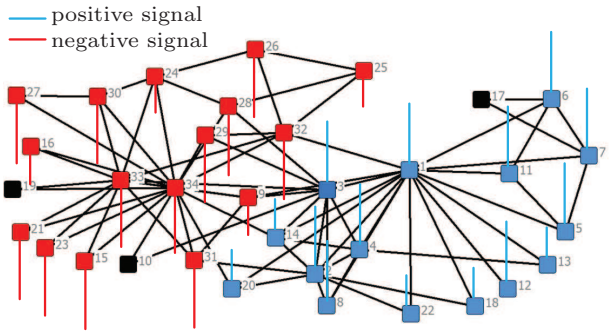


Fig. 2. Example of the Zachary club network. After an overall evaluation of each node, the Zachary club network can be represented as a graph with node signals on it. There is no signal on the black nodes. Red means negative signal, while blue means positive signal. The height is the absolute value of each signal.

3. Graph signal processing centrality (GSPC) method

In order to unite node attributes with the network topology together to identify influential nodes, we propose a graph signal processing centrality method in this section.

Signal processing on graphs focuses on the interplay between the graph topology and the graph signals.[26] For signal processing on graphs, a spectral graph theory has been explored as a tool to define expansion bases for graph Fourier transforms.[27] Many important mathematical ideas and intuitions can be extended from the classical Fourier analysis to the graph setting. By this means, main bases are picked up to do compression, filter, and so on.

Let $\boldsymbol{A}$ denote the adjacency matrix of a network with N vertexes, E edges, and a node evaluation set f, graph $G = (N, E, f)$. When a link is presented between two nodes v_i and v_j, both $A_{i,j}$ and $A_{j,i}$ are set to 1, otherwise they are set to 0. Since $\boldsymbol{A}$ is a real symmetric matrix and diagonalizable, it has a complete set of orthogonal eigenvectors denoted by $\boldsymbol{u}_l$ ($l = 1, \ldots, N$) and corresponding eigenvalues λ_l ($l = 1, \ldots, N$, $\lambda_1 \leq \lambda_2 \leq \cdots \lambda_N$), which satisfy $\boldsymbol{A}\boldsymbol{u}_l = \lambda_l \boldsymbol{u}_l$.

According to the classic Fourier transform, it turns the time domain signal into the frequency domain signals, which turns complex things into independent components. That is,

$$\tilde{f}(\xi) := \langle f, \mathrm{e}^{2\pi \mathrm{i} \xi t} \rangle = \int_R f(t)\, \mathrm{e}^{-2\pi \mathrm{i} \xi t}\, \mathrm{d}t. \tag{2}$$

According to the spectrum, the main frequency band is reserved by an appropriate filter. Thus, the main component of the spectrum is picked up, which could be used to reconstruct the original signal $f(t)$.

Analogously, treating the evaluation of node attributes as signals, the graph Fourier transform is utilized to find the principal eigenvectors of the adjacency matrix $\boldsymbol{A}$. Since the eigenvectors of the adjacency matrix are used to measure the topology importance of the nodes,[14–16,28] we define the graph Fourier transform to identify influential nodes as

$$\tilde{f}(\lambda_l) := \langle f, \boldsymbol{u}_l \rangle = \sum_{i=1}^{N} f(i) \boldsymbol{u}_l^*(i). \tag{3}$$

This graph Fourier transform gives us a way to represent both the network topology and the node attributes in the graph spectral domain. According the the Fourier analysis, the signals are compressible when the graph Fourier coefficients decay rapidly. If it is compressible, then several principal eigenvectors selected according to the signals in the spectral domain are used to represent the network under the node heterogeneity.

Thus, according to the spectrum, p appropriate eigenvectors are selected to represent the main information of the network topology. Choosing an appropriate p is important. We choose p analogous to that in the EVC or PCC method. The first idea is similar to the EVC method, we choose the eigenvector with the highest spectral value, that is,

$$p = \max\left(\tilde{f}(\lambda_p)\right). \tag{4}$$

This is suitable for the case where one spectral value is much more higher than the others. Otherwise, we list spectral values in descending order and then choose p according to

$$p = \max(\tilde{f}(\lambda_{p+1}) - \tilde{f}(\lambda_p)). \tag{5}$$

That is, if the next value $\tilde{f}(\lambda_{p+1})$ is much less than current one $\tilde{f}(\lambda_p)$, then the spectral values after the p-th would be ignored, and so are the corresponding eigenvectors. In this way, the principle eigenvectors are picked up.

Combing the node heterogeneity, the centrality of the nodes is calculated as

$$c_{N\times 1} = |f_{N\times 1}| \odot \sqrt{(\boldsymbol{u}_{N\times p} \odot \boldsymbol{u}_{N\times p})(\lambda_{p\times 1} \odot \lambda_{p\times 1})}, \quad (6)$$

where $\odot$ is the Hadamard operator.

Equation (6) states that the node centrality is determined by both the node signals and the network topology. The larger the node signal is, the higher the node centrality is. The more important the network position a node occupies, the higher the node centrality is.

We now analyz the time complexity. The GSPC method contains four parts step by step, that is, eigenvalues and eigenvectors calculation ($O(N^2)$), Fourier transform ($O(N\log N)$), selection ($O(N\log N)$), and centrality computing ($O(pN)$, where p is the number of selected eigenvectors and $p \ll N$). Then, the time complexity of the GSPC method is $O(N^2)$.

The time complexity of the most simple degree centrality is $O(E)$. For the most popular PageRank algorithm, its time complexity is $O(EL)$, where L stands for the iterations. The time complexity for the betweenness centrality is $O(NE)$. The time complexity of the recently proposed PCC algorithm is $O(N^2\log^2 N)$ or $O(N^2\log^3 N)$ to different networks topologies and PCC could be used in large-scale networks like Facebook data with 3097165 users.[16] The time complexity of the GSPC method is higher than the former two and lower than the latter two.

4. Simulation and discussion

In this section, we first identify influential nodes by our GSPC method in two small real networks with ground truth. These networks contain not only the network topology information but also the node attributes information. Then the compatibility and stability of our GSPC method are analyzed. At last, spreading analysis is implemented both in a larger synthetic network and a larger real social network.

4.1. Zachary club network

The Zachary club network[29] is a social network of friendships among 34 members of a karate club at a US university in the 1970s. During the observation period, for the disagreement between the administrator and instructor of the club, the club splits into two groups and one group leaves to start their own club. Mapping the network to the topological graph, the administrator and the instructor are represented by vertices 1 and 34, respectively.

Figure 3(a) shows the well known Zachary club network. It shows that two groups surround their own centers, nodes 1 and 34, respectively, and it has a very clearly divided community structure.

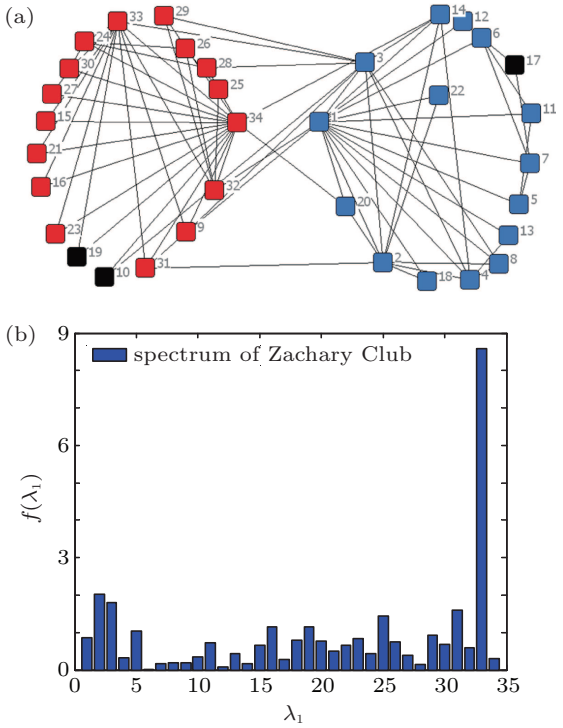


Fig. 3. (color online) (a) The network topology and (b) spectrum of the Zachary club network.

First, we obtain the node signal as follows. As shown in Table 2, the support attribute belongs to the sign category and determines the sign to be negative or positive. In this case, nodes are set as $+1$, -1, and 0 corresponding to their support on nodes 1, 34, and none, respectively. The strength attribute belongs to the ability category and determines the degree of support. Thus, the node evaluation is

$$f = \mathrm{sgn}(g_{\mathrm{support}}) \times |a_{\mathrm{strength}}|.$$

By the graph Fourier transform, we calculate $\tilde{f}(\lambda_l)$ and obtain the spectrum in Fig. 3(b). Given that the spectral value at λ_{33} is much higher than the rest, we set $p = 1$. Then we pick up λ_{33} and its corresponding eigenvector to rank the node centrality as

$$c_{N\times 1} = |f_{N\times 1}| \odot \sqrt{(\boldsymbol{u}_{33} \odot \boldsymbol{u}_{33})(\lambda_{33} \odot \lambda_{33})}.$$

The results in Table 2 show that the node centrality obtained by the GSPC method conforms to the real situation of the Zachary club network because the centrality of nodes 1 and 34 are the two highest. In addition, we can find out that the network topology based methods distinguish nodes 1 and 34, which may be due to the Zachary karate club displaying a divided community structure and hence a distinct topology.

Table 2. Node attributes, the corresponding node evaluation and comparison among the network topology based centrality methods and our GSPC method of the Zachary club network.

ID	Support	g_{support}	Strength	a_{strength}	$f(i)$	c_{degree}	c_{close}	c_{between}	c_{cluster}	PCC	EVC	GPSC
1	1	+1	2	2	2	**16**	**56.897**	**43.764**	**48.8483**	**9.423842**	**0.355491**	**3.850873**
2	1	+1	2	2	2	9	48.529	5.394	24.1363	4.991352	0.26596	2.677064
3	1	+1	2	2	2	10	**55.932**	14.366	37.5942	4.977279	0.317193	1.305582
4	1	+1	2	2	2	6	46.479	1.191	9.9104	3.602245	0.21118	2.51786
5	1	+1	2	2	2	3	37.931	0.063	4.9552	0.705172	0.075969	1.33283
6	1	+1	2	2	2	4	38.372	2.999	7.9057	0.811322	0.079483	1.449893
7	1	+1	2	2	2	4	39.372	2.999	7.9057	0.811322	0.079483	1.449893
8	1	+1	2	2	2	4	44.000	0.000	4.1000	2.403499	0.17096	2.079812
9	34	–1	1	1	–2	5	51.563	5.593	18.6574	2.41419	0.227404	0.273812
10	0	0	0	0	2	2	43.421	0.085	27	0.534194	0.102674	0
11	1	+1	2	2	2	3	37.931	0.063	4.9552	0.705172	0.075969	1.33283
12	1	+1	2	2	2	1	36.667	0.000	16	0.276036	0.052856	0.773772
13	1	+1	1	1	1	2	37.079	0.000	2.2000	0.730469	0.084255	0.639807
14	1	+1	1	1	1	5	51.563	4.586	14.5689	2.768084	0.226473	0.669317
15	34	–1	2	2	–2	2	37.079	0.000	2.9000	0.948361	0.101403	1.39029
16	34	–1	1	1	–1	2	37.079	0.000	2.9000	0.948361	0.101403	0.695145
17	0	0	0	0	0	2	28.448	0.000	0.8000	0.110134	0.023635	0
18	1	+1	1	1	2	2	37.500	0.000	2.5000	0.816276	0.0924	0.655801
19	0	0	0	0	0	2	37.079	0.000	2.9000	0.948361	0.101403	0
20	1	+1	1	1	1	3	50.000	3.248	19.4947	1.071002	0.147913	0.285212
21	34	–1	2	2	–2	2	37.079	0.000	2.9000	0.948361	0.101403	1.39029
22	1	+1	1	1	1	2	37.500	0.000	2.5000	0.816276	0.0924	0.655801
23	1	+1	1	1	1	2	37.079	0.000	2.9000	0.948361	0.101403	1.39029
24	34	–1	1	1	–1	5	39.286	1.761	15.9243	2.186356	0.150119	1.080259
25	34	–1	1	1	–1	3	37.500	0.221	6.0341	0.226071	0.057052	0.28077
26	34	–1	2	2	–2	3	37.500	0.384	6.4982	0.299713	0.059206	0.751388
27	34	–1	2	2	–2	2	36.264	0.000	2.1000	0.591516	0.075579	1.154335
28	34	–1	2	2	–2	4	45.833	2.233	23.8452	1.068969	0.133477	1.025778
29	34	–1	2	2	–2	3	45.205	0.179	15.3172	0.893935	0.131078	0.683328
30	34	–1	2	2	–2	4	38.372	0.292	7.7560	1.881037	0.134961	2.056314
31	34	–1	2	2	–2	4	45.833	1.441	13.5978	1.613071	0.174758	0.96244
32	34	–1	2	2	–2	6	54.098	13.828	34.0717	1.90971	0.191034	1.017658
33	34	–1	2	2	–2	12	51.563	14.525	38.7579	6.918462	0.308622	3.230677
34	34	–1	2	2	–2	**17**	55.000	**30.407**	**50.4219**	**9.707766**	**0.373362**	**3.688899**

4.2. Krack-High-Tec managers network

The Krack-High-Tec managers network[30] was collected from the managers of a high-tec company on the west coast of the United States. It has just over 100 employees with 21 managers. Mapping to topological graph, it has 21 nodes and the most important is node 7.

The topology is shown in Fig. 4(a) in which the nodes are too mixed with each other to tell the central node intuitively.

Table 3 involves four attributes of the managers: age (in years), tenure (length of time employed by the company, in years), level in corporate hierarchy (coded 1,2,3), department of the company (coded 0,1,2,3,4). These attributes all belong to the ability category and we use the modulated 2-norm model to obtain the node signal.

In particular, age and department attributes are not in accordance with the tenure and level attributes (contribute to role of an individual a lot), so their coefficients are set rather small. For the level attribute, the smaller, the better, we turn the values of level upside down, that is, $a_{\text{level}} = 4 - a_{\text{level}}$. Then the node signal is

$$f(i) = |0.05\boldsymbol{a}_{\text{age}} + 0.45\boldsymbol{a}_{\text{tenure}} + 0.45\boldsymbol{a}_{\text{level}} + 0.05\boldsymbol{a}_{\text{department}}|.$$

Then by the graph Fourier transform, we obtain the spectrum in Fig. 4(b). According to Fig. 4(b), the spectral value at λ_{21} exceeds the rest by a lot, thus $p = 1$ and λ_{21}, u_{21} is picked up. The node centrality is calculated as

$$c_{N\times 1} = f_{N\times 1} \odot \sqrt{(\boldsymbol{u}_{21} \odot \boldsymbol{u}_{21})(\lambda_{21} \odot \lambda_{21})}.$$

In this way, we obtain the node centrality of the Krack-High-Tec network shown in Table 3. It shows that the node centrality obtained by our GSPC method conforms the ground truth of the Krack-High-Tec managers because the centrality of node 7 is the highest. For network topology based methods, the centrality of node 15 is the highest, which suggests that node 15 is the most important person in the network and turns out to be false. This indicates that the node attributes are dispensable in identifying influential nodes in complex networks.

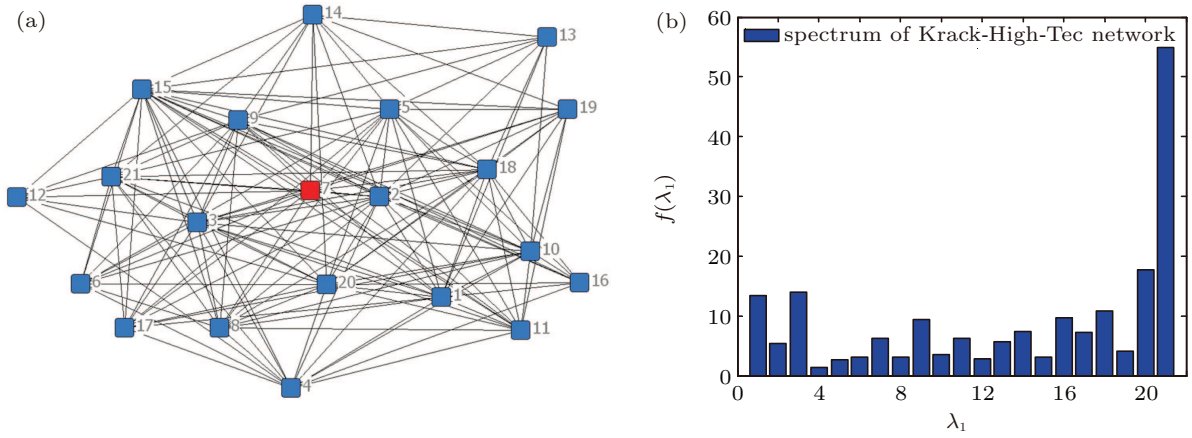


Fig. 4. (color online) (a) Krack-High-Tec network topology and (b) spectrum.

Table 3. Node attributes and comparison among the network topology based centrality methods and our GSPC method of the Krack-High-Tec network.

ID	Age	Tenure	Level	Dept	c_{degree}	c_{close}	$c_{between}$	$c_{cluster}$	PCC	EVC	GSPC
1	33	9.333	3	4	16	83.333	1.851	14.7649	12.75127	1.151004	27.96641
2	42	19.583	2	4	19	95.238	3.374	5.5776	16.85136	1.35962	54.07824
3	40	12.75	3	2	17	86.957	2.522	52.8649	14.11032	1.221511	35.93398
4	33	7.5	3	4	14	76.923	1.728	49.0277	11.11032	1.014661	22.01761
5	32	3.333	3	2	16	83.333	2.666	58.5232	13.54738	1.153372	17.61681
6	59	28	3	1	10	66.667	0.290	15	8.039628	0.734529	42.64824
7	55	30	1	0	14	76.923	1.827	41.7427	10.06142	1.015297	**56.94068**
8	34	11.333	3	1	14	76.923	0.722	26.5587	11.86671	1.009591	27.92019
9	62	5.417	3	2	16	83.333	3.415	56.7016	14.06929	1.155795	30.20686
10	37	9.25	3	3	15	80.000	1.556	40.0722	11.27648	1.082213	27.17873
11	46	27	3	3	13	74.074	0.568	13.9140	11.32891	0.940387	50.074253
12	34	8.917	2	1	7	60.606	0.171	2.8000	4.532365	0.529557	12.53320
13	48	0.25	3	2	8	62.500	0.149	17.6298	4.041031	0.597747	10.71188
14	43	10.417	2	2	11	68.966	0.659	7.3903	7.779661	0.805718	24.20278
15	40	8.417	3	2	**20**	**100.000**	**4.776**	**68.0254**	**18.14423**	**1.431469**	33.417183
16	27	4.667	3	4	9	64.516	0.160	11.8494	6.607813	0.665345	12.00284
17	30	12.417	3	1	11	68.966	0.439	14.9994	9.799927	0.802904	22.62137
18	33	9.083	2	3	17	86.957	2.274	52.1560	14.32734	1.220295	30.19475
19	32	4.833	3	2	11	68.966	0.348	29.6890	7.218864	0.803322	15.16665
20	38	11.667	3	2	17	86.957	3.101	53.4047	15.03659	1.222632	33.15351
21	36	12.5	2	1	15	80.000	1.617	38.8546	11.20849	1.08303	31.72183

4.3. Compatibility analysis

We first test the compatibility of our GSPC method to the eigenvector based centrality methods such as the PCC and EVC methods. Assume that the nodes are all homogeneous and the node signals are all set to be 1. Then, our GSPC method is degenerated to

$$v_{N\times 1} = \sqrt{(\boldsymbol{u}_{N\times p} \odot \boldsymbol{u}_{N\times p})(\lambda_{p\times 1} \odot \lambda_{p\times 1})}, \tag{7}$$

which is of the same form as that in the PCC method. And when $p = 1$, PCC equals a scaled version of EVC. This indicates that our GSPC method is compatible to both the PCC and EVC methods under different p selection and could be effective with only topological information.

The GPSC method under homogenous nodes is used to calculate the node centrality in the above two networks, which is then compared to the PCC and EVC methods. The parameter r_1 stands for the correlation coefficient, if it is more closer to 1, then the variables are more correlated; and r_2 for testing the hypothesis of no correlation, the smaller the better.

Table 4 shows the correlation of our GSPC with the PCC and EVC in the Zachary club network. Since we chose $p = 1$ in our GSPC method, the result equals to that of the EVC method and is highly correlated to that of the PCC method. We also find that the top five nodes are the same, $1, 2, 3, 33, 34$.

Similar to Table 4, Table 5 shows that our method is highly correlated with the PCC and EVC methods in the Krack-High-Tec network. We then find that the top five nodes are the same too, 2,3,15,18,20. However, the most important node 7 is not spotted, which suggests that the node attributes do affect the node centrality and are dispensable for the node centrality ranking.

Table 4. The correlation of our GSPC method with the PCC and EVC methods in the Zachary club network.

	GSPC vs. PCC	GSPC vs. EVC
r_1	0.9382	1.0000
r_2	0.0000	0.0000

Table 5. The correlation of our GSPC method with the PCC and EVC methods in the Krack-High-Tec network.

	GSPC vs. PCC	GSPC vs. EVC
r_1	0.9734	1.0000
r_2	0.0000	0.0000

4.4. Stability analysis

This subsection is mainly divided into two parts. First, we analyze the modulation of parameters of our GSPC model. Second, we analyze the stability of the topology and node signals.

In the first place, we vary the value of η to test the stability of our method satisfying the two rules aforementioned in Section 2. The test group includes $[0.05, 0.45, 0.45, 0.05]$, $[0.1, 0.4, 0.4, 0.1]$, and $[0.2, 0.3, 0.3, 0.2]$ in the Krack-High-Tec network. It turns out that as long as η for the important attributes are larger than the others, we can still maintain the same node centrality ranking order and correctly find the influential nodes.

We test the effect of p on the node centrality ranking. The results in Table 6 and Table 7 show that when p varies from 1 to 3, both in the Zachary club network and the Krack-High-Tec network, the influential nodes are picked out in accordance to the ground truth.

Then, we analyze the stability of topology and node signals. For a null model, we randomly shuffle the edges while conserving the degree distribution. We then shuffle the node signals. The results are shown in Tables 6 and 7, respectively.

Table 6. The results under varying p values and null models of our GSPC method in the Zachary club network.

ID	$p=1$	$p=2$	$p=3$	Edges shuffled			Signals shuffled		
				group 1	group 2	group 3	group 1	group 2	group 3
1	**3.850873**	**4.299519**	**5.51313**	0.072414	0.278841	**0.937754**	0.106504	0.118667	0
2	2.677064	2.778501	2.780522	0.181323	0.083753	0.691747	0.019232	0.079635	0.315771
3	1.305582	3.94108	4.027363	0.217123	0.106793	0.07166	0.039087	0.117277	0.17312
4	2.51786	2.5216	2.544286	0.214279	0.256613	0.283612	0.473427	0.109638	0.252959
5	1.33283	1.375128	1.517813	0.13178	0.164772	0.37297	0.011511	0.002208	0.162289
6	1.449893	1.505904	1.709865	0.084305	0.203551	0.213824	0.218223	0	0
7	1.449893	1.505904	1.709865	**0.604687**	0.025515	0.279857	0.109112	0.053938	0.031353
8	2.079812	2.097048	2.467846	**0.580132**	**0.666425**	0.303555	0	0.223824	0.300268
9	0.273812	0.416077	0.56275	0.158483	0.142367	0.094196	0.792949	0.473795	**0.557679**
10	0	0	0	0	0	0	0.102934	0.016944	0.010658
11	1.33283	1.375128	1.517813	0.489038	0.32089	0.346764	0.011511	0.004415	0.324577
12	0.773722	0.951961	1.461824	0.485135	0.347531	0.253823	0.146593	0.082175	0.374027
13	0.639807	0.705469	0.8648	0.111943	0.315458	0.137818	0.179222	0.003126	**0.976698**
14	0.669317	0.810788	0.960556	0.284985	0.053255	0.079608	0.41156	0.044454	0.235035
15	1.39029	1.390403	1.718669	0.157699	0.270076	0.120427	0.187129	0.13066	0.10778
16	0.695145	0.695201	0.859335	0.300528	0.128804	0.06946	0.187129	0.13066	0.10778
17	0	0	0	0	0	0	0.300365	0.074703	0.074697
18	0.655801	0.760551	0.951474	0.226524	0.066486	0.176218	0.120122	0.137321	0.378293
19	0	0	0	0	0	0	0.093564	0.06533	0
20	0.285212	0.291818	0.524875	0.201591	0.163587	0.089014	0.029423	0.042049	0.457278
21	1.39029	1.390403	1.718669	0.321209	**0.771564**	0.188137	0.187129	0.06533	0.10778
22	0.655801	0.760551	0.951474	0.055406	0.226908	0.148877	0.120122	0.068661	0.378293
23	1.39029	1.390403	1.718669	0.332278	0.44973	0.203964	0.187129	0.06533	0.05389
24	1.080259	1.162918	1.282248	0.335727	0.267655	0.087132	**1.086381**	0.179442	0.149123
25	0.28077	0.364015	0.393408	0.049587	0.014293	0.081891	0.323993	**0.825722**	0.302904
26	0.751388	1.061788	1.14234	0.057194	0.098761	0.181926	0	0.183843	0.033851
27	1.154335	1.316409	1.317985	0.423292	0.277133	0.22613	1.072856	0.683812	0.057241
28	1.025778	2.345697	2.386404	0.108934	0.194609	0.226198	0	0.464484	0.377039
29	0.683328	2.200206	2.200362	0.094312	0.238081	0.029407	0.248874	0.614439	0.211713
30	2.056314	2.056886	2.15482	0.366599	0.627223	0.253332	**1.298869**	0.728451	0.326688
31	0.96244	1.048185	1.608283	0.387569	0.090405	0.455105	0.706369	0.14229	0.234282
32	1.017658	1.621272	1.622698	0.286837	0.247125	0.166335	1.022392	**0.911763**	0.182198
33	3.230677	3.960855	4.39234	0.176104	0.385531	**0.959989**	0.265979	0	0.197976
34	**3.688899**	**4.31075**	**4.580168**	0.045914	0.056484	0.082502	0.130025	0	0.350714

Table 7. The results under varying p values and null models of our GSPC method in the Krack-High-Tec network.

ID	$p=1$	$p=2$	$p=3$	Edges shuffled			Signals shuffled		
				group 1	group 2	group 3	group 1	group 2	group 3
1	27.966416	28.47635	28.48135211	25.8445027	24.65963878	26.07307182	22.9031048	17.28050863	15.95844788
2	54.078242	54.14728	54.39154693	48.1878734	48.14869407	47.74763974	34.5218221	18.26630676	34.8180976
3	35.933985	36.45396	36.95384982	33.3946474	33.54975772	34.50351452	26.2008831	26.51576866	33.73905724
4	22.017606	22.09691	23.73962084	20.7075654	20.3343147	19.66946461	13.861327	28.24301391	51.78790362
5	17.616808	17.81233	19.07608945	15.0338772	16.33893412	15.84471554	22.8819165	29.24060695	28.00408052
6	42.648239	45.67402	52.51284264	40.2831239	41.21678046	39.10876301	20.1973386	38.35919254	20.47031161
7	**56.94068**	**59.1293**	**61.4200087**	**53.35207**	**53.5043618**	**53.3948814**	36.089897	21.53591568	25.37971725
8	27.92019	28.01968	29.14639046	26.5625637	26.44379686	26.69773187	**55.98474**	27.96803766	39.2401403
9	30.206858	30.31407	33.7299319	24.2712205	24.56290009	24.94032355	17.3316891	**56.2580461**	17.33168913
10	27.178735	28.22466	28.24531467	24.7789865	24.38243371	24.8684881	30.5593074	24.29259421	**53.1159853**
11	50.074253	50.41131	54.50170649	46.1940409	47.06689785	48.62730469	22.5174018	21.48419239	22.26174309
12	12.533199	14.91496	17.83447127	12.4830334	11.97714603	12.17554585	13.696503	13.99049648	10.76773065
13	10.711875	11.69747	12.01364483	7.08992669	7.352463344	7.277806526	9.42674447	23.08380418	13.45343666
14	24.202781	24.20997	26.41142652	20.3374089	19.17525081	19.05439612	21.7825853	17.04572335	12.37431031
15	33.417183	33.42244	33.94059087	29.9404627	29.948002	29.89758896	16.2853652	16.28536518	29.66662924
16	12.002844	12.71817	14.42239533	10.287036	10.49929119	10.62502498	15.4489574	37.81921051	16.53332429
17	22.62137	22.7295	26.5975199	21.5186463	21.73097171	21.91710165	19.1236231	19.75108921	45.17969814
18	30.194751	30.47732	30.68746642	24.8631146	24.76894539	24.40503618	62.7710586	24.11235573	29.15903608
19	15.166653	15.5027	15.73490915	13.3883717	13.60264092	13.05474242	20.5221544	20.52215435	17.63786531
20	33.153513	33.15352	35.1583761	31.6949399	31.1030841	31.84247287	27.0313019	18.81257944	26.47312424
21	31.721828	33.0662	33.14612244	29.0068932	28.64717258	27.11461729	52.8585959	27.24688651	12.96421342

For the Zachary club network, the most influential nodes always change in both conditions. As in both cases, the node signals in one community are of different signs. The sign of the node signal greatly affects the node centrality ranking.

For the Krack-High-Tec network, we find that the node attributes play a more important role in node centrality ranking. When we shuffle the edges randomly, we still identify node 7 to be the most influential, and when we shuffle the attributes randomly, the node with the attributes of original node 7 is found to be the most influential.

In other words, it indicates that both network topology and node heterogeneity play an important role in influential node identification. In networks with a divided community, like the Zachary club network, the sign attributes are essential because they chaos the network topology so that neighbor nodes may be opponents instead of supporters. In networks with homogeneous topology, like the Krack-High-Tec network, the node heterogeneity becomes much more important in influential node identification.

4.5. Spreading analysis

As the previous works[12,15] have pointed out, the influential nodes are verified by their efficiency in dynamics like spreading. That is, the nodes are taken as sources and their corresponding spreading efficiency is used to measure their influential importance. We apply the SI model[31,32] to test the spreading efficiency.

In the SI model, the nodes are classified into two states, infected or suspectable. At the beginning, some nodes are chosen as the source and in the infected state while all others are in the susceptible state. The source nodes will spread to all her neighbors. In every time step and for every link connected to the infected nodes, the suspectable nodes become infected at a fixed probability λ.

We use two datasets in this section. The first one is a synthetic network developed by a BA model with gradual aging. The BA model[33] captures the essential scale-free property of real networks by growth and preferential attachment. Since nodes have a finite lifetime or a finite capacity,[34] the probability a new node connecting to an old one is not only proportional to its degree but also depends on its age, decaying as $\tau^{-\gamma}$, $\gamma < 1$. Thus, with age decaying considered, the node signal is calculated as $\tau^{-\gamma}$, $\gamma = 0.5$.

We then apply a real social network, a Facebook like forum network[35] which originates from an online community for students at the University of California, Irvine. The dataset includes 899 users and the numbers of messages posted by the users. Then, the number of posted messages is referred to as the node signal.

By using the methods aforementioned and our GSPC method, we obtain the rank of nodes. Then several influential nodes are chosen as the source nodes for the SI model.

Figure 5 shows that in both cases, nodes spread faster where the sources are the most influential nodes chosen by our GSPC method compared to the other methods. As shown in Fig. 5(a), the GSPC method outperforms the other methods be-

cause it adds node attributes into consideration. The closeness and PCC methods gain a better performance than the other three. Since the BA model based networks have small cluster coefficients, the influential nodes obtained by the closeness method are scattered. The PCC methods identify influential nodes at a global view. The EVC method performs the worst, as the influential nodes under the EVC method gather in one restricted region of the network topology.

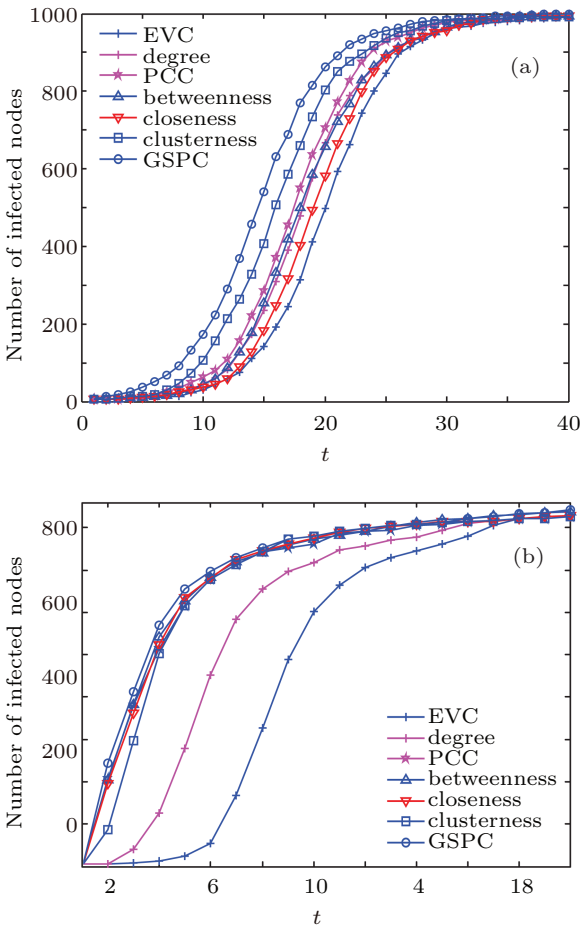


Fig. 5. (color online) SI spreading in a BA model based network and the Facebook like forum network. The results are the average over 200 independent realizations: (a) network size $N = 1000$, $\lambda = 0.3$; (b) network size $N = 899$, $\lambda = 0.3$.

As shown in Fig. 5(b), the GSPC method performs well in spreading. It is applicable in real social networks and indicates that the top five influential nodes are identified effectively. The EVC method results in bad performance, as the influential nodes gather in one restricted region of the network topology, which is the same as that in Fig. 5(a). Due to the positive correlation between the number of posted messages and the network topology in the Facebook like forum network, the influential nodes identified by PCC, betweenness , closeness, and cluster centrality are close to those by the GSPC method.

In summary, the GSPC method has a positive role in spreading. It gains the insight of balance between the network topology and the node heterogeneity. Thus, it is very useful in network dynamics where both the network topology and the node heterogeneity matter.

5. Conclusion

Identifying influential nodes for complex networks is an important task. A graph signal processing based centrality method is proposed in this paper, which combines the network topology and the node attributes together to distinguish the influential nodes. First a modulated 2-norm model is presented to obtain an overall evaluation of the node attributes. Then, the graph Fourier transform is utilized to select principal eigenvectors. Together, the node centrality is calculated with both the network topology and the node attributes. The results show that the GSPC method is effective in influential node identification and it is compatible to the previous eigenvector based centrality methods, the PCC and EVC methods. In addition, we verify the stability of the GSPC method by varying parameters and null models. Also, the spreading performance is upgraded.

Our GSPC method takes node heterogeneity into account and extends the scope of the node centrality calculation. This paper shows that our method can identify influential nodes in signed graphs (the Zachary club network). There are some possible applications. For example, many real complex networks can be regarded as weighted networks. Ranking node centrality in weighted networks attracts lots of interest. By converting the weights on edges into the graph signal or use the weight matrix other than the adjacency matrix for the graph Fourier transform, our method is easy to transplant to weighted networks. For many real networks, nodes with constraints (for example, limited online time in social networks, node with limited capacity in power grid) have a strong impact on the node behaviors. Our GSPC method could explore these constraints and calculate the node centrality combining the node properties with the network topology.

References

[1] Barabási A L 2007 *IEEE Control System Magazine* **27** 33
[2] Yeung C H and Saad D 2013 *J. Phys. A: Math. Theor.* **46** 103001
[3] Lu Y L, Jiang G P and Song Y R 2012 *Chin. Phys. B* **21** 100207
[4] Wu Y , Hu Y , He X H and Deng K 2014 *Chin. Phys. B* **23** 060101
[5] Fakhteh G and Konstantin K 2012 *Europhys. Lett.* **99** 58006
[6] Halu A, Zhao K , Baronchelli A and Bianconi G 2013 *Europhys. Lett.* **102** 16002
[7] Chen S M, Pang S P and Zou X Q 2013 *Chin. Phys. B* **22** 058901
[8] Leskovec J, Adamic L A and Huberman B A 2007 *ACM Transactions on the Web* **1** 5
[9] Goldenberg J, Han S, Lehmann D R and Hong J W 2009 *Journal of Marketing* **73** 2
[10] Sabidussi G 1966 *Psychometrika* **31** 581
[11] Freeman L C 1979 *Social Networks* **1** 215
[12] Chen D B, Lü L Y, Shang M S, Zhang Y C and Zhou T 2012 *Physica A* **391** 1777
[13] Bonacich P 2007 *Social Networks* **29** 555
[14] Page L, Brin S, Motwani R and Winograd T 1999 *Stanford InfoLab*

[15] Lü L Y, Zhang Y C, Yeung C H and Zhou T 2011 *PLoS ONE* **6** e21202
[16] Ilyas M U, Shafiq M Z, Liu A X and Radha H 2011 *INFOCOM, 2011 Proceedings IEEE* p. 561
[17] Liang N C, Chen P C, Sun T, Chen L J and Mario G 2006 *Systems, Man and Cybernetics 2006 IEEE International Conference on* p. 187
[18] Katiyar V, Chand N and Soni S 2011 *International Journal of Advanced Networking and Applications* **2** 4
[19] Yang Z C and John C S L 2011 *ACM SIGMETRICS Performance Evaluation Review* **39** 52
[20] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[21] Theodorakopoulos G and Baras J S 2006 *IEEE Journal on Selected Areas in Communications* **24** 318
[22] Cha M, Haddadi H, Benevenuto F and Krishna P 2010 *in ICWSM'10: Proceedings of international AAAI Conference on Weblogs and Social Media* p. 10
[23] Parantapa B, Muhammad B Z, Niloy G, Saptarshi G and Krishna G 2014 *ACM Recommender System Conference* to appear
[24] Malcolm G 2000 *The tipping point: How Little things can make a big difference* pp. 33–41
[25] Wu S, J M H, Mason W A and Watts D J 2011 *in Proc 20th Intl Conf WWW* pp. 705–714
[26] Shuman D I, Narang S K, Frossard P, Ortega A and Vandergheynst P 2013 *Signal Processing Magazine, IEEE* **30** 83
[27] Sandryhaila A and Moura J 2013 *Signal Processing, IEEE Transactions on* **61** 1644
[28] Bonacich P 1987 *American Journal of Sociology* 1170
[29] Zachary W 1977 *Journal of Anthropological Research* **33** 452
[30] Krackhardt D 1987 *Social Networks* **9** 109
[31] Barrat A, Barthelemy M and Vespignani A 2008 *Dynamical Processes on Complex Networks* (Cambridge: Cambridge University Press)
[32] Zhou T, Liu J G, Bai W J, Chen G R and Wang B H 2006 *Phys. Rev. E* **74** 056109
[33] Barabási A L and Albert R 1999 *Science* **286** 509
[34] Dorogovtsev S N and Menders J F F 2000 *Phys. Rev. E* **62** 1842
[35] Opsahl T 2010 *Social network* **35** 159