Spectral Learning Algorithm Reveals Propagation Capability of Complex Networks

Shuang Xu, Pei Wang, *Member, IEEE*, Chun-Xia Zhang, and Jinhu Lü, *Fellow, IEEE*

***Abstract*—In network science and the data mining field, a long-lasting and significant task is to predict the propagation capability of nodes in a complex network. Recently, an increasing number of unsupervised learning algorithms, such as the prominent PageRank (PR) and LeaderRank (LR), have been developed to address this issue. However, in degree uncorrelated networks, this paper finds that PR and LR are actually proportional to in-degree of nodes. As a result, the two algorithms fail to accurately predict the nodes' propagation capability. To overcome the arising drawback, this paper proposes a new iterative algorithm called SpectralRank (SR), in which the nodes' propagation capability is assumed to be proportional to the amount of its neighbors after adding a ground node to the network. Moreover, a weighted SR algorithm is also proposed to further involve a priori information of a node itself. A probabilistic framework is established, which is provided as the theoretical foundation of the proposed algorithms. Simulations of the susceptible-infected-removed model on 32 networks, including directed, undirected, and binary ones, reveal the advantages of the SR-family methods (i.e., weighted and unweighted SR) over PR and LR. When compared with other 11 well-known algorithms, the indices in the SR-family always outperform the others. Therefore, the proposed measures provide new insights on the prediction of the nodes' propagation capability and have great implications in the control of spreading behaviors in complex networks.**

***Index Terms*—Complex network, important node, influential spreader, propagation capability, SpectralRank (SR).**

Manuscript received May 13, 2017; revised April 5, 2018; accepted July 25, 2018. This work was supported by the National Key Research and Development Program of China under Grant 2016YFB0800401, in part by the National Natural Science Foundation of China under Grant 61773153, Grant 11671317, Grant 61621003, Grant 61532020, and Grant 11472290, in part by the Key Scientific Research Projects in Colleges and Universities of Henan under Grant 17A120002, and in part by the Basal Research Fund of Henan University under Grant yqpy20140049. This paper was recommended by Associate Editor J. Cao. *(Corresponding author: Pei Wang.)*

S. Xu and C.-X. Zhang are with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: shuangxu@stu.xjtu.edu.cn; cxzhang@mail.xjtu.edu.cn).

P. Wang is with the School of Mathematics and Statistics, Henan University, Kaifeng 475004, China, and also with the Institute of Applied Mathematics, Laboratory of Data Analysis Technology, Henan University, Kaifeng 475004, China (e-mail: wp0307@126.com).

J. Lü is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100083, China, also with the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100083, China, also with the Beijing Advanced Innovation Center for Big Data and Brain Machine Intelligence, Beihang University, Beijing 100083, China, and also with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jhlu@iss.ac.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2861568

I. Introduction

With the advent of the big data era [1]–[3], researchers are prone to focus on more complex data [4], where graph-based or system-based data have attracted much attention [5]. Graph-based data can be extracted from various fields, such as biology [6]–[11]; social technology [12], [13]; and industry [14]. This paper on machine learning methods on graph-based data provides complex network practitioners with plenty of tools in data mining [15]. Identifying important nodes [16], [17] is one of the most prevailing applications, such as estimating the nodes' propagation capability [18]; finding vital proteins or genes [9], [11], [19], [20]; mining network values of customers [21]; and pining control of multiagent systems [22].

In complex networks, ranking nodes according to their importance is an unsupervised learning problem. Node importance is equivalent to its propagation capability under many circumstances. Evaluation of the propagation capability of a node relies on spreading dynamics. Spreading dynamics on complex networks are so ubiquitous that the investigation of them might shed some light on controlling real-world networks [23]. To approximate the propagation capability, there are many traditional measures, including degree; k-core [24], [25]; H-index [26]; and many other Markov chain-based methods. In particular, degree [18] is a basic measure but of little relevance in many cases. The k-core decomposing algorithm [24] identifies the core and periphery of network, and assigns each node with a layer (called coreness number) by removing low degree nodes repeatedly. It has been reported that coreness outperforms degree to some extent [25]. Coreness can also be obtained by applying the H-index operation $\mathcal{H}$ [26] iteratively. The zero-order H-index is just the degree, while the steady state of the H-index operation is just the coreness. The degree, H-index, and coreness consist of the H-indices family, which provides pretty prediction results of node importance for complex networks.

Markov chain-based methods are another class of learning algorithms, where the most representative one is PageRank (PR) [27]. It assumes that an Internet surfer walks randomly on the web and chooses one of the web links stochastically. In the meantime, the surfer does not click on a hyperlink but jumps instead to a random web with a small probability. Even though PR is originally designed to rank webs, it has been applied to rank images, genes, and scientists [16]. Similar to the PR, the LeaderRank (LR) and its variants [28]–[30] are designed to mine the leaders in social networks. PR and LR have been widely used to evaluate the node's spreading influence [16].

However, they are misused in network science, especially in learning propagation capability. According to mean field analysis and empirical analysis in degree uncorrelated and correlated networks (see [31, App. A], and the supplementary material), it shows that PR and LR are proportional to the node's in-degree. This yields the conflict that in-degree actually can hardly extract the information of a node's propagation capability [26], [32].

To learn the nodes' propagation capability in complex networks, an effective algorithm should take the nature of spread dynamics into account. In this paper, we study a class of node ranking spectral algorithms. Our main contributions are as follows.

1) The parameter-free learning algorithms, called the SpectralRank (SR)-family, are proposed to elaborately measure the nodes' propagation capability. It is shown that indexes in the SR-family are closely related to the dominant eigenvector of the augment network, that is, the original network with a ground node.
2) In 32 representative networks (15 directed ones, 12 undirected ones, and 5 binary ones), we compare SR with 11 existing popular algorithms, including degree, H-index, coreness, mixed degree decomposition (MDD) [33], PR, LR, weighted LR (WLR), adaptive LR (ALR), ClusterRank (CluR) [32], eigenvector centrality (EC) [34], and cumulative nomination (CN) [35]. The obtained results reveal that the SR-family methods have advantages over the other measures in all of the considered networks.
3) We develop a probabilistic framework for a class of spectral-based algorithms, including EC, CN, and SR-family. We prove that the dominant eigenvector is the optimal solution under the probabilistic framework, which guides the application of spectral-based algorithms. Furthermore, this framework is able to explain why the addition of the ground node can enhance the performance of an algorithm.

This paper is organized as follows. Section II discusses the drawbacks of PR and LR in detail. Section III proposes the new algorithms and develops a probabilistic framework for node ranking problems. Results of some experiments are reported in Section IV. At last, discussions and conclusions are provided in Section V.

II. Preliminaries

A. Propagation Capability

Overwhelming evidence has revealed that different nodes and edges play heterogeneous roles in dynamics, control, evolution, and function [22], [36]. Propagation capability measures a node's spreading impact and it is defined as the number of infected nodes if we set a node as a single infection source. A node leading to a larger spreading scope has higher propagation capability. Nonetheless, this measure cannot be obtained unless disease breaks out. Hence, researchers have made great efforts to more efficiently learn the node's propagation capability. A promising way is to predict and estimate it by unsupervised learning algorithms, such as PR, LR, EC, and so on. For each algorithm, a score referring to the relative propagation capability is assigned to each node. A standard metric, Kendall τ correlation coefficient (as shown in the supplementary material), is used to quantify the accuracy of prediction algorithms [37]. We usually apply classical epidemic models to real networks so as to obtain a good approximation of the node's propagation capability. Accordingly, we employ Kendall τ between the propagation capability and algorithm score to estimate the accuracy. In the same network, an algorithm with higher τ indicates more accuracy and $\tau = 1$ means perfect prediction. Hence, τ is called algorithmic accuracy or simply accuracy.

B. Drawbacks of PageRank and LeaderRank

Consider a network $G(V, E)$, where V is the node set and E is the edge set. Meanwhile, $N = |V|$ and $M = |E|$ are the numbers of nodes and edges, respectively. The adjacency matrix $\mathbf{A}$ captures the wiring diagram, where the (i, j)th entry $a_{ij} = 1$ if node i points to j and 0 otherwise. For PR, each node is initially assigned an importance score $s_i(0) = 1(i = 1, 2, \ldots, N)$. Then, the score for node i is updated according to the following iterative process [27]:

$$s_i(t+1) = q\sum_{j=1}^{N} a_{ji}\frac{s_j(t)}{k_j^{\text{out}}} + (1-q)\frac{1}{N} \tag{1}$$

where k_j^{out} is the out-degree and q is a parameter, usually set as 0.85 [39]. When PR converges, the steady state $s(t_c)$ is employed to evaluate node importance.

However, PR is criticized for some drawbacks [39]. First, the ranking result of PR is not unique if the considered network has disconnected components. As a result, some schemes attempt to overcome such a defect, and one prevalent algorithm is the LR [28]. In LR, a ground node that connects all other nodes via bidirectional edges is added. Consequently, an augmented strongly connected network with $N + 1$ nodes and $M + 2N$ edges is obtained. The updating rule of node importance score is designed as [28]

$$s_i(t+1) = \sum_{j=1}^{N+1} a_{ji}\frac{s_j(t)}{k_j^{\text{out}}}. \tag{2}$$

Remark that s_g or s_{N+1} denotes the score for the ground node and we set $s_g(0) = 0$ for the ground node, $s_i(0) = 1(i = 1, 2, \ldots, N)$ for ordinary nodes. Different from PR, LR can ensure the uniqueness of a node's importance score.

Second, although LR improves PR, a growing number of empirical analyses reveals that both PR and LR may fail under some circumstances [32]. Currently, it is still questionable about whether it is proper to use them to estimate the node's propagation capability in all kinds of complex networks. Based on the mean field analysis, it is reported that k^{in} is a good approximation of PR when the network is degree uncorrelated [31]. In addition, there is a similar conclusion for LR.

Theorem 1: In a degree uncorrelated network, the average LR score for nodes within the node group with degree $\mathbf{k} =$

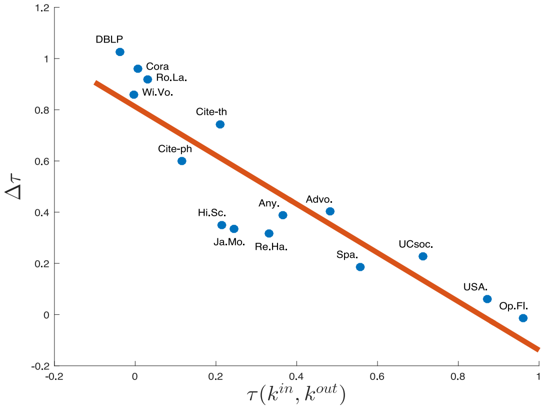


Fig. 1. k^{out} is a better indicator in the 15 directed networks. The fitted line is based on the ordinary least square regression. Here, $\tau(k^{\text{in}}, k^{\text{out}})$ is the Kendall correlation coefficient between k^{in} and k^{out}. $\Delta\tau = \tau_{k^{\text{out}}} - \tau_{k^{\text{in}}}$ is the accuracy difference of k^{out} and k^{in}.

$(k^{\text{out}}, k^{\text{in}})$ is proportional to k^{in}

$$\bar{s}_{\mathbf{k}}(t+1) \approx \theta k^{\text{in}}. \tag{3}$$

Here, $\theta = N/[(N+1)\langle k^{\text{in}}\rangle]$.

The proof of Theorem 1 can be found in Appendix A. As for degree correlated networks, we conduct an empirical analysis in the supplementary material, which demonstrates that correlations between k^{in} and PR / LR are statistically significant as well.

Nevertheless, k^{in} is not a good indicator of the node's propagation capability. Such a conclusion can be drawn by the accuracy difference index $\Delta\tau = \tau_{k^{\text{out}}} - \tau_{k^{\text{in}}}$ ($\tau_{k^{\text{out}}}$ and $\tau_{k^{\text{in}}}$ are prediction accuracy of k^{out} and k^{in}, respectively) in the 15 directed networks, as shown in Fig. 1. Fig. 1 indicates k^{in} is always inferior to k^{out} and this effect may be weakened only when k^{out} and k^{in} are strongly correlated. Thus, we conclude that k^{in} cannot effectively extract the information on propagation capability. At the same time, this phenomenon also implies that PR and LR fail to predict the node's propagation capability.

The drawbacks of PR and LR urge us to propose new effective algorithms to evaluate the node's propagation capability in any type of complex network.

III. New Algorithms

A. SpectralRank and Its Generlizations

1) SpectralRank: The nodes' propagation capability actually depends on their outgoing edges, which explains why PR and LR fail to learn nodes' propagation capability in networks where k^{in} and k^{out} sequences are not strongly correlated. In view of this point, methods that taking outgoing edges into account have been proposed, such as the EC and CN. But similar to the PR, they also have some drawbacks, such as nonunique rankings and dangling nodes. Consequently, it is urgent to develop a learning algorithm with universal applicability and high accuracy.

Hereinafter, we propose a new method called SR. In SR, each node is assigned a score representing its propagation capability. To cope with the above-mentioned problems, we insert a ground node (i.e., the node connects to all other nodes via bidirectional edges) to obtain a strongly connected network. Nodes with larger k^{out} have more successive neighbors. However, these successive neighbors play heterogeneous roles in propagation process. Hence, the node with more successive influential neighbors owns greater propagation capability. Specifically, we let the score of node i be proportional to the sum of its successive neighbors' score, that is,

$$\text{SR}_i = s_i = c\sum_{j=1}^{N+1} a_{ij}s_j. \tag{4}$$

The matrix formation of (4) follows:

$$\mathbf{s} = c\widetilde{\mathbf{A}}\mathbf{s} \tag{5}$$

where $\widetilde{\mathbf{A}}$ is the adjacency matrix of the augmented network

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} \tag{6}$$

and c is a tuning parameter and is usually selected as the reciprocal of dominant eigenvalue of $\widetilde{\mathbf{A}}$, that is $c = 1/\lambda_1$. On the basis of this fact, the SR score reduces to the dominant eigenvector of $\widetilde{\mathbf{A}}$.

The eigenvector can be easily obtained by the iterative power method. Initially, we set score $s_i(0) = 1 (i = 1, 2, \ldots, N)$ for ordinary nodes and $s_g(0) = s_{N+1}(0) = 0$ for the ground node. Then the iteration includes two operations, i.e., linear transformation and normalization. The updating rule is

$$\hat{s}_i(t+1) = \sum_{j=1}^{N+1} a_{ij}s_j(t),\ s_i(t+1) = \frac{\hat{s}_i(t+1)}{\max_k \hat{s}_k(t+1)}. \tag{7}$$

The matrix form can be written as

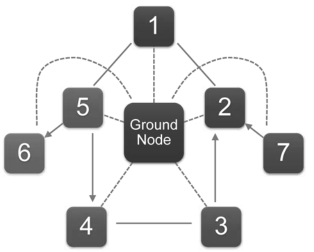
$$\hat{\mathbf{s}}(t+1) = \widetilde{\mathbf{A}}\mathbf{s}(t),\ \mathbf{s}(t+1) = \frac{\hat{\mathbf{s}}(t+1)}{\max \hat{\mathbf{s}}(t+1)}. \tag{8}$$

The Perron–Frobenius theorem guarantees the iteration process (7) to be converged within finite steps.

Remark 1: In (5), we intuitively set the tuning parameter c to be $1/\lambda_1$ without any technical proof. In Section III-B, we will build a probabilistic framework for EC and SR. It can be proved that $c = 1/\lambda_1$ is the optimal result under our probabilistic framework.

Remark 2: It is worth noticing that the SR procedures are similar to the EC whose updating rule follows $\mathbf{s}(t+1) = \mathbf{A}\mathbf{s}(t)$. The only difference is the iterative matrix in the former is augmented by a ground node. (As a summary, Fig. 2 demonstrates the relation and difference among PR, LR, EC, and SR.) But there are two technical questions. First, how do we guarantee that the ground node improves our algorithm's performance? Second, the topology of original network is changed by the ground node, so how do we guarantee that the SR scores stand for the importance of the nodes in the original network? These questions can be solved by our probabilistic framework in Section III-B.

Remark 3: It is feasible to add a ground node even for a large-scale network (N is very large) since the sparsity of the augmented network is $M/N^2 + 2/N$, which is very close to the sparsity of the original network, i.e., M/N^2. The cost of adding a ground node is very low.



Consider node 2

- PR: $\hat{s}_2(t+1) = q\left(s_1(t) + s_2(t) + s_7(t)\right) + \frac{1-q}{7}$;
- LR: $\hat{s}_2(t+1) = s_1(t) + s_2(t) + s_7(t) + s_g(t)$;
- EC: $\hat{s}_2(t+1) = s_1(t)$;
- SR: $\hat{s}_2(t+1) = s_1(t) + s_g(t)$.

Consider node 5

- PR: $\hat{s}_5(t+1) = qs_1(t) + \frac{1-q}{7}$;
- LR: $\hat{s}_5(t+1) = s_1(t) + s_g(t)$;
- EC: $\hat{s}_5(t+1) = s_1(t) + s_4(t) + s_6(t)$;
- SR: $\hat{s}_5(t+1) = s_1(t) + s_4(t) + s_6(t) + s_g(t)$.

Fig. 2. Toy network with seven nodes demonstrates the differences among PR, LR, EC, and SR. Directional edges are with arrows and bidirectional edges are without arrows. Solid lines represent the true topology edges and dashed lines represent the bidirectional edges between the ground node and original ones. Specially, nodes 2 and 5 are focused on which have large in-degree and out-degree, respectively.

Algorithm 1 WSR

```
Initialize importance score $\mathbf{s}(0)$ and set $t = 0$;
Construct the iteration matrix $\mathbf{W} = \widetilde{\mathbf{A}} + \mathbf{P}$;
repeat
    Update importance score $\hat{\mathbf{s}}(t+1) = \widetilde{\mathbf{A}}\mathbf{s}(t)$;
    Normalize importance score $\mathbf{s}(t+1) = \frac{\hat{\mathbf{s}}(t+1)}{\max \hat{\mathbf{s}}(t+1)}$;
    Set $t = t + 1$;
until change of $\mathbf{s}(t)$ is smaller than a predefined threshold
```

2) Weighted SpectralRank: Since SR assumes that the nodes propagation capability is proportional to the sum of its successive neighbors, it suffers the *boundary effect* to some extent. In other words, the SR may (not always) possibly underestimate a node as the best spreader and overestimate it as the worst spreader. When learning propagation capability, it is appropriate to consider information from a node itself. To achieve this, we can add the priori knowledge into SR. The iteration matrix $\widetilde{\mathbf{A}}$ can be replaced by $\mathbf{W} = \widetilde{\mathbf{A}} + \mathbf{P}$, where $\mathbf{P}$ is a diagonal matrix and its (i, i)th entry encodes our a priori knowledge of node i. In general, the weighted matrix $\mathbf{P}$ should be carefully selected. If we have no node information, $\mathbf{P}(i, i)$ can be set as 1 ($i = 1, 2, \ldots, N$). We can also consider results from some existing algorithms, such as degree k, H-index h, and coreness k_s. Notice that for the ground node, we always set $\mathbf{P}(N+1, N+1) = 0$. The new algorithm with weighted matrix $\mathbf{P}$ can be called weighted SR (WSR). Since different kinds of a priori information correspond to different WSR algorithms, we denote the WSR with degree k, H-index h, and coreness k_s as diagonal elements of $\mathbf{P}$ as WSR-k, WSR-h, and WSR-k_s, respectively. In the subsequent discussions, we use SR-family methods to refer to SR and WSR. Actually, SR can be viewed as a special case of the WSR, where all entities of $\mathbf{P}$ are zeros.

B. Probabilistic Explanation

In the above section, we proposed heuristic algorithms to predict the node importance. Now, we discuss the physical mechanism of the new algorithms in undirected networks. At first, we build a data-driven framework for the node ranking problem, which bridges the gap between the heuristic algorithms (EC and SR-family) and statistical theory. Then, we prove that $c = 1/\lambda_1$ is an optimal parameter in our framework. Last, we explain the reason why we should add a ground node and the weighted matrix $\mathbf{P}$ from the perspective of machine learning.

1) Data-Driven Framework for Node Ranking: For decades, researchers have proposed numerous heuristic node ranking algorithms, including PR, LR, EC, and so on, and empirical experiments demonstrated that they are very effective in many applications. Nonetheless, to our best knowledge, there is still a lack of literature explaining why these heuristic algorithms work.

Here, inspired by the preferential attachment and statistical mechanics, we build a novel framework to provide a theoretical understanding of the EC and SR-family. Preferential attachment is a classical growth model in complex network theory [36]. Namely, newly added nodes tend to connect to the nodes with a specific property, such as large degree nodes. This phenomenon has been proved by numerous evidence, one of which is called the Matthew effect or Gibrat's law, that is, the rich get richer. For example, the Barabási–Albert model [36] generates an undirected scale-free network, where the probability that a new node i connects to node j is proportional to the degree of j. Here, we consider a so-called fitness model [48], where the link between i and j is created with a probability $p(i, j)$. Specifically, we assume that $p(i, j)$ is proportional to the product of their importance scores, $p(i, j) \propto s_i s_j$, where $s_i \geq 0, \forall i \in V$. Let $\mathcal{A}$ denote the network space which contains all possible complex networks constructed by nodes in V. Thus, the probability of observing $\mathbf{A}$ is given by the Boltzmann distribution [49]

$$p(\mathbf{A}; \mathbf{s}) = \frac{e^{-H(\mathbf{A}; \mathbf{s})}}{Z_{\mathcal{A}}} \tag{9}$$

where the energy is given by the Hamiltonian function $H(\mathbf{A}; \mathbf{s}) = -\sum_{i,j \in V} a_{ij} s_i s_j$ and $Z_{\mathcal{A}} = \sum_{\mathbf{A} \in \mathcal{A}} p(\mathbf{A}; \mathbf{s})$ is the partition constant. Note that this distribution coincides with the Ising model without an external field [49].

In practice, our observed data is network $\mathbf{A}$ and our task is to infer unknown parameters $\mathbf{s}$, that is, importance score. This issue can be solved by the maximum likelihood principle, $\mathbf{s}^* = \arg\max_{\mathbf{s}} \ \log p(\mathbf{A}; \mathbf{s})$.

Theorem 2: The maximum likelihood estimate of importance score $\mathbf{s}^*$ in the fitness model (9) is exactly the EC of network $\mathbf{A}$. Furthermore, $c = 1/\lambda_1$ is the necessary condition of the maximum likelihood estimation.

Proof: The objective function can be rewritten as follows:

$$\begin{aligned} \mathbf{s}^* &= \arg\max_{\mathbf{s}} \ \log p(\mathbf{A}; \mathbf{s}) \\ &= \arg\max_{\mathbf{s}} \sum_{i,j \in V} a_{ij} s_i s_j \\ &= \arg\max_{\mathbf{s}} \ \mathbf{s}^T \mathbf{A} \mathbf{s}. \end{aligned} \tag{10}$$

Now, without loss of generality, we must add the constraint, $\mathbf{s}^T \mathbf{s} = 1$, otherwise the quadratic form $\mathbf{s}^T \mathbf{A} \mathbf{s}$ may go to the infinite. Therefore, the Lagrange function of our problem can be constructed as follows:

$$L = \mathbf{s}^T \mathbf{A} \mathbf{s} - \lambda\big(\mathbf{s}^T \mathbf{s} - 1\big) \tag{11}$$

where λ is the Lagrange multiplier. Let $\partial L/\partial \mathbf{s} = 0$ and we know $\mathbf{s}$ ought to satisfy $\mathbf{A}\mathbf{s} = \lambda\mathbf{s}$, implying that λ must be the eigenvalue of $\mathbf{A}$. Thus, the original problem is equivalent to

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} \ \lambda\mathbf{s}^T\mathbf{s}, \quad \text{subject to } \mathbf{s}^T\mathbf{s} = 1. \tag{12}$$

So, λ must be the largest eigenvalue of $\mathbf{A}$ and $\mathbf{s}^*$ is the dominant eigenvector. The proof is completed. ■

Theorem 2 provides a theoretical foundation of why we chose $c = 1/\lambda_1$ in (5). What is more, it gives new insightful understanding about the EC algorithm from the data-driven perspective. Besides, Theorem 2 also guides the application of EC, i.e., EC only works in the fitness model defined by (9).

2) Theoretical Foundation of Weighted Matrix and Ground Node: As stated in Section III-A, the main differences between EC and SR-family are the weighted matrix $\mathbf{P}$ and ground node. Despite empirical studies showing these two differences may improve the accuracy of ranking algorithms [28]–[30], [35], researchers still do not know the working mechanism. Here, from the viewpoint of Bayesian statistics, we study the roles that the weighted matrix $\mathbf{P}$ and ground node play.

In Bayesian statistics, we would encode our a priori knowledge by a priori distribution of $\mathbf{s}$. If we assume $\mathbf{s}$ is governed by the conjugate priori of (9)

$$p(\mathbf{s}) = \frac{1}{Z_{\mathcal{S}}} e^{-\mathbf{s}^T\mathbf{P}\mathbf{s}} \tag{13}$$

where $Z_{\mathcal{S}} = \int p(\mathbf{s})d\mathbf{s}$ is a partition constant, our task is to make the posterior estimate reach its maximum

$$\begin{aligned}\mathbf{s}^* &= \arg\max_{\mathbf{s}} \ \log p(\mathbf{s}|\mathbf{A})\\ &= \arg\max_{\mathbf{s}} \ \log p(\mathbf{A}|\mathbf{s}) + \log p(\mathbf{s})\\ &= \arg\max_{\mathbf{s}} \ \mathbf{s}^T\mathbf{A}\mathbf{s} + \sum_{i=1}^{N} s_i^2\mathbf{P}(i,i).\end{aligned} \tag{14}$$

From the above equations, we know that $\mathbf{P}$ brings a weighted L_2-norm penalty (also called the ridge penalty), which is widely used in machine learning and pattern recognition [50]. The effect of the L_2 norm penalty [or the priori $p(\mathbf{s})$] is to prevent over-fitting. In (14), $\mathbf{P}(i,i)$ is a penalty parameter. It is obvious that, based on optimization theory, larger/smaller $\mathbf{P}(i,i)$ leads to larger/smaller s_i.

As for the ground node, we have similar conclusions, which can be drawn if we replace $\mathbf{A}$ by $\widetilde{\mathbf{A}}$. Thus, the problem turns into

$$\begin{aligned}\left(\mathbf{s}^*, s_g^*\right) &= \arg\max \ \mathbf{s}^T\mathbf{A}\mathbf{s} + 2s_g\mathbf{s}^T\mathbf{1}\\ &\quad\text{subject to } \ \mathbf{s}^T\mathbf{s} + s_g^2 = 1\end{aligned} \tag{15}$$

whose Lagrange function is

$$L = \mathbf{s}^T\mathbf{A}\mathbf{s} + 2s_g\mathbf{s}^T\mathbf{1} - \lambda\left(\mathbf{s}^T\mathbf{s} + s_g^2 - 1\right). \tag{16}$$

Let $\partial L/\partial s_g = 0$; then, we have $\lambda s_g = \mathbf{s}^T\mathbf{1}$. After plugging this equation into our objective function, we have

$$\mathbf{s}^* = \arg\max \ \mathbf{s}^T\mathbf{A}\mathbf{s} + \mathbf{s}^T\left(\frac{2}{\lambda}\mathbf{1}^T\mathbf{1}\right)\mathbf{s}. \tag{17}$$

Obviously, ground node implies that we have a priori knowledge about importance score, which is encoded by $p(\mathbf{s}) = (1/Z)\exp(-2\mathbf{s}^T\mathbf{1}^T\mathbf{1}\mathbf{s}/\lambda)$ and, from the perspective of machine learning, the addition of the ground node can prevent over-fitting.

In summary, although $\mathbf{P}$ and the ground node change the topology of the original network, the (W)SR scores still stand for the importance of nodes in the original network. But a natural question is whether the a priori information is correct, or in other words, whether the weighted matrix $\mathbf{P}$ and ground node improve the accuracy of a node identification algorithm. In the next section, the empirical studies show that the two over-fitting restriction strategies indeed enhance the eperformances of algorithms.

IV. Experiments

A. Prediction of Propagation Capability

Thirty two representative real-world networks, including 15 directed networks, 12 undirected ones, and 5 binary ones, are considered. The 32 networks cover social, biological, technological, and transportation fields, with the numbers of nodes ranging from tens to tens of thousands. Detailed information about the 32 networks can be found in the supplementary material.

SIR model is applied to 32 representative real-world networks to obtain the propagation capability of nodes therein and the Kendall τ correlation coefficient is employed to evaluate algorithm accuracy. To show the superiority of the new learning algorithms, we selected 11 existing algorithms as benchmarks, including degree k, H-index h, coreness k_s, MDD, LR, WLR, ALR, CluR, PR, CN, and EC.

We introduce two measures to evaluate the performance of the proposed algorithms. The first one is the median accuracy τ_m; it is the median of an algorithm's Kendall τ over all considered networks, measuring the average performance of algorithm x. That is, $\tau_m(x) = \text{median}_i \tau_i(x)$, where i and x are the indices of network and algorithm, respectively. $\tau_i(x)$ represents the accuracy of the algorithm x in network i. The second index is the relative loss L, defined as $L(x) = (1/n)\sum_{i=1}^{n} \tau_i(x) - \tau_i^{\text{opt}}$, where n is the number of the considered networks, and $\tau_i^{\text{opt}} = \max_x \tau_i(x)$ is the best accuracy among the considered algorithms for network i. In this way, an algorithm with larger L indicates that it achieves better performance.

Table I lists the accuracy of the considered algorithm on selected networks (for all the 32 networks, refer to the supplementary material). Table II reports the overall performance as assessed by τ_m and L. It can be observed that the SR-family methods show excellent predicting ability among all types of networks, where most of the optimal results are offered by the SR-family approaches. The SR-family methods also tend to have larger τ_m and L. The PR and LR-family always lead to the worst prediction results in all types of networks. This implies that the in-linkage-based methods can poorly predict the propagation capability, especially in degree uncorrelated networks. As an example, Fig. 3 shows the colormap for the James

TABLE I
ACCURACY MEASURED BY τ FOR THE 16 ALGORITHMS IN 15 SELECTED NETWORKS. WE LISTED THE RESULTS FOR 5 BINARY, 5 UNDIRECTED, AND 5 DIRECTED NETWORKS, FOR THE RESULTS OF ALL THE 32 NETWORKS, SEE SUPPLEMENTARY INFORMATION. THE BEST AND THE SECOND BEST PREDICTIONS ARE SHOWN IN BOLD FACE AND WITH UNDERLINE, RESPECTIVELY. NAMES OF SOME NETWORKS ARE REPLACED BY THEIR ABBREVIATIONS. INSTEAD OF k, h, AND k_s IN UNDIRECTED AND BINARY NETWORKS, k^{out}, h^{out}, AND k_s^{out} ARE CONSIDERED IN DIRECTED ONES

Algo.	Binary networks					Undirected networks					Directed networks				
	Am.Re.	Le.sh.	SE	WB	WN	Ham.	Email	David	USAir	NS	Advo.	Ja.Mo.	Wi.Vo.	DBLP	Any.
k	0.5196	0.8049	0.5121	0.5431	0.5782	0.7032	0.7794	0.8374	0.7256	0.5092	0.8180	0.5785	0.8162	0.9877	0.6515
h	0.5111	0.7206	0.5439	0.5700	0.5985	0.7359	0.8103	0.8525	0.754	0.5178	0.8296	0.6100	0.7848	0.8818	0.6925
k_s	0.5100	0.7008	0.5558	0.5719	0.6006	0.7445	0.8021	0.8158	0.7529	0.4747	0.8276	0.6053	0.7786	0.8082	0.6975
MDD	0.5196	0.8217	0.5208	0.5476	0.5823	0.7127	0.7893	0.8444	0.7308	0.5199	0.8063	0.7445	0.7643	0.8081	0.6772
LR	0.4536	0.7154	0.3620	0.3031	0.6206	0.6612	0.7440	0.7992	0.6697	0.4541	0.2832	0.2201	-0.0507	-0.0550	0.3903
WLR	0.2824	0.8299	0.6382	0.5402	0.5708	0.7399	0.7959	0.8478	0.7719	0.5707	0.3231	0.2444	-0.0479	-0.0408	0.3922
ALR	0.5651	0.8172	0.7060	0.6973	0.6600	0.7938	0.8306	0.87	0.8106	0.5672	0.4554	0.3353	0.2673	0.4359	0.4667
CluR	0.7860	0.8961	**0.7790**	0.7908	0.7551	0.7631	0.7347	0.7721	0.6061	0.4644	0.8582	0.7663	0.8195	0.9462	0.7721
PR	0.3647	0.6391	0.1707	0.3448	0.3542	0.4875	0.6828	0.7535	0.5371	0.3357	0.6284	0.3802	0.7685	0.9324	0.6483
CN	0.7931	0.8468	0.7521	0.8064	0.7559	0.8203	0.8202	0.8729	0.822	0.5273	0.9105	0.5188	0.8052	0.9169	0.8820
EC	0.7042	0.8574	0.7357	0.7916	0.7529	0.8202	0.8202	0.8733	0.822	0.5203	0.9083	0.5172	0.7780	0.7899	0.8786
SR	**0.8563**	0.9062	0.6730	0.8256	0.7680	0.8402	0.8295	0.8987	0.8285	0.6209	0.8852	**0.8426**	0.8209	0.9906	0.8820
WSR-k	0.7903	0.8702	0.7363	0.7178	0.7233	0.7303	0.7573	0.7738	0.7655	**0.6867**	0.8278	0.7186	0.5356	0.6592	0.8571
WSR-h	0.8311	0.8850	0.6738	**0.8263**	**0.7756**	0.8349	**0.8476**	0.8858	0.8222	0.6251	**0.9130**	0.7742	0.5610	0.6668	0.8836
WSR-k_s	0.8311	0.9062	0.6693	0.8250	0.7746	**0.8445**	0.8393	**0.9**	0.8252	0.6135	0.8911	0.8211	0.5188	0.6854	**0.8909**
WSR-1	0.8560	**0.9083**	0.6720	0.8256	0.7692	0.8405	0.8303	0.8977	**0.8286**	0.6237	0.9107	**0.8426**	**0.8229**	**0.9907**	0.8821

TABLE II
TWO METRICS OF THE 16 ALGORITHMS. THE BEST AND THE SECOND BEST PREDICTIONS ARE SHOWN IN BOLD FACE AND WITH UNDERLINE, RESPECTIVELY

Algo.	τ_m-D	L-D	τ_m-U	L-U	τ_m-B	L-B
k	0.7517	-0.1158	0.6976	-0.1318	0.5431	-0.1380
h	0.7638	-0.1307	0.7219	-0.1132	0.57	-0.1280
k_s	0.6975	-0.1924	0.7252	-0.1275	0.5719	-0.1446
MDD	0.7260	-0.1531	0.7042	-0.1422	0.5476	-0.1418
LR	0.2201	-0.6559	0.6516	-0.3051	0.4536	-0.3211
WLR	0.2444	-0.6211	0.7338	-0.2196	0.5708	-0.2624
ALR	0.3752	-0.4496	0.7959	-0.1514	0.6973	-0.1746
CluR	0.7663	-0.1188	0.7286	-0.0948	0.786	-0.0713
PR	0.5730	-0.2729	0.4684	-0.3036	0.3542	-0.2830
CN	0.7369	-0.1424	0.8133	-0.0456	0.7931	-0.0562
EC	0.7369	-0.1566	0.8132	-0.0482	0.7529	-0.0661
SR	**0.8820**	-0.0096	0.8290	-0.0427	0.8256	-0.0210
WSR-k	0.7038	-0.1691	0.7199	-0.1000	0.7363	-0.0897
WSR-h	0.7742	-0.0958	0.8104	-0.0531	**0.8263**	-0.0421
WSR-k_s	0.8211	-0.0864	0.8029	-0.0605	0.825	-0.0486
WSR-1	0.8815	**-0.0067**	**0.8295**	**-0.0324**	0.8256	**-0.0194**

Moody network. It is observed that the colormap of SR can match that of spread range. Nevertheless, PR and LR failed. Furthermore, it can be seen that one of the WSRs may outperform the SR in a certain network, but the SR always adapts to all kinds of cases and offers outstanding results. For instance, WSR-*h* offers the best result in Advogato with $\tau = 0.9130$ and in Email with $\tau = 0.8476$; while for SR, whose accuracy for the two networks are $\tau = 0.8852$ and $\tau = 0.8295$, respectively, has similar performance with WSR-*h*. Nevertheless, the performances of WSR-*h* are far beyond that of SR in Cora and WikiVote (see Table I). Furthermore, Table II reports that the performance of SR is the best on average.

B. Application in Biological Networks

Recently, identification of key nodes (e.g., genes, proteins) in biological networks has attracted much attention [19], [20]. In the following, two real-world directed biological networks are considered, including the *C. Elegans* Neural [40] (CEN) network with 280 nodes and 2194 edges, and the *E. Coli* Transcriptional [41] (ECT) regulatory network with 1706 nodes and 3870 edges. It has been known that 10 command interneurons (AVER, AVEL, AVAR, AVBL, AVBR, AVAL, AVDL, PVCR, AVDR, PVCL) in CEN; 18 global regulators (fnr, crp, fis, fur, mlc, ompR, cpxR, hns, arcA, narL, soxR, soxS, purR, lrp, rob, phoB, CspA, IHF); and 7 key global regulators (fnr, crp, fis, arcA, narL, lrp, IHF) in ECT are the key nodes, and they play profound biological roles in normal life activities [19]. By taking the mentioned key nodes as gold standards, and in order to evaluate the accuracy of an algorithm, ROC analysis is employed (see Appendix B). We can find that degree, PR, *H*-index, and MDD may outperform in certain cases, but SR gets the fourth, third, and first positions in the three cases, respectively (see Fig. 4). Overall, SR achieves a better balance between precision and generalization.

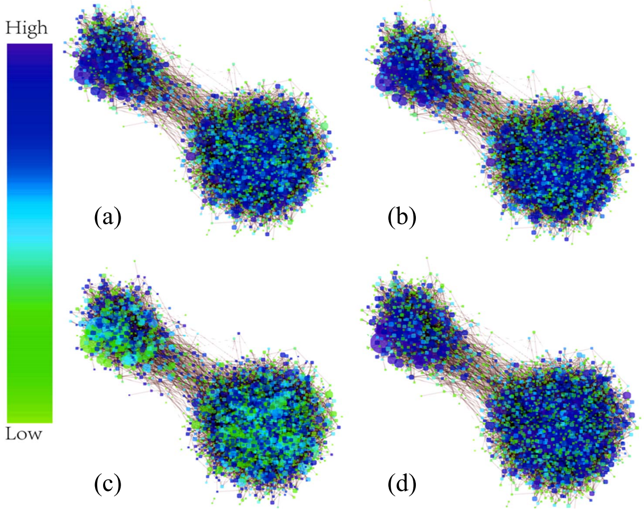


Fig. 3. Colormaps for the JamesMoody network. Node sizes are proportional to degrees. (a)–(d) Node colors are proportional to real spread ranges, SR values, PR values, and LR values, respectively.

The examples suggested that the proposed SR can also be applied to biological networks, which will help us to robustly find key nodes in bio-molecular networks.

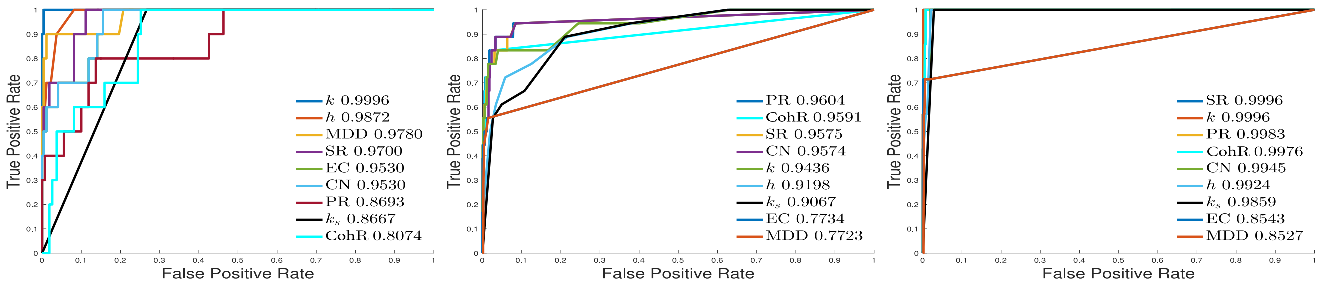


Fig. 4. ROCs and areas under curves (AUCs) for the CEN and ECT. (Left) Command interneurons in CEN as important nodes. (Middle) Global regulators in ECT as important nodes. (Right) Key global regulators in ECT as important nodes. The numbers following the legends are AUCs, and algorithms are sorted by AUCs.

V. Conclusion

Aided by the advanced data mining and machine learning techniques, the applications of complex networks have made great progress in various domains. For instance, this paper of coarse-graining of complex networks [42]; link prediction [43]; recommender systems [44]; community identification [45], [46]; and vital node mining [16]. This paper concentrates on identification of super spreaders, i.e., prediction of the nodes' propagation capability. PR and LR are two popular algorithms. Nevertheless, some evidence has shown that they may fail in some situations. Taking both the merits and the drawbacks of PR and LR into account, the SR algorithm is proposed. The proposed SR reveals that the nodes' propagation capability depends on the spectrum λ_1 of augmented network $\widetilde{\mathbf{A}}$. The simulations of SIR model on 32 real networks reveal that SR is very competitive compared with some other 11 renowned algorithms.

We established a probabilistic framework for the node ranking problem. Under this framework, we provide a theoretical foundation of the parameter chosen in the spectral-based algorithms. Our framework also tells us that spectral-based algorithms are the maximum likelihood or maximum *a posteriori* estimates in the fitness network. Note that researchers' empirical studies show that ground node can improve the algorithm's performance. Nevertheless, no literature can explain the working mechanism of the ground node. Our framework gives a theoretical foundation with the addition of the ground node. Namely, it encodes a priori knowledge that helps our algorithms to prevent over-fitting.

The proposed SR can not only be applied to any type of complex network but also shows superiority in the identification of key neurons and transcription factors in biological networks. In view of these facts, we conclude that SR can help us to better understand the pattern of spread dynamics and to better identify important nodes in various complex networks.

Appendix A
Proof of Theorem 1

We apply the mean field analysis to LR. Nodes can be classified by its degrees, i.e., nodes in the same class $\mathbf{k}$ have the same out-degree and in-degree $(k^{\text{out}}, k^{\text{in}})$. We consider the average LR score for a class of nodes

$$\bar{s}_{\mathbf{k}}(t+1) \equiv \frac{1}{\text{NP}(\mathbf{k})} \sum_{i \in \mathbf{k}} s_i(t+1) \tag{18}$$

where $P(\mathbf{k})$ denotes the frequency of node with degree $\mathbf{k} = (k^{\text{out}}, k^{\text{in}})$. The updating rule of LR follows:

$$s_i(t+1) = \sum_{j=1}^{N+1} a_{ji} \frac{s_j(t)}{k_j^{\text{out}}}. \tag{19}$$

Hence, we have

$$\bar{s}_{\mathbf{k}}(t+1) = \frac{1}{\text{NP}(\mathbf{k})} \sum_{i \in \mathbf{k}} \sum_{j=1}^{N+1} a_{ji} \frac{s_j(t)}{k_j^{\text{out}}}. \tag{20}$$

Notice that the nodes in the same class have the same out-degree. So in the right side of (20), we split the sum over j into two sums, one over all the degree classes $\mathbf{k}'$ and the other over all the nodes within each degree class $\mathbf{k}'$

$$\bar{s}_{\mathbf{k}}(t+1) = \frac{1}{\text{NP}(\mathbf{k})} \sum_{\mathbf{k}'} \frac{1}{k'^{\text{out}}} \sum_{i \in \mathbf{k}} \sum_{j \in \mathbf{k}'} a_{ji} s_j(t). \tag{21}$$

We apply the mean field theory, i.e., it is assumed that the LR scores of node i's predecessors that belong to class $\mathbf{k}'$ can be replaced by the mean value of LR scores for class $\mathbf{k}'$

$$\sum_{i \in \mathbf{k}} \sum_{j \in \mathbf{k}'} a_{ji} s_j(t) \simeq \bar{s}_{\mathbf{k}'}(t) \sum_{i \in \mathbf{k}} \sum_{j \in \mathbf{k}'} a_{ji} = \bar{s}_{\mathbf{k}'}(t) E_{\mathbf{k}' \to \mathbf{k}} \tag{22}$$

where $E_{\mathbf{k}' \to \mathbf{k}}$ denotes the amount of edges from nodes within class $\mathbf{k}'$ to nodes within class $\mathbf{k}$

$$E_{\mathbf{k}' \to \mathbf{k}} = k^{\text{in}} P(\mathbf{k}) N \frac{E_{\mathbf{k}' \to \mathbf{k}}}{k^{\text{in}} P(\mathbf{k}) N} = k^{\text{in}} P(\mathbf{k}) \text{NP}_{\text{in}}\left(\mathbf{k}' | \mathbf{k}\right) \tag{23}$$

where $P_{\text{in}}(\mathbf{k}'|\mathbf{k})$ is the frequency that the start node of an edge is with degree $\mathbf{k}'$, but the end node of the edge has degree $\mathbf{k}$. Therefore, we have

$$\bar{s}_{\mathbf{k}}(t+1) = \sum_{\mathbf{k}'} \frac{k^{\text{in}} P_{\text{in}}\left(\mathbf{k}'|\mathbf{k}\right) \bar{s}_{\mathbf{k}'}(t)}{k'^{\text{out}}}. \tag{24}$$

If the network is uncorrelated (ground node does not affect the degree correlation of the original network), the conditional degree distribution $P_{\text{in}}(\mathbf{k}'|\mathbf{k})$ does not depend on $\mathbf{k}$ and we have

$$P_{\text{in}}\left(\mathbf{k}'|\mathbf{k}\right) = \frac{k'^{\text{out}} P(\mathbf{k}')}{\langle k^{\text{in}} \rangle} \tag{25}$$

where $\langle k^{\text{in}} \rangle$ denotes the average in-degree. Furthermore, we have

$$\bar{s}_{\mathbf{k}}(t+1) = \frac{k^{\text{in}}}{\langle k^{\text{in}} \rangle} \sum_{\mathbf{k}'} \bar{s}_{\mathbf{k}'}(t) P(\mathbf{k}'). \tag{26}$$

According to the definition of statistical expectation, we have $\sum_{\mathbf{k}'} \bar{s}_{\mathbf{k}'}(t)P(\mathbf{k}') = E[\bar{s}_{\mathbf{k}'}(t)]$. Notice that $E[\bar{s}_{\mathbf{k}'}(t)] = N/(N+1)$. Thus, when N tends to infinity

$$\bar{s}_{\mathbf{k}}(t+1) \approx \frac{k^{\text{in}}}{\langle k^{\text{in}} \rangle} \frac{N}{N+1} = \theta k^{\text{in}}. \tag{27}$$

Here, θ is a constant. Within finite steps, the iteration process converges, so the average LR score for nodes within class $\mathbf{k} = (k^{\text{out}}, k^{\text{in}})$ is proportional to k^{in}.

Appendix B
ROC Curve Analysis

We call the classification of node i determined by score s_i through an algorithm the prediction standard, and the real classification as the gold standard. Thresholds s_t are chosen and nodes with a score larger than s_t are classified as important, and unimportant otherwise. Moreover, we define the false positive rate and true positive rate as

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{28}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{29}$$

where FP denotes the number of false positive nodes (i.e., those nodes are actually unimportant but predicted as important). TN denotes the number of true negative nodes (i.e., actually important but predicted as unimportant). TP and FN denote the number of true positive and false negative nodes, respectively. Given a s_t, a point $(\text{FPR}_t, \text{TPR}_t)$ can be obtained. After taking a series of different thresholds, an ROC curve [19] can be obtained. A larger area under the ROC curve means more precise prediction results.

Appendix C
SIR Model and Parameter Settings

The SIR model [38] is employed to evaluate the nodes' spreading scope. There are three possible states for each node, that is, susceptible, infected, and recovered. The susceptible node may be infected by its infected neighbors with probability β, and the infected ones may recover with probability μ. To obtain the propagation capability of node i, we set i as a single infection seed and count the number of recovered nodes at the steady state of the SIR process. For each node, we average the propagation capability over 100 independent simulation runs. The spreading rate β is shown in Table S1 in the supplementary material, and the recover rate μ is set as 1.

References

[1] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Nat. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.

[2] G. Wen, W. Yu, Z. Li, X. Yu, and J. Cao, "Neuro-adaptive consensus tracking of multiagent systems with a high-dimensional leader," *IEEE Trans. Cybern.*, vol. 47, no. 7, pp. 1730–1742, Jul. 2017.

[3] G. Wen, T. Huang, W. Yu, Y. Xia, and Z.-W. Liu, "Cooperative tracking of networked agents with a high-dimensional leader: Qualitative analysis and performance evaluation," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2060–2073, Jul. 2018.

[4] B. Yoon and Y. Park, "A text-mining-based patent network: Analytical tool for high-technology trend," *J. High Technol. Manag. Res.*, vol. 15, no. 1, pp. 37–50, 2004.

[5] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.

[6] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nat. Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, 2009.

[7] J. A. Dunne, R. J. Williams, and N. D. Martine, "Food-web structure and network theory: The role of connectance and size," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 20, pp. 12917–12922, 2002.

[8] Z.-H. You, M. Zhou, X. Luo, and S. Li, "Highly efficient framework for predicting interactions between proteins," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 731–743, Mar. 2017.

[9] Z. Wang *et al.*, "Multi-gene co-transformation can improve comprehensive resistance to abiotic stresses in B napus L," *Plant Sci.*, vol. 274, pp. 410–419, Sep. 2018.

[10] P. Wang, D. Wang, and J. Lü, "Controllability analysis of a gene network for Arabidopsis thaliana reveals characteristics of functional gene families," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, to be published, doi: 10.1109/TCBB.2018.2821145.

[11] P. Wang *et al.*, "Transcriptomic basis for drought-resistance in Brassica napus L," *Sci. Rep.*, vol. 7, Jan. 2017, Art. no. 40532.

[12] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[13] G. Mei *et al.*, "Compressive-sensing-based structure identification for multilayer networks," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 754–764, Feb. 2018.

[14] G. A. Pagani and M. Aiello, "The power grid as a complex network: A survey," *Physica A Stat. Mech. Appl.*, vol. 392, no. 11, pp. 2688–2700, 2013.

[15] M. Zanin *et al.*, "Combining complex networks and data mining: Why and how," *Phys. Rep.*, vol. 635, pp. 1–44, May 2016.

[16] L. Lü *et al.*, "Vital nodes identification in complex networks," *Phys. Rep.*, vol. 650, pp. 1–63, Sep. 2016.

[17] W. Liu *et al.*, "Mining top K spread sources for a specific topic and a given node," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2472–2483, Nov. 2015.

[18] P. Wang, C. Tian, and J. Lu, "Identifying influential spreaders in artificial complex networks," *J. Syst. Sci. Complex.*, vol. 27, no. 4, pp. 650–665, 2014.

[19] P. Wang, J. Lü, and X. Yu, "Identification of important nodes in directed biological networks: A network motif approach," *PLoS ONE*, vol. 9, Aug. 2014, Art. no. e106132.

[20] P. Wang, X. Yu, and J. Lü, "Identification and evolution of structurally dominant nodes in protein-protein interaction networks," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 1, pp. 87–97, Feb. 2014.

[21] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, San Francisco, CA, USA, 2001, pp. 57–66.

[22] W. Xu, D. W. C. Ho, L. Li, and J. Cao, "Event-triggered schemes on leader-following consensus of general linear multiagent systems under different topologies," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 212–223, Jan. 2017.

[23] Z.-K. Zhang *et al.*, "Dynamics of information diffusion and its applications on complex networks," *Phys. Rep.*, vol. 651, pp. 1–34, Sep. 2016.

[24] S. B. Seidman, "Network structure and minimum degree," *Soc. Netw.*, vol. 5, no. 3, pp. 269–287, 1983.

[25] M. Kitsak *et al.*, "Identification of influential spreaders in complex networks," *Nat. Phys.*, vol. 6, pp. 888–893, Aug. 2010.

[26] L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, "The H-index of a network node and its relation to degree and coreness," *Nat. Commun.*, vol. 7, Jan. 2016, Art. no. 10168.

[27] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw.*, vol. 56, no. 18, pp. 3825–3833, 2012.

[28] L. Lü, Y. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS ONE*, vol. 6, Jun. 2011, Art. no. e21202.

[29] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted LeaderRank," *Physica A Stat. Mech. Appl.*, vol. 404, pp. 47–55, Jun. 2014.

[30] S. Xu and P. Wang, "Identifying important nodes by adaptive LeaderRank," *Physica A Stat. Mech. Appl.*, vol. 469, pp. 654–664, Mar. 2017.

[31] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, "Approximating PageRank from in-degree," in *Proc. Int. Workshop Algorithms Models Web Graph*, Banff, AB, Canada, 2006, pp. 59–71.

[32] D. Chen, H. Gao, L. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: The role of clustering," *PLoS ONE*, vol. 8, no. 10, 2013, Art. no. e77455.

[33] A. Zeng and C.-J. Zhang, "Ranking spreaders by decomposing complex networks," *Phys. Lett. A*, vol. 377, no. 14, pp. 1031–1035, 2013.

[34] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *J. Math. Soc.*, vol. 2, no. 1, pp. 113–120, 1972.

[35] R. Poulin, M.-C. Boily, and B. R. Mâsse, "Dynamical systems to define centrality in social networks," *Soc. Netw.*, vol. 22, no. 3, pp. 187–220, 2000.

[36] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[37] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, pp. 81–93, 1938.

[38] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Rev. Mod. Phys.*, vol. 87, no. 3, pp. 925–979, 2015.

[39] D. F. Gleich, "PageRank beyond the Web," *SIAM Rev.*, vol. 57, no. 3, pp. 321–363, 2015.

[40] B. L. Chen, D. H. Hall, and D. B. Chklovskii, "Wiring optimization can relate neuronal structure and function," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 12, pp. 4723–4728, 2006.

[41] A. M. Huerta, H. Salgado, D. Thieffry, and J. Collado-Vides, "RegulonDB: A database on transcriptional regulation in Escherichia coli," *Nucleic Acids. Res.*, vol. 26, no. 1, pp. 55–59, 1998.

[42] S. Xu and P. Wang, "Coarse graining of complex networks: A k-means clustering approach," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Yinchuan, China, 2016, pp. 4113–4118.

[43] L. Lü *et al.*, "Toward link predictability of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 8, pp. 2325–2330, 2015.

[44] T. Zhou *et al.*, "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 10, pp. 4511–4515, 2010.

[45] L. Yang, X. Cao, D. Jin, X. Wang, and D. Meng, "A unified semi-supervised community detection framework using latent space graph regularization," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2585–2598, Nov. 2015.

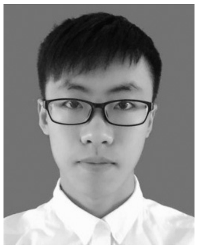
[46] T. He and K. C. C. Chan, "MISAGA: An algorithm for mining interesting subgraphs in attributed graphs," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1369–1382, May 2018.

[47] S. Xu, P. Wang, and J. Lü, "Iterative neighbour-information gathering for ranking nodes in complex networks," *Sci. Rep.*, vol. 7, Jan. 2017, Art. no. 41321.

[48] D. Garlaschelli and M. I. Loffredo, "Fitness-dependent topological properties of the world trade Web," *Phys. Rev. Lett.*, vol. 93, no. 18, 2004, Art. no. 188701.

[49] R. Metzner, "Fundamental of statistical and thermal physics," *Phys. Today*, vol. 20, no. 12, pp. 85–87, 1967.

[50] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.



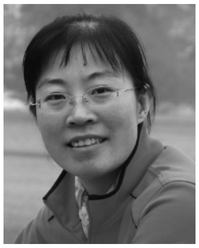
Shuang Xu is currently pursing the Ph.D. degree in statistics with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

His current research interests include statistics, data mining, and complex network and system.

Pei Wang (M'18) received the M.Sc. and Ph.D. degrees in computational mathematics from the School of Mathematics and Statistics, Wuhan University, Wuhan, China, in 2009 and 2012, respectively.

He is currently an Associate Professor with the School of Mathematics and Statistics, Henan University, Kaifeng, China. He was a Visiting Research Fellow with the School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Melbourne, VIC, Australia. He has authored and co-authored over 30 journal papers. His current research interests include biostatistics, systems biology, and complex systems and networks.

Dr. Wang serves as a Reviewer of *American Mathematical Reviews*. He is currently serving as a Technique Committee Member of the Complex Systems and Complex Networks Society and the Chinese Society for Industrial and Applied Mathematics.



Chun-Xia Zhang received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2010.

She is currently an Associate Professor with the School of Mathematics and Statistics, Xi'an Jiaotong University. She has authored and co-authored over 30 journal papers on ensemble learning techniques and nonparametric regression. Her current research interests include area of ensemble learning, variable selection, and deep learning.

Jinhu Lü (M'03–SM'06–F'13) received the Ph.D. degree in applied mathematics from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, in 2002.

He was a Professor with RMIT University, Melbourne, VIC, Australia, and a Visiting Fellow with Princeton University, Princeton, NJ, USA. He is currently the Dean of the School of Automation Science and Electrical Engineering, Beihang University, Beijing. He is also a Professor with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He is the Chief Scientist of the National Key Research and Development Program of China and a Leading Scientist of Innovative Research Groups of the National Natural Science Foundation of China. His current research interests include nonlinear circuits and systems, complex networks, multiagent systems, and big data.

Dr. Lü was a recipient of the Prestigious Ho Leung Ho Lee Foundation Award in 2015; the State Natural Science Award three times from the Chinese Government in 2008, 2012, and 2016, respectively; the Australian Research Council Future Fellowships Award in 2009; the National Natural Science Fund for Distinguished Young Scholars, and a Leading Scientist of the Ten Thousand Talents Program of China; and the Highly Cited Researcher Award in engineering in 2014–2017. He was a member of the Evaluating Committees of the IEEE Circuits and Systems Society, the IEEE Industrial Electronics Society, and the IEEE Computational Intelligence Society. He was the General Co-Chair of the 43rd Annual Conference of the IEEE Industrial Electronics Society in 2017. He was an Editor in various ranks for 15 SCI journals, including seven IEEE TRANSACTIONS journals. He is also the Co-Editor-in-Chief of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS.