Information Sciences 454–455 (2018) 229–241

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Identifying influential genes in protein–protein interaction networks

Peng Gang Sun[a,b,*], Yi Ning Quan[a,*], Qi Guang Miao[a], Juan Chi[c]

[a] School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[b] Center for Complex Data and Network Science, Xidian University, Xi'an 710071, China
[c] The 61st Research Institute of PLA, Beijing 100039, China

ARTICLE INFO

Article history:
Received 7 December 2017
Revised 23 April 2018
Accepted 29 April 2018
Available online 30 April 2018

Keywords:
Influential gene
Protein–protein interaction network

ABSTRACT

Influential nodes in influence maximization problems are of great importance for the spread of information in complex networks. In this study, we identify influential nodes, called influential genes, in protein–protein interaction (PPI) networks. In theory, information can percolate through an entire network when influential genes are activated. We propose a new framework by taking the asymmetry of influence into account to identify genes that are more influential in PPI networks. In the framework, we identify influential genes by considering the heterogeneity of influence. As such, the minimal set of influential genes in the influence maximization problem can be mapped onto the optimal set of genes in the optimal percolation problem. We identify the influential genes in the PPI networks of five species, and the results show that the genes identified by our method are more influential and tend to be located in the core of a PPI network. In addition, we find that influential genes tend to be more significantly enriched in essential yeast genes, tumor suppressor genes, and drug target genes.

1. Introduction

Network science has grown rapidly [1,39], and has attracted more attention in recent years. Complex networks can be used to model a wide range of real systems, including communications, marketing, and transportation systems [11,12], and they help us to understand the mechanisms underlying complex phenomena [1,39]. Subtopics, including community detection [17,22], network controllability [16,23,24,28–30,40,43,45], and the information spreading problem [8,21,25,26,31–37,42], provide us with new insight into complex systems from multiple perspectives.

Network science considerably facilitates the development of bioinformatics [4,18,41]. Theories of complex networks provide effective tools to analyze biologically networked systems, such as identifying functional units/modules or protein complexes in protein–protein interaction (PPI) networks [5,6,41] and predicting disease-causing genes [18,19,27,46] and drug target genes [46,49], which are of great importance for revealing the mechanisms of biochemical processes as well as those of diseases at a biomolecular level.

Recently, the influence maximization problem in complex networks has attracted more attention [21,25,26,31–35,42], with extensive application prospects, including product promotion [38] and behavior adoption [8]. This problem focuses

[*] Corresponding authors at: School of Computer Science and Technology, Xidian University, Xi'an 710071, China.
E-mail addresses: psun@mail.xidian.edu.cn (P.G. Sun), ynquan@mail.xidian.edu.cn (Y.N. Quan).

https://doi.org/10.1016/j.ins.2018.04.078
0020-0255/© 2018 Elsevier Inc. All rights reserved.

on the search for the minimal set of influential nodes (or influencers), whose activation can cause information to permeate through the whole network, or whose removal can break down the whole network into many disconnected pieces [21,25,26,31–35,42]. Influential nodes can be attributed to the heterogeneity of complex networks—that is, that some nodes located at more important positions are more influential than others, and more likely to trigger information diffusion on a large scale [21,25,26,31–35,42].

Currently, identifying influential nodes can be classified into two subtopics [32,33]: (1) identifying a single influential node as the spreading origin of, say, an epidemic or a rumor, and (2) identifying the minimal set of influential nodes as the spreading origin (e.g., of an advertisement) that maximizes the scale of product promotion. Pei and Makse [32] and Pei et al. [33] summarized related methods for identifying influential nodes in a review. For a single influential node, nodes with a high-degree [3,10] or high-betweenness centrality [14,15] intuitively tend to be located in very important positions, and are able to provoke epidemic diffusion on a large scale. However, Kitsak et al. [21] demonstrated that the most efficient influential nodes tend to be located within the core of a network, and can be identified by k-shell decomposition [9]. In spreads that begin with multiple influential nodes, Morone et al. [25] and Morone and Makse [26] identified the minimal set of influential nodes for maximizing influence by mapping the problem onto the optimal percolation problem. Their results showed that certain low-degree nodes in specific positions are in fact super-influential nodes.

In this study, we identify influential nodes, called influential genes, in PPI networks. In theory, information can percolate through an entire network when influential genes are activated. We propose a new framework by taking the asymmetry of influence into account to identify those genes that are more influential in PPI networks. In the framework, we identify influential genes by considering the heterogeneity of influence. As such, the minimal set of influential genes in the influence maximization problem can be mapped onto the optimal set of genes in the optimal percolation problem, which has been fully discussed by Morone et al. [25] and Morone and Makse [26] in complex networks. We used our method to study genes that are influential in the PPI networks of five species.

The rest of the paper is organized as follows. In Section 2, we present definitions. In Section 3, we introduce our framework for identifying influential genes. In Section 4, we use our method to identify the influential genes in the PPI networks of five species. The paper is concluded in Section 5.

2. Preliminaries

Consider a PPI network that contains n nodes and m links denoted by its adjacency matrix, $\mathbf{A}$.

Definition 1. (Adjacency matrix)

$$\mathbf{A} = (a_{ij})_{n\times n} \tag{1}$$

where $a_{ij} = 1$ if gene i and gene j are connected by a link, and $a_{ij} = 0$ otherwise, and where $i, j = 1, 2,\ldots, n$.

Definition 2. (Degree)

The degree of gene i, $k(i)$, is defined as the number of other genes that are adjacent to it, and the average degree of a PPI network, $\mathbf{A}$, is denoted by $\langle k\rangle$.

$$k(i) = \sum_{j=1}^{n} a_{ij} \tag{2}$$

Definition 3. (Coreness) Coreness is a metric that can be used to measure the locations of genes in a PPI network. By k-shell decomposition analysis [9,21], each gene is assigned an integer value, ks, according to successive layers of the PPI network to indicate its location. In Fig. 1, we illustrate the locations of genes in a toy PPI network, and we can see that gene i ($ks = 1$) is located in the periphery of the network, and the genes with large ks values tend to be located in the core of the network. The average coreness of a PPI network, $\mathbf{A}$, is denoted by $\langle ks\rangle$.

3. Influence problem in PPI networks

In this section, we first describe the influence maximization problem in complex networks, and then introduce our framework, which extends the optimal percolation theory introduced by Morone et al. [25] and Morone and Makse [26] to identify influential genes in PPI networks.

3.1. Influence maximization

The influence maximization problem aims to find the minimal set of influential nodes [25,26]. If these nodes are activated, information can percolate through the whole network. Conversely, if these nodes are removed, the whole network can be broken down into several disconnected pieces. However, the heterogeneity of PPI networks determines the asymmetry of influence, i.e., the mutual influence of two genes is asymmetrical [21,25,26,32,33,40]. For example, the ability of gene i to influence gene j is often different from the ability of gene j to influence gene i. Moreover, a gene in a more important position tends to have greater influence on its closely associated neighbors. Therefore, for the spread of information between

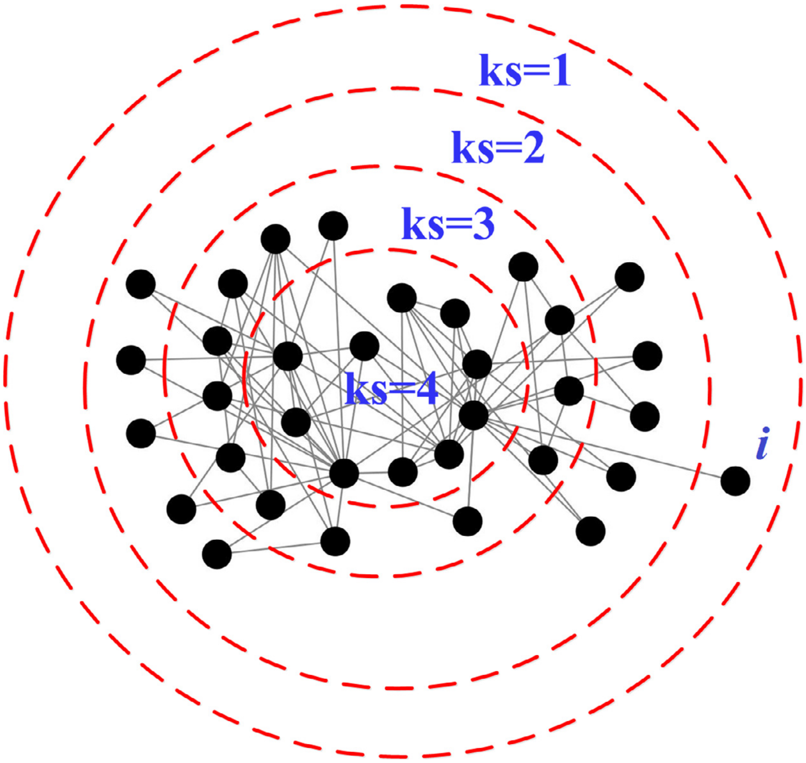

Fig. 1. Illustration of coreness by a toy PPI network.

genes, a gene's ability to influence should be considered in the influence maximization problem in order to identify genes that are more influential in PPI networks.

3.2. Framework for identifying influential genes

We propose a new framework to identify genes that are more influential by taking the asymmetry of influence into account. In the framework, we first construct a network transformed from PPI networks by considering the heterogeneity of influence to describe the spreading pathways. Here, we consider a PPI network that contains n nodes and m links denoted by its adjacency matrix, $\mathbf{A}$. We define the ability of gene i to influence gene j as $p_{i\mapsto j}$,

$$p_{i\mapsto j} = \left[\frac{a_{ij} + |\theta(i,j)|}{k_j}\right] \times a_{ij} \tag{3}$$

where $|\theta(i, j)|$ corresponds to the number of shared neighbors between gene i and gene j, and the greater the value of $|\theta(i, j)|$, the stronger the relevance.

By using $p_{i\mapsto j}$, each gene in a PPI network is activated as a spreading origin, and the weight of a directed link indicates the influence of the activated gene on its neighbor [40]. We construct a spreading network using a threshold p. We illustrate the asymmetry of influence [40] in a toy network, to show that information can percolate from gene 2 to gene 1 at $p \geq 0.3$, but not from gene 1 to gene 2 (see the top of Fig. 2).

For the influence maximization problem, Morone et al. [25] and Morone and Makse [26] proposed a framework to identify the minimal set of influential nodes that maps the problem onto the optimal percolation problem by minimizing the energy of a many-body system. Here, we extend the framework to identify the minimal set of influential genes in PPI networks by $\mathbf{A} \to \mathbf{A}_t$ (see Fig. 2).

The problem of optimal percolation is to find the minimal set of nodes required to maintain the global connectivity of a network. Without at least these nodes, the network will dismantle, i.e., removing these nodes collapses the giant component in the network [25,26,32–35]. Therefore, the total fraction of removed nodes in a network, $\mathbf{A}$, is denoted by $q_{\mathbf{A}}$,

$$q_{\mathbf{A}} = 1 - \frac{1}{n}\sum_{i=1}^{n} x_i^{\mathbf{A}} \tag{4}$$

where vector $\mathbf{x}^{\mathbf{A}} = (x_1^{\mathbf{A}}, x_2^{\mathbf{A}}, \ldots, x_n^{\mathbf{A}})$ indicates that $x_i^{\mathbf{A}} = 0$ if node i is removed, and $x_i^{\mathbf{A}} = 1$, otherwise.

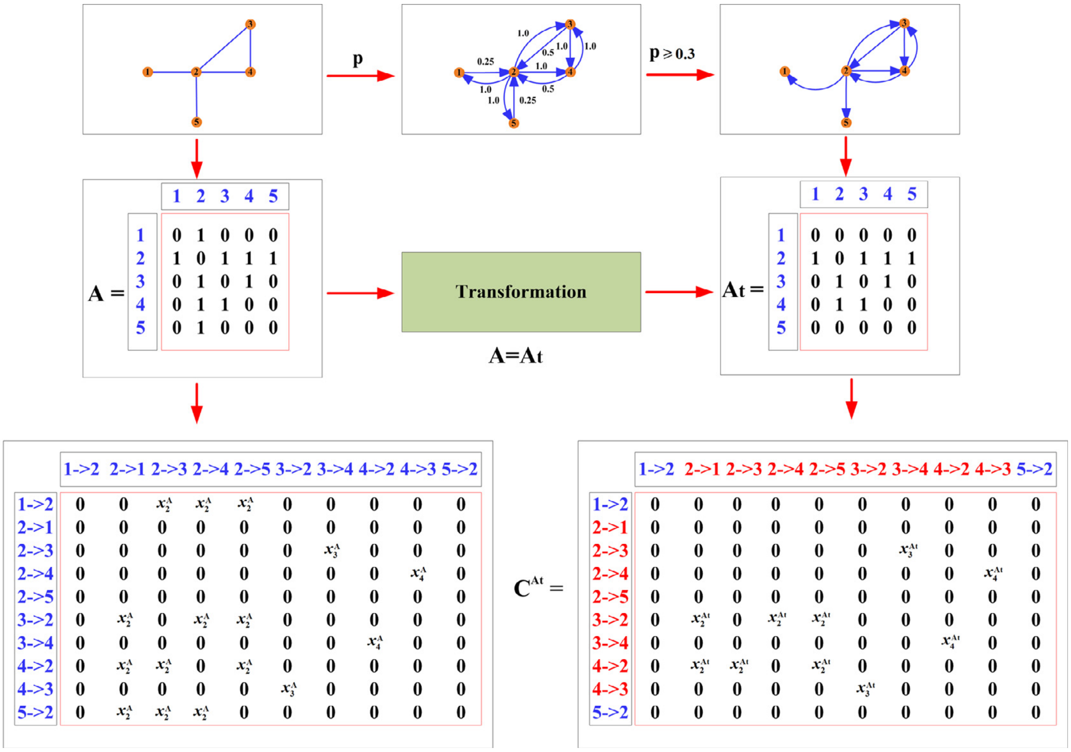

Fig. 2. Illustration of the proposed framework. Spreading networks are shown at the top. At the bottom, the influential genes in PPI networks are determined by optimal percolation theory based on $\mathbf{A} \to \mathbf{A}_t$.

The fraction of nodes belonging to the giant component after removing $q_{\mathbf{A}}$ of them together from the network, $\mathbf{A}$, is denoted by $G_{\mathbf{A}}(q_{\mathbf{A}})$:

$$G_{\mathbf{A}}(q_{\mathbf{A}}) = \frac{1}{n}\sum_{i=1}^{n} y_i^{\mathbf{A}} \tag{5}$$

where vector $\mathbf{y}^{\mathbf{A}} = (y_1^{\mathbf{A}}, y_2^{\mathbf{A}}, \ldots, y_n^{\mathbf{A}})$ indicates that $y_i^{\mathbf{A}} = 1$ if $i \in G_{\mathbf{A}}$, and $y_i^{\mathbf{A}} = 0$, otherwise.

The optimal influence problem corresponds to finding the minimum fraction $q_{\mathbf{A}}^*$ of influential nodes to fragment the network, $q_{\mathbf{A}}^* = min\{q_{\mathbf{A}} \in [0,1] | G_{\mathbf{A}}(q_{\mathbf{A}}) = 0\}$, which can be determined by the largest eigenvalue [25,26], $\lambda(\mathbf{x}^{\mathbf{A}}; q_{\mathbf{A}})$, of the coupling matrix, $C^{\mathbf{A}} = (c^{\mathbf{A}}_{u\to v, i\to j})_{2m\times 2m}$, defined on $2m \times 2m$ directed links, $c^{\mathbf{A}}_{u\to v,i\to j} \equiv \frac{\partial y^{\mathbf{A}}_{i\to j}}{\partial y^{\mathbf{A}}_{u\to v}}|_{\{y^{\mathbf{A}}_{i\to j}=0\}}$, where $y^{\mathbf{A}}_{i\to j} = 0$ if $i \notin G_{\mathbf{A}}$ in the case of the removal of j. Here, $c^{\mathbf{A}}_{u\to v,i\to j} = x_i^{\mathbf{A}} b^{\mathbf{A}}_{u\to v,i\to j}$, where $b^{\mathbf{A}}_{u\to v,i\to j} = 1$, if $v = i$ and $u \neq j \neq i$, and $b^{\mathbf{A}}_{u\to v,i\to j} = 0$, otherwise (see the bottom of Fig. 2).

Morone et al. [25] and Morone and Makse [26] demonstrated that the optimal configuration of $nq_{\mathbf{A}}^*$ influential nodes, $\mathbf{x}^{\mathbf{A}*}$, is therefore obtained when the minimum of the largest eigenvalue satisfies $\lambda(\mathbf{x}^{\mathbf{A}*}; q_{\mathbf{A}}^*) = 1$, and the largest eigenvalue is then calculated with the power method [7]:

$$\lambda(\mathbf{x}^{\mathbf{A}}; q_{\mathbf{A}}) = \lim_{L\to\infty}\left[\frac{|\mathbf{w}_L(\mathbf{x}^{\mathbf{A}})|}{|\mathbf{w}_0|}\right]^{\frac{1}{L}} \tag{6}$$

where $|\mathbf{w}_L(\mathbf{x}^{\mathbf{A}})|$ is the Lth iteration of $C^{\mathbf{A}}$ on $\mathbf{w}_0$, and $|\mathbf{w}_L(\mathbf{x}^{\mathbf{A}})| = |C_L^{\mathbf{A}}\mathbf{w}_0|$. The best configuration of the vector $\mathbf{x}^{\mathbf{A}}$ corresponds to the minimized cost function, $|\mathbf{w}_L(\mathbf{x}^{\mathbf{A}})|$, for minimizing the largest eigenvalue of matrix $C^{\mathbf{A}}$ to achieve influence maximization [25,26].

Morone et al. [25] and Morone and Makse [26] simplified the problem by an approximation of $|\mathbf{w}_L(\mathbf{x}^{\mathbf{A}})|^2$:

$$|\mathbf{w}_L(\mathbf{x}^{\mathbf{A}})|^2 = \sum_{i=1}^{n}(k(i)-1)\sum_{j\in\partial\mathbf{Ball}(i,2L-1)}\left(\prod_{r\in\mathbf{Path}_{2L-1}(i,j)} x_r^{\mathbf{A}}\right)(k(j)-1) \tag{7}$$

where $\partial\mathbf{Ball}(i, L)$ indicates the frontier of a ball of radius L in terms of the shortest path centered at node i, and $\mathbf{Path}_L(i, j)$ is the shortest path of length L connecting node i and node j (see Fig. 3).

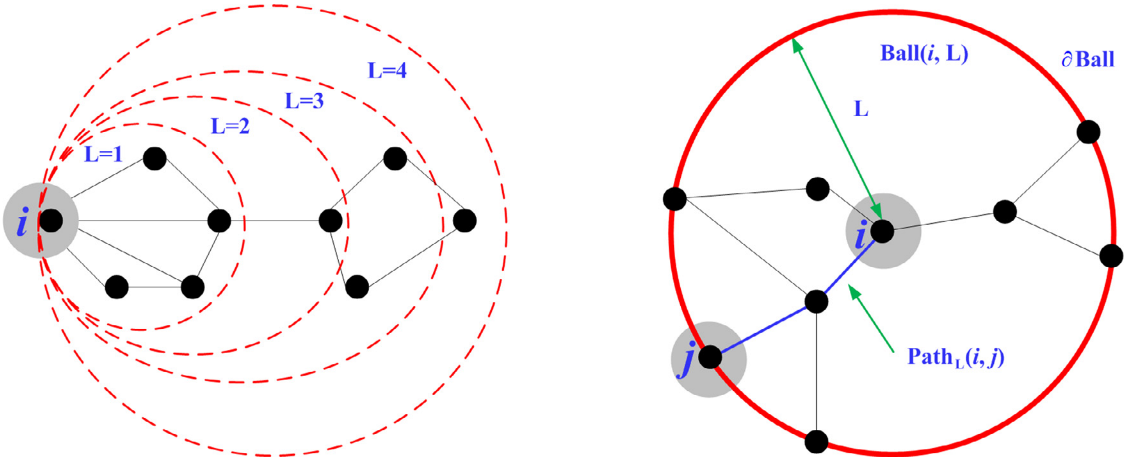

Fig. 3. Illustration of L-neighborhood of node i (left) and the ∂**Ball** (right).

Further, Morone et al. [25] and Morone and Makse [26] defined an energy function for each configuration, $\mathbf{x^A}$:

$$\mathbf{E}_L(\mathbf{x^A}) = \sum_{i=1}^{n}(k(i)-1)\sum_{j\in\partial\mathbf{Ball}(i,L)}\left(\prod_{r\in\mathbf{Path}_L(i,j)} x_r^{\mathbf{A}}\right)(k(j)-1) \tag{8}$$

$$\mathbf{E}_L(\mathbf{x^A}) = \begin{cases} |\mathbf{w}_{\frac{L+1}{2}}(\mathbf{x^A})|^2, & L \text{ is odd} \\ |\mathbf{w}_{\frac{L}{2}}(\mathbf{x^A})|^2, & L \text{ is even} \end{cases} \tag{9}$$

Moreover, they rewrote $\mathbf{E}_L(\mathbf{x^A})$ as the sum of the collective influence (**CI**) of each node:

$$\mathbf{E}_L(\mathbf{x^A}) = \sum_{i=1}^{n}(\mathbf{CI}(i)-1) \tag{10}$$

$$\mathbf{CI}_L^{\mathbf{A}}(i) = (k(i)-1)\sum_{j\in\partial\mathbf{Ball}(i,L)}(k(j)-1) \tag{11}$$

where L denotes the neighborhood of a node based on the shortest paths (see Fig. 3), and $\mathbf{CI}_L^{\mathbf{A}}(i)$ corresponds to the collective influence of node i at length L.

The **CI** algorithm removes those nodes that can cause the largest decrease in the energy function $\mathbf{E}_L(\mathbf{x^A})$ [25,26,31–35]. In the **CI** algorithm, the node with the largest CI value is first removed at each iteration, and then the CI values of the remaining nodes are recalculated [25,26,31–35]. This procedure is repeated until the giant component is fragmented [25,26,31–35]. The pseudocodes for the **CI** algorithm and the extended **CI** algorithm based on our framework are given in Algorithms 1 and 2, respectively.

Algorithm 1 Pseudocode for the **CI** algorithm to identify influential nodes.

```
Input: The network, $\mathbf{A}$;
Output: The vector, $\mathbf{x^A} = (x_1^{\mathbf{A}}, x_2^{\mathbf{A}}, \ldots, x_n^{\mathbf{A}})$;
 1: Initialization:
 2: For all $i \in \{1, 2, \ldots, n\}$;
 3:    $x_i^{\mathbf{A}} = 1$;
 4: Repeat:
 5: For all $i \in \{1, 2, \ldots, n\}$;
 6:    If $x_i^{\mathbf{A}} = 1$;
 7:       calculate $\mathbf{CI}_L^{\mathbf{A}}(i)$;
 8: Remove node $i$ with the maximum $\mathbf{CI}_L^{\mathbf{A}}(i)$;
 9: Set $x_i^{\mathbf{A}} = 0$;
10: Stop: If $G_{\mathbf{A}}(q_{\mathbf{A}}) = 0$,
11:          Else Go to step 4;
```

The analysis above shows the optimal percolation theory for identifying the minimal set of influential nodes [25,26], which can be easily extended by transformation, $\mathbf{A} \to \mathbf{A}_t$, in order to identify influential genes in PPI networks based on our framework (see Fig. 2).

Algorithm 2 Pseudocode for the extended **CI** algorithm based on our framework for identifying influential genes ($\mathbf{A} \rightarrow \mathbf{A}_t$).

```
Input: The network, $\mathbf{A}_t$;
Output: The vector, $\mathbf{x}^{\mathbf{A}_t} = (x_1^{\mathbf{A}_t}, x_2^{\mathbf{A}_t}, \ldots, x_n^{\mathbf{A}_t})$;
 1: Initialization:
 2: For all $i \in \{1, 2, \ldots, n\}$;
 3:    $x_i^{\mathbf{A}_t} = 1$;
 4: Repeat:
 5: For all $i \in \{1, 2, \ldots, n\}$;
 6:    If $x_i^{\mathbf{A}_t} = 1$;
 7:       calculate $\mathbf{CI}_L^{\mathbf{A}_t}(i)$;
 8: Remove node $i$ with the maximum $\mathbf{CI}_L^{\mathbf{A}_t}(i)$;
 9: Set $x_i^{\mathbf{A}_t} = 0$;
10: Stop: If $G_{\mathbf{A}_t}(q_{\mathbf{A}_t}) = 0$,
11:          Else Go to step 4;
```

Table 1
Properties of PPI networks for the five species.

PPI networks	n	m	$\langle k \rangle$	$\langle ks \rangle$
Human	11,150	46,963	8.42	4.34
Worm	8435	14,955	3.55	1.85
Fly	8566	26,965	6.30	3.32
Yeast	5490	54,161	19.73	10.21
E. coli	2927	14,428	9.86	5.19

4. Results and discussions

In this section, we first describe the PPI data of five species used in this study, and then identify influential genes in these PPI networks. Finally, we analyze the biological significance of the identified influential genes.

4.1. Experimental data

The experimentally determined data of protein–protein interactions—viz., Homo sapiens (human), Caenorhabditis elegans (worm), Drosophila melanogaster (fly), Saccharomyces cerevisiae (yeast), and Escherichia coli (E. coli)—were collected from the IntAct database [20]. Table 1 shows the properties of the PPI networks of the five species.

4.2. Identifying influential genes in PPI networks

To identify the influential genes in these species, we first used the original framework introduced by Morone et al. [25] and Morone and Makse [26]. The results are shown in Fig. 4, where Fig. 4(a) and Fig. 4(b) correspond to $L = 1$ and $L = 2$, respectively. From the figure, we can see that for the PPI network of the worm, $G_{\mathbf{A}}(q_{\mathbf{A}})$, decreases quickly as $q_{\mathbf{A}}$ increases, compared with the PPI networks of other species. That is, the giant component in the PPI network of the worm becomes highly fragmented as $q_{\mathbf{A}}$ increases, such that removing the fraction of influential genes collapses the connectivity of the PPI network of the worm. For the influence maximization of PPI networks, the natural measure of influence is the size of the giant connected components in PPI networks as the influential genes are removed [25,26]. The most influential genes are the ones forming the minimal set that maintains a global connection of PPI networks. The genes in this set are the optimal influential genes in the PPI network. We aim to find the minimal set of influential genes, such that when these genes are activated, information percolates through the whole PPI network. By contrast, if they are immunized, information diffusion will be prevented (i.e., if these genes are removed, the giant connected component of the PPI network will be broken down). In Fig. 4(c) and (d), we compare the results for variable L, showing that it becomes easier to dismantle the networks as L increases. As mentioned by Morone et al. [25] and Morone and Makse [26], richer topological information in the networks tends to be considered as L increases, and optimal influential genes are more easily approximated. Fig. 5 shows the influential genes identified by our framework in the PPI networks of the five species, with similarities to the results in Fig. 4 (i.e., for the PPI network of the worm, $G_{\mathbf{A}_t}(q_{\mathbf{A}_t})$ decreases quickly as $q_{\mathbf{A}_t}$ increases, compared with the PPI networks of other species).

In Fig. 6, we compare the influential genes identified by the proposed framework with those identified by the original framework. From the figure, we can see that the influential genes identified by our framework are more influential, and their removal will cause the networks to collapse. That is, our framework extracts the optimal influential genes in PPI networks for influence maximization. In our framework, we take the asymmetry of mutual influence into account, which can be observed in the weighted version of the PPI networks based on Eq. (3). By choosing p, information percolates from gene i to

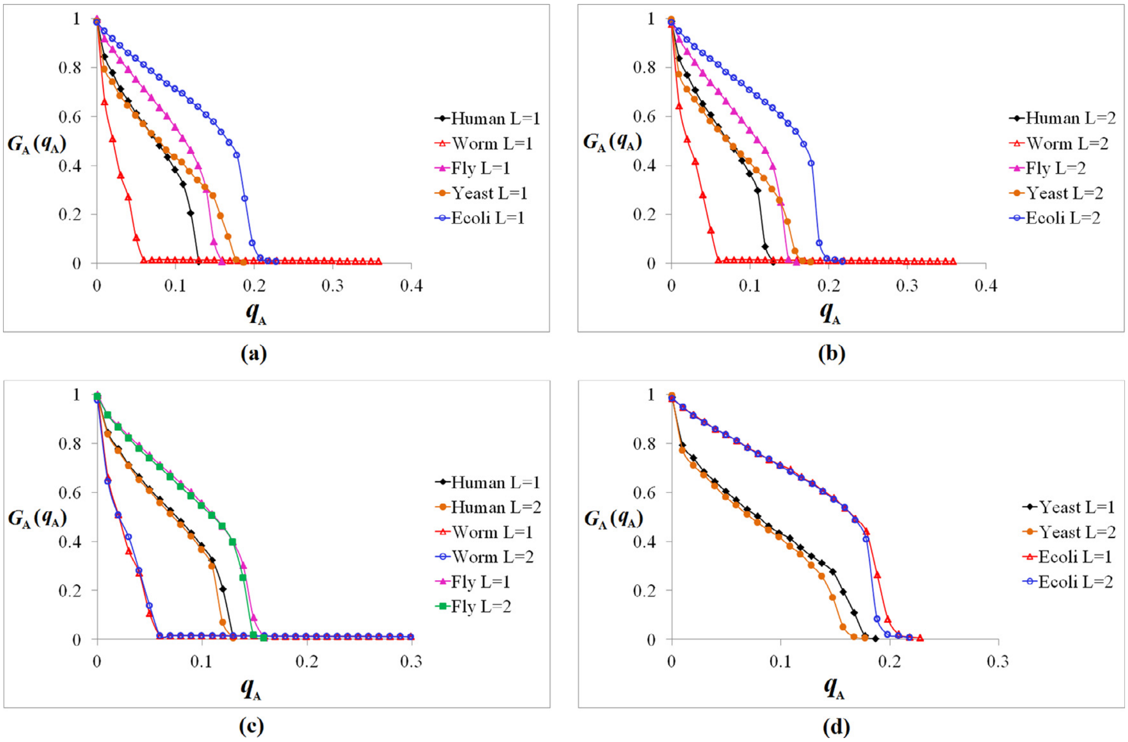

Fig. 4. Influential genes in the PPI networks of five species based on the original framework. Here, (a) and (b) correspond to L =1 and L = 2, respectively. In (c) and (d), we compare the results for variable L.

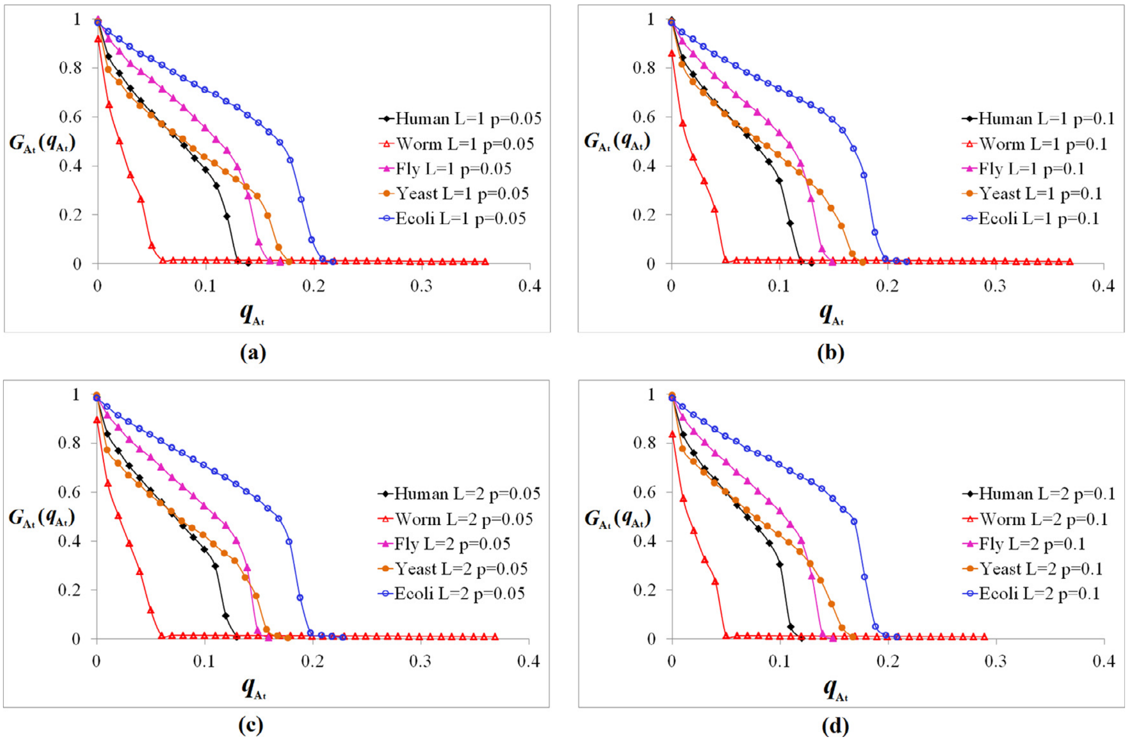

Fig. 5. Influential genes in the PPI networks of five species based on our framework. Here, (a) and (b) correspond to L = 1 and L = 2, respectively. In (c) and (d), we compare the results for variable L.

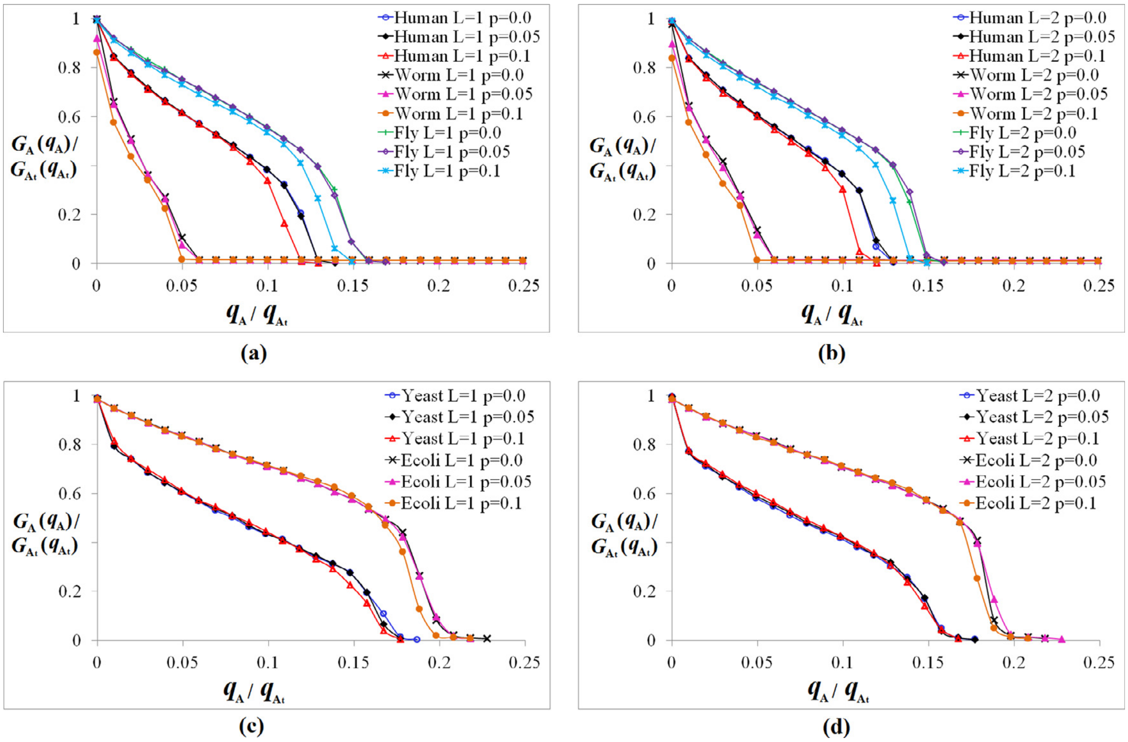

Fig. 6. Comparison of influential genes in our framework and the original framework. When $p = 0.0$, the results are the same ($\mathbf{A} = \mathbf{A}_t$).

Table 2
Illustration of Fisher's exact test for influential genes and essential genes.

	Influential genes	Non-influential genes	Row total
Essential genes	*a*	*b*	*a*+*b*
Non-essential genes	*c*	*d*	*c*+*d*
Column total	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d* (=*n*)

Table 3
Enrichment of influential genes in essential yeast genes.

Methods	*L*	*p*	Value of Fisher's exact test
Original framework	1	0.00	1.96×10^{-93}
	2	0.00	1.69×10^{-63}
Proposed framework	1	0.05	1.58×10^{-89}
	1	0.10	4.08×10^{-90}
	2	0.05	2.13×10^{-62}
	2	0.10	6.39×10^{-64}

gene *j*, but not from gene *j* to gene *i*, insofar as gene *i* tends to have a more important position and holds greater influence on gene *j*. When $p = 0.0$, the results from our framework are equivalent to those of the original framework, because $\mathbf{A} = \mathbf{A}_t$. Here, we selected $p = 0.05$ and $p = 0.1$, because as *p* increases, the weighted version of the PPI networks tends to be sparse. That is, some nodes tend to disjoin from the weighted version of the PPI networks. We aim to find the minimal set of nodes needed to maintain the global connectivity of the PPI network. This is equivalent to determining which nodes will cause the network to dismantle when they are removed (i.e., the nodes whose removal will collapse the giant component in the network). By assigning *p* a lower value, we can keep the global connectivity of the weighted version the same as the original networks, i.e., $G_{\mathbf{A}_t}(q_{\mathbf{A}_t} = 0) = 1$, which equitably verifies the effectiveness of removing influential nodes to dismantle the network.

Table 4
Enrichment of influential genes in human tumor suppressor genes.

Methods	L	p	Value of Fisher's exact test
Original framework	1	0.00	3.01×10^{-28}
	2	0.00	2.03×10^{-23}
Proposed framework	1	0.05	4.44×10^{-30}
	1	0.10	1.27×10^{-28}
	2	0.05	6.30×10^{-24}
	2	0.10	8.95×10^{-24}

Table 5
Enrichment of influential genes in human drug target genes.

Methods	L	p	Value of Fisher's exact test
Original framework	1	0.00	2.63×10^{-11}
	2	0.00	1.53×10^{-8}
Proposed framework	1	0.05	2.77×10^{-11}
	1	0.10	4.66×10^{-12}
	2	0.05	5.51×10^{-10}
	2	0.10	2.40×10^{-10}

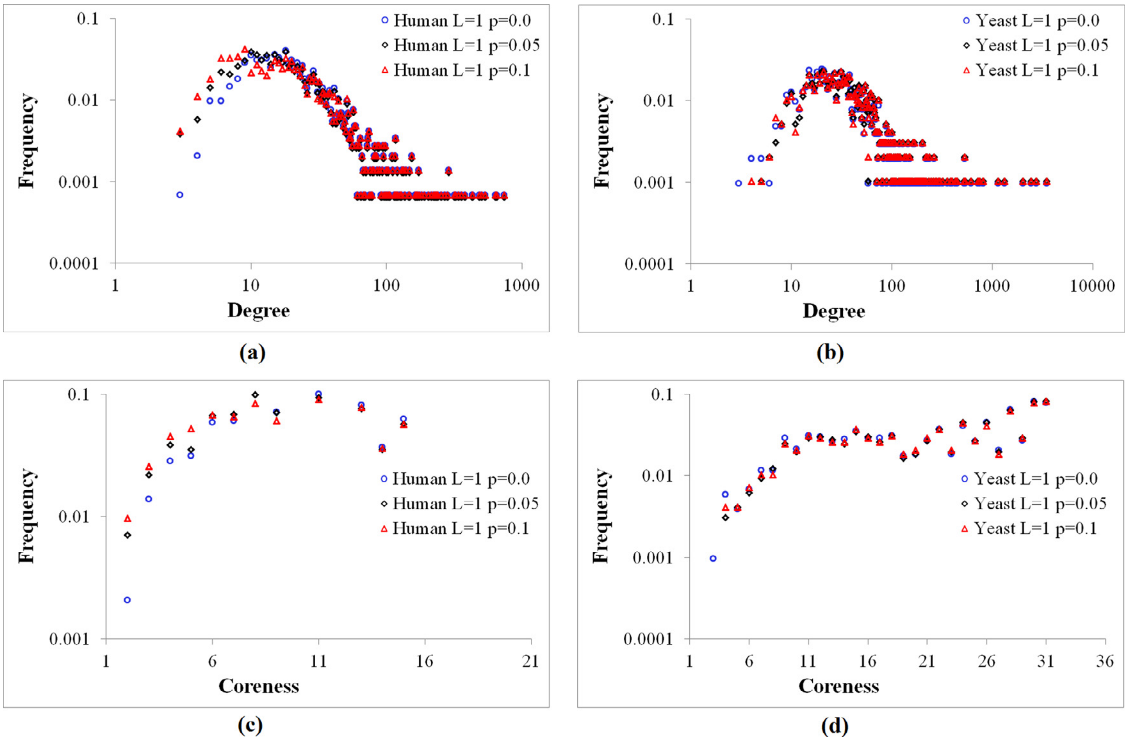

Fig. 7. Degree and location distributions of influential genes in human and yeast PPI networks. Here, (a) and (b) correspond to the degree distribution of influential genes, and (c) and (d) correspond to the location distribution of influential genes. When $p = 0.0$, the results of our framework are equivalent to those of the original framework ($\mathbf{A} = \mathbf{A}_t$).

4.3. *Biological significance of influential genes*

In order to indicate the biological significance of the identified influential genes in PPI networks, we collected 1129 essential yeast genes from [47], 1217 human tumor suppressor genes from [48], and 2456 drug targets from [44]. Then, we applied Fisher's exact test [13].

Fisher's exact test [13] is a statistical significance test that is used in the analysis of contingency tables. The test is useful for categorical data that result from classifying objects in two different ways. It is used to examine the significance of the association (contingency) between the two types of classification. The significance of the deviation from a null hypothesis (e.g., p-value) can be calculated exactly. Here, we take the following table, a 2×2 contingency table, as an example to

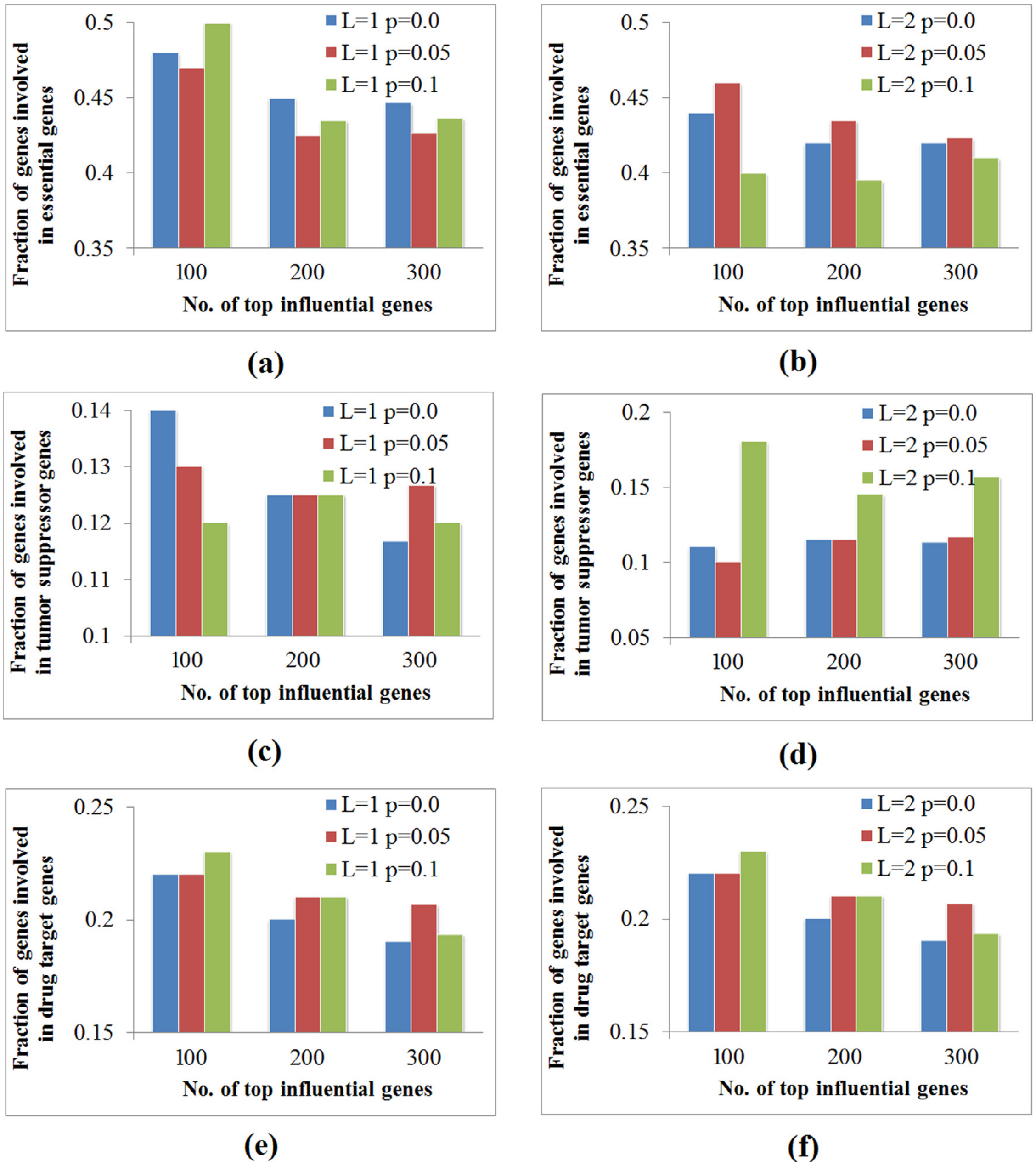

Fig. 8. Top influential genes involved in essential yeast genes, human tumor suppressor genes, and drug target genes. Here, (a) and (b) correspond to the fraction of influential genes involved in essential yeast genes for a fixed L; (c) and (d) correspond to the fraction of influential genes involved in human tumor suppressor genes for a fixed L; and (e) and (f) correspond to the fraction of influential genes involved in human drug target genes for a fixed L. When $p = 0.0$, the results of our framework are equivalent to those of the original framework ($\mathbf{A} = \mathbf{A}_t$).

test the statistical significance of influential genes in essential genes. In Table 2, "a+c" corresponds to the total number of influential genes, and "a" corresponds to the number of influential genes that are essential genes.

The p-value of Fisher's exact test [13] is computed under a null hypothesis of independence to a hypergeometric distribution of the numbers in the cells of the table.

$$p - value(Fisher) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{12}$$

Table 3 shows the enrichment of influential genes in essential yeast genes. We can see that the influential genes identified by our framework and by the original framework are more significantly enriched in essential yeast genes, and that as L

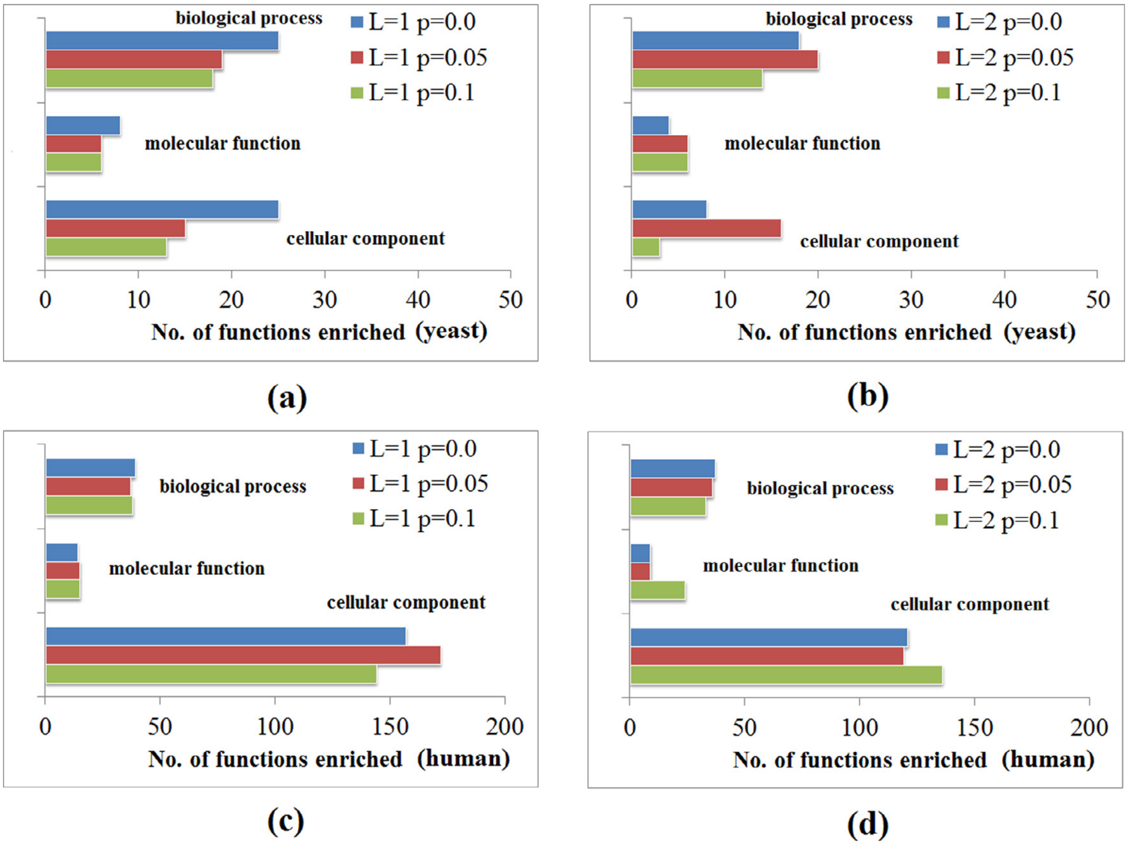

Fig. 9. Top influential genes enriched in GO annotations. Here, (a) and (b) correspond to the number of functional annotations enriched in yeast influential genes with a fixed L; (c) and (d) correspond to the number of functional annotations enriched in human influential genes with a fixed L. When $p = 0.0$, the results of our framework are equivalent to those of the original framework ($\mathbf{A} = \mathbf{A}_t$).

and p increase, our framework exhibits better performance. Tables 4 and 5 correspond to the enrichment of the influential genes in human tumor suppressor genes and drug target genes, respectively, and we can observe that the influential genes are also more significantly enriched in human tumor suppressor genes and drug target genes.

Fig. 7 shows the degree and location distributions of the influential genes. Fig. 7(a) and (b) correspond to the degree distribution of the influential genes in human and yeast PPI networks, respectively, and we can see that the influential genes tend to avoid low-degree nodes. Fig. 7(c) and (d) correspond to the location distribution of the influential genes in human and yeast PPI networks, respectively, and we can see that the influential genes tend to be located in the core of the PPI network.

Fig. 8 shows the top influential genes involved in essential yeast genes, human tumor suppressor genes, and drug target genes. Fig. 8(a) and (b) correspond to the fraction of influential genes involved in essential yeast genes with a fixed L, and we can see that the identified influential genes based on our framework when $L = 2$ and $p = 0.05$ are more likely to be essential yeast genes. Similarly, we can observe that as L and p increase, the identified influential genes based on our framework tend to be human tumor suppressor genes (see Fig. 8(c) and (d)) and drug target genes (see Fig. 8(e) and (f)).

Fig. 9 shows the top influential genes enriched in gene ontology (GO) [2]. Fig. 9(a) and (b) correspond to the number of functional annotations enriched in yeast influential genes with a fixed L, and we can see that more GO functions are enriched in yeast influential genes based on our framework when $L = 2$ and $p = 0.05$. Fig. 9(c) and (d) correspond to the number of functional annotations enriched in human influential genes with a fixed L, and we observe that as L and p increase, human influential genes based on our framework tend to be enriched in more GO functions. In Fig. 10, a heatmap illustrates the GO functions enriched in the influential genes.

5. Conclusion

We proposed a new framework by taking the asymmetry of influence into account to identify genes that are relatively more influential in PPI networks. The proposed framework identifies influential genes on spreading networks transformed from PPI networks by considering the heterogeneity of mutual influence. The minimal set of influential genes in the influence maximization problem can thus be mapped onto the optimal set of genes in the optimal percolation problem. We

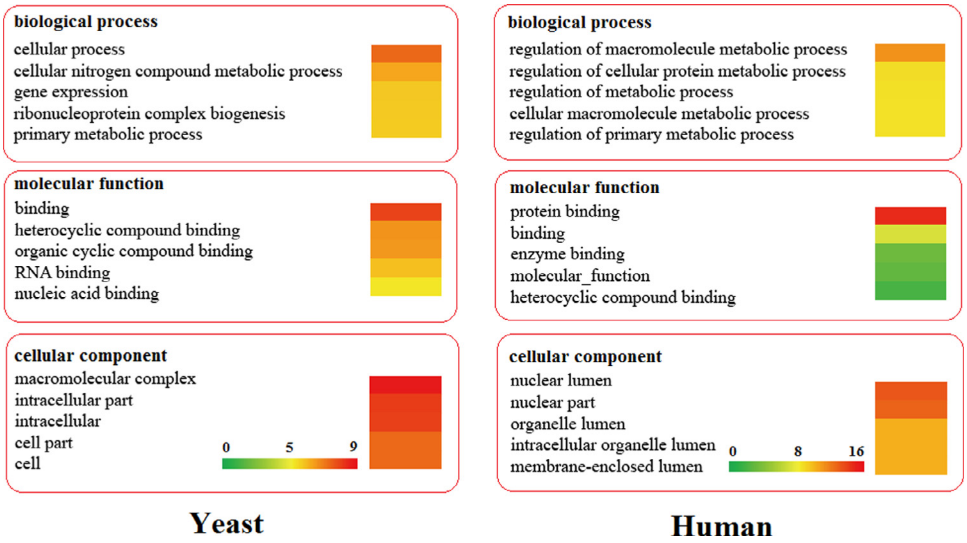

Fig. 10. Heatmap of influential genes enriched in GO annotations.

identified influential genes in the PPI networks of five species, and the results showed that the genes identified by our method were more influential and tended to be located in the core of the PPI network. In addition, we found that the influential genes tend to be more significantly enriched in essential yeast genes, human tumor suppressor genes, and drug target genes. In future research, we will apply our framework to identify influential genes in multiple biological networks.

Acknowledgments

This work was supported by the China Scholarship Council (CSC) for Studying Abroad (Grant No. 201306965011), the National Natural Science Foundation of China (Grant Nos. 61202175 and 61771369), the Fundamental Research Funds for the Central Universities (Grant Nos. BDY181417, JB180308), the Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20120203120015), the Natural Science Basic Research Plan in Shaanxi Province of China (Grant Nos. 2016JQ6044 and 2017JM6030), and the Natural Science Basic Research Plan in Ningbo City (Grant No. 2016A610034).

References

[1] R. Albert, A.L. Barabsi, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (2002) 47–97.
[2] M. Ashburner, C.A. Ball, J.A. Blake, et al., The gene ontology consortium gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[3] R. Albert, H. Jeong, A.L. Barabsi, Error and attack tolerance of complex networks, Nature 406 (2000) 378–482.
[4] A.L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet. 5 (2004) 101–113.
[5] D. Bu, Y. Zhao, L. Cai, et al., Topological structure analysis of the protein–protein interaction network in budding yeast, Nucleic Acids Res. 31 (2003) 2443–2450.
[6] G.D. Bader, C.W. Hogue, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinf. 4 (2003) 1–17.
[7] N.P. Bhatia, G.P. Szego, Stability Theory of Dynamical Systems, Springer-Verlag, Berlin, Heidelberg, 2002.
[8] D. Centola, The spread of behavior in an online social network experiment, Science 329 (5996) (2010) 1194–1197.
[9] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, A model of internet topology using k-shell decomposition, Proc. Natl. Acad. Sci. 104 (2007) 11150–11154.
[10] R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Breakdown of the internet under intentional attack, Phys. Rev. Lett. 86 (2001) 3682–3685.
[11] W.B. Du, Y. Gao, C. Liu, Z. Zheng, Z. Wang, Adequate is better: particle swarm optimization with limited-information, Appl. Math. Comput. 268 (2015) 832–838.
[12] W.B. Du, X.L. Zhou, O. Lordand, Z. Wang, C. Zhao, Y.B. Zhu, Analysis of the chinese airline network as multi-layer networks, Transp. Res. Part E Logist. Transp. Rev. 89 (2016) 108–116.
[13] R.A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, 1954.
[14] L.C. Freeman, Centrality in social networks: conceptual clarification, Soc. Netw. 1 (1979) 215–239.
[15] N.E. Friedkin, Theoretical foundations for centrality measures, Am. J. Sociol. 96 (1991) 1478–1504.
[16] J. Gao, Y.Y. Liu, R.M. D'Souza, A.L. Barabsi, Target control of complex networks, Nat. Commun. 5 (2014) 5415.
[17] K. He, Y. Li, S. Soundarajan, J.E. Hopcroft, Hidden community detection in social networks, Inf. Sci. 425 (2018) 92–106.
[18] L.H. Hartwell, J.J. Hopfield, S. Leibler, et al., From molecular to modular cell biology, Nature 402 (1999) C47–C52.
[19] M. Jalili, Graph theoretical analysis of Alzheimer's disease: discrimination of AD patients from healthy subjects, Inf. Sci. 384 (2017) 145–156.
[20] S. Kerrien, et al., The intact molecular interaction database in 2012, Nucleic Acids Res. 40 (2012) D841–D846.
[21] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (2010) 888–893.
[22] Z. Li, R.S. Wang, S. Zhang, X.S. Zhang, Quantitative function and algorithm for community detection in bipartite networks, Inf. Sci. 367 (368) (2016) 874–889.
[23] Y.Y. Liu, A.L. Barabsi, Control principles of complex networks, Rev. Mod. Phys. 88 (2016) 035006.
[24] Y.Y. Liu, J.J. Slotine, A.L. Barabsi, Controllability of complex networks, Nature 473 (2011) 167–173.

[25] F. Morone, B. Min, L. Bo, R. Mari, H.A. Makse, Collective influence algorithm to find influencers via optimal percolation in massively large social media, Sci. Rep. 6 (2016) 30062.
[26] F. Morone, H.A. Makse, Influence maximization in complex networks through optimal percolation, Nature 524 (2015) 65–68.
[27] P. Maji, E. Shah, S. Paul, RelSim: an integrated method to identify disease genes using gene expression profiles and PPIN based similarity measure, Inf. Sci. 384 (2017) 110–125.
[28] J.C. Nacher, T. Akutsu, Structurally robust control of complex networks, Phys. Rev. E 91 (2015) 012826.
[29] J.C. Nacher, T. Akutsu, Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control, New J. Phys. 14 (2012) 073005.
[30] J.C. Nacher, T. Akutsu, Minimum dominating set-based methods for analyzing biological networks, Methods 102 (2016) 57–63.
[31] M.E.J. Newman, Spread of epidemic disease on networks, Phys. Rev. E 66 (2002) 016128.
[32] S. Pei, H.A. Makse, Spreading dynamics in complex networks, J. Stat. Mech. Theory Exp. 12 (2013) P12002.
[33] S. Pei, F. Morone, H.A. Makse, Theories for influencer identification in complex networks, Springer Nature, 2017.
[34] S. Pei, X. Teng, J. Shaman, F. Morone, H.A. Makse, Efficient collective influence maximization in threshold models of behavior cascading with first-order transitions, Sci. Rep. 7 (2017) 45240.
[35] S. Pei, L. Muchnik, J.S. Andrade Jr, Z. Zheng, H.A. Makse, Searching for superspreaders of information in real-world social media, Sci. Rep. 4 (2014) 5547.
[36] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (2001) 3200–3203.
[37] E.M. Rogers, Diffusion of Innovation, fourth ed., Free Press, 1995.
[38] M. Richardson, P. Domingos, Mining the network value of customers, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 61–70.
[39] S.H. Strogatz, Exploring complex networks, Nature 410 (2001) 268–276.
[40] P.G. Sun, X. Ma, Understanding control of network spreading from network controllability, J. Stat. Mech. Theory Exp. (2017) 093405.
[41] P.G. Sun, The human drug–disease–gene network, Inf. Sci. 306 (2015) 70–80.
[42] X. Teng, S. Pei, F. Morone, H.A. Makse, Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks, Sci. Rep. 6 (2016) 36043.
[43] S. Wuchty, Controllability in protein interaction networks, Proc. Natl. Acad. Sci. 11 (2014) 7156–7160.
[44] D. Wishart, C. Knox, A. Guo, et al., DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906.
[45] G. Yan, P.E. Vertes, E.K. Towlson, Y.L. Chew, D.S. Walker, W.R. Schafer, A.L. Barabsi, Network control principles predict neuron function in the C. elegans connectome, Nature 550 (2017) C519-C523.
[46] S. Zhao, S. Li, A co-module approach for elucidating drug-disease associations and revealing their molecular basis, Bioinformatics 28 (7) (2012) 955–961.
[47] R. Zhang, H.Y. Ou, C.T. Zhang, DEG, a database of essential genes, Nucleic Acids Res. 32 (2004) D271–D272.
[48] M. Zhao, J. Sun, Z. Zhao, TSGEne: a web resource for tumor suppressor genes, Nucleic Acids Res. 41 (2013) D970–D976.
[49] S. Zu, T. Chen, S. Li, Global optimization-based inference of chemogenomic features from drug-target interactions, Bioinformatics 31 (15) (2015) 2523–2529.