Physica A 433 (2015) 344–355

Contents lists available at ScienceDirect

Physica A

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

Using global diversity and local topology features to identify influential network spreaders

Yu-Hsiang Fu [a], Chung-Yuan Huang [b,*], Chuen-Tsai Sun [a]

[a] Department of Computer Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan

[b] Department of Computer Science and Information Engineering, School of Electrical and Computer Engineering, College of Engineering, Chang Gung University, 259 Wen Hwa 1st Road, Taoyuan 333, Taiwan

HIGHLIGHTS

- We combine global diversity and local features to identify influential nodes.
- A robust and reliable two-step framework is presented as a node ranking measure.
- Results from a series of experiments indicate our method performs well and stably.

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form 30 January 2015
Available online 16 April 2015

Keywords:
Node diversity
Entropy
Social network analysis
k-shell decomposition
SIR epidemic model

ABSTRACT

Identifying the most influential individuals spreading ideas, information, or infectious diseases is a topic receiving significant attention from network researchers, since such identification can assist or hinder information dissemination, product exposure, and contagious disease detection. Hub nodes, high betweenness nodes, high closeness nodes, and high *k*-shell nodes have been identified as good initial spreaders. However, few efforts have been made to use node diversity within network structures to measure spreading ability. The two-step framework described in this paper uses a robust and reliable measure that combines global diversity and local features to identify the most influential network nodes. Results from a series of Susceptible–Infected–Recovered (SIR) epidemic simulations indicate that our proposed method performs well and stably in single initial spreader scenarios associated with various complex network datasets.

1. Introduction

The network-spreading phenomenon is the focus of studies ranging from information diffusion via online social media sites, to viral marketing, to epidemic disease identification and control, to cascading failures in electrical power grids and the Internet, among many others [1–9]. Strategies for identifying key spreaders are being established and tested to accelerate information dissemination, increase product exposure, detect contagious disease outbreaks, and execute early intervention strategies [10]. Topological structure is a core concept in this network spreading identification process [1,2].

In social network analyses, centrality measures for identifying influential network nodes are broadly categorized as local or global [3,7,11]. Degree centrality, defined as the number of nodes that a focal node is connected to, measures node involvement in a network. However, techniques favored by most researchers for measuring the influence of network nodes fail to consider the importance of global topological structures. The two most widely used global centrality measures for

* Corresponding author. Tel.: +886 3 211 8800x3474; fax: +886 3 211 8700.
E-mail address: gscott@mail.cgu.edu.tw (C.-Y. Huang).

http://dx.doi.org/10.1016/j.physa.2015.03.042

overcoming these limitations are betweenness and closeness. Betweenness centrality, which assesses the degree to which a node lies on the shortest path between two other nodes, determines network flow. Closeness centrality is defined as the inverse sum of the shortest distances from a focal node to all other nodes. Influence is tied to the occupation of advantageous network positions. Three basic sources of advantages are high degree, high closeness, and high betweenness. In simple network structures, these advantages tend to vary individually. In complex networks, the potential exists for significant disjunctures among these position characteristics, meaning that a spreader's location may be advantageous in some ways and disadvantageous in others.

In addition to centrality measures, results from a k-shell decomposition analysis indicate that network nodes located in core layers are capable of spreading throughout a much broader area than nodes located in peripheral layers [1,2]. Although the spreading capability of each node differs, those with similar k-shell values are perceived as having equal importance. A method for ranking the network spreading ability of nodes in terms of degree centrality in identical k-shell layers for purposes of adjusting rank lists has been proposed [8]. To rank spreaders, a method referred to as mixed degree decomposition (MDD) adds otherwise ignored degree nodes to the decomposition process [3,6,12]. Still, researchers have shown a tendency to overlook the importance of network topology and node diversity, despite their positive correlations with factors such as community economic development [13]. Further, the entropy values of locations visited by users are positively correlated with the numbers of social ties those same users have in a social network [14]. The combined entropy values of node degree, betweenness, and closeness centralities have been applied to create complex network visualizations [15,16].

Inspired by past studies of network topology and node diversity, we used the entropy concept to develop a robust and reliable method for measuring the spreading capability of nodes, and for identifying super-spreader nodes in complex networks. This measure can be used to analyze the numbers of global network topological layers and local neighborhood nodes that are affected by specific individual nodes. Our assumption is that k-shell decomposition [1,2] can be used for purposes of global analysis, with nodes having high degrees of global diversity and local centrality capable of penetrating multiple global layers and influencing large numbers of neighbors in the local layers of a complex network. To measure node influence, we propose a two-step framework for acquiring global and local node information within complex networks. In the first step, global node information is obtained using algorithms such as a community detection algorithm for complex networks [5,17,18] or a k-shell decomposition algorithm for core/periphery network layers, after which entropy is used to evaluate the global diversity of network nodes. In the second step, local node information is acquired through the use of various types of local centrality, including degree centrality. Last, global diversity and local features are combined to determine node influence. In our experiments, spreading ability was measured as the total number of recovered nodes over time. We compared the spreading capability of the proposed measure and the local/global centralities of the social network using an SIR (susceptible–infective–recovered) epidemic simulation [2,19,20] with various social network datasets [21–25].

2. Background

To represent a complex social network, let an undirected graph $G = (V, E)$, where V is the node set and E the edge set of the network. Let $n = |V|$ indicate the number of network nodes and $m = |E|$ the number of edges. Network structure is represented as an adjacency matrix $A = \{a_{ij}\}$ and $a_{ij} \in R^n$, where $a_{ij} = 1$ if a link exists between nodes i and j, otherwise $a_{ij} = 0$.

Degree (or local) centrality is a simple yet effective method for measuring node influence in a complex network. Let $C_d(i)$ denote the degree centrality of node i. A high-degree centrality indicates a large number of connections between a node and its neighbors. $NB_h(i)$ denotes the set of neighbors of node i at a h-hop distance. The degree centrality of node i is therefore defined as

$$C_d(i) = |NB_h(i)| = \sum_{j=1}^{n} a_{ij} \tag{1}$$

where $|NB_h(i)|$ is the number of neighbors of node i at the h-hop distance; in most cases, $h = 1$ [7].

Betweenness centrality or dependency measures the proportion of the shortest paths going through a node in a complex network. Let $C_b(i)$ denote the betweenness centrality of node i. A high betweenness value indicates that a complex network node is located along an important communication path. Accordingly, the betweenness centrality of node i is defined as

$$C_b(i) = \sum_{s \neq t \neq v \in V} \frac{Q_{st}(i)}{Q_{st}} \tag{2}$$

where $Q_{st}(i)$ is the number of shortest paths from node s to node t through node i, and Q_{st} the total number of shortest paths from node s to node t [3,7,11].

Closeness (also known as global) centrality measures the average length of the shortest paths from one node to other nodes. Let $C_l(i)$ denote the closeness centrality of node i. A high closeness centrality value indicates that a node is located in

Fig. 1. The SIR (susceptible–infected–recovered) epidemic simulation model.

the center of a complex network, and that the average distance from that node to other nodes is shorter compared to nodes with low closeness centrality. The closeness centrality of node i is defined as

$$C_l(i) = \frac{1}{l_i}, \quad l_i = \frac{1}{n} \cdot \sum_{j=1}^{n} d_{ij} \tag{3}$$

where l_i is the average length of the shortest paths from node i to the other nodes, and d_{ij} is the distance from node i to node j [11].

The k-shell decomposition [1,2] iteratively assigns a k-shell layer value to every node in a complex network. During the first step, let $k = 1$ and remove all nodes where $C_d(n) = k = 1$. Following removal, the degrees of some remaining network nodes may be $k = 1$. Nodes are continuously pruned from the network until there are no $k = 1$ nodes. All removed nodes are assigned a k-shell value of $ks = 1$. The next step entails a similar process: let $k = 2$, prune nodes, and assign a k-shell value of 2 to all removed nodes. This procedure is repeated until all network nodes are removed and assigned k-shell indexes. The method reveals the significant features of a complex network—for example, all Internet nodes can be classified as nuclei, peer-connected components, or isolated components [1].

There are slight differences in the k-shell index and k-core layer concepts. The k-shell index is a global indicator representing the network core layer that a node is located in. A higher k-shell index represents inner core layer nodes that are more important than periphery layer nodes. The k-core layer is a sub-network consisting of nodes having ks k-shell indexes that exceed a given value k [1,2]. High k-shell index nodes are capable of infecting a larger number of neighbors than nodes with identical degree centrality values [2,6,8].

According to the global measures described above, node network positions are determined by analyzing whole network structures and the relative relationships of neighboring node network positions. Node network position information is acquired by computing the shortest paths of all pairs in terms of their betweenness centrality and the average length of all shortest paths in their closeness centrality. Decomposing k-shell network structure from periphery layer to core layer indicates the global property involving the entire network structure; in contrast, "local" indicators involve node information such as the number of connections.

The SIR epidemic model shown in Fig. 1 [2,19,20] has been widely used in multiple fields to study the theoretical spreading processes of information, rumors, biological diseases, and other phenomena within populations. The "infectious" concept is a general property of the spreading phenomena described above. Using epidemiology jargon, ideas and rumors can be represented as disease pathogens causing infected individuals to spread a disease among a larger population. The individual may have pathogen antibodies after recovering or receiving a vaccine, and therefore become incapable of reinfection. Due to this characteristic, the SIR epidemic model is widely used as a general-purpose model to describe the states of a disease and to study contagious spreading processes. Regarding a population's social network structure, the SIR model can be applied to network spreading dynamics, with disease transmission an example of a general network spreading phenomenon. The SIR model has been modified and extended (e.g., SEIR [26] and SIHR [27]) to study the spreading dynamics of different diseases and rumors within networks [28].

The SIR model consists of three states: susceptible (S), infective (I), and recovered (R). S set nodes are susceptible to information or diseases, I set nodes are capable of infecting neighbors, and R set nodes are immune and cannot be reinfected. Initially, almost all network nodes are in the S set, with a small number of infected nodes (sometimes a single individual) acting as spreaders. During each time step, I nodes infect their neighbors at a pre-established infection rate β, after which they become recovered nodes at a recovery rate of γ. Let $S(t)$ denote the number of susceptible nodes at time t, $I(t)$ the number of infected nodes at time t, $R(t)$ the number of recovered nodes at time t, and $\rho(t) = R(t)/N$ the proportion of immune nodes. The total number of nodes in an SIR model is $S(t) + I(t) + R(t) = n$.

3. The proposed measure

The influential social network spreader should satisfy two network topology conditions: high global diversity and high local features. First, node global diversity is determined by the network positions of neighbors. Greater differences in neighbor position information increase node diversity—that is, neighbors are distributed equally in all network layers (communities), otherwise they have similar network positions. There is an expectation that high global diversity nodes will spread information, ideas, or rumors very quickly in the early stages of the spreading process. Second, the local features of nodes are measured using the sum of neighbor connections. Here the expectation is that high local feature nodes will trigger an early and rapid accumulation of contagious transmissions among a large number of candidate nodes.

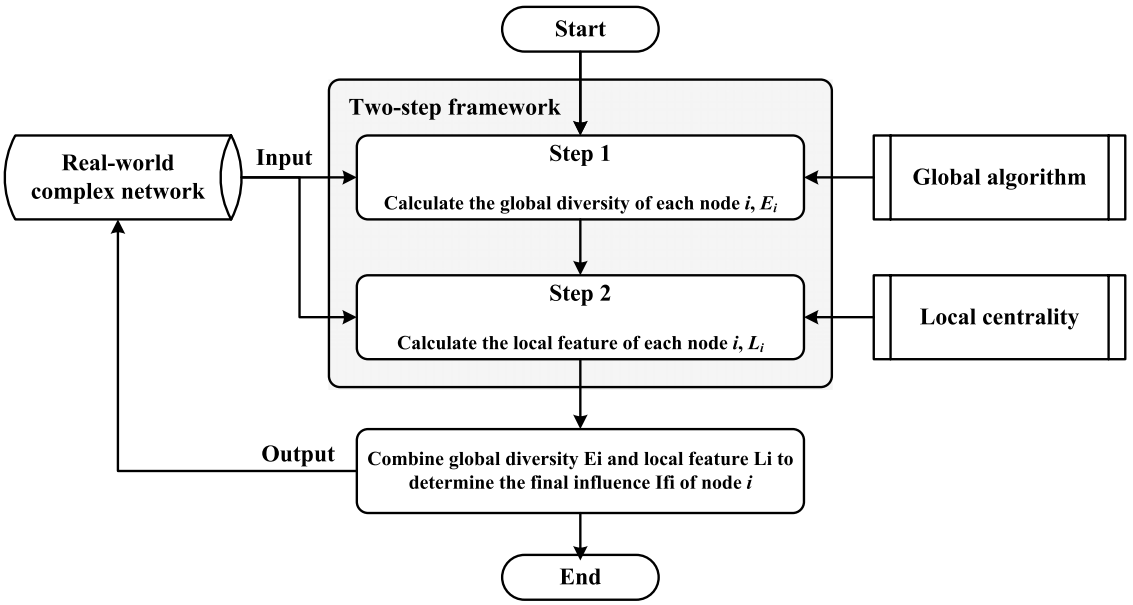

Fig. 2. The proposed two-step framework for computing the influence of network nodes.

The two-step framework shown in Fig. 2 is our proposed method for obtaining global and local node information in a complex network. In step 1, global algorithms (e.g., community detection, graph clustering, k-shell decomposition) are used to analyze the global features of nodes in a complex network. Results are used to compute the global diversity of nodes. In step 2, degree centrality is used to measure local node features. Last, global diversity and local features are combined to determine the final influence of complex network nodes.

In step 1, the k-shell decomposition method was used as an example for obtaining global information about nodes in a complex network. The k-shell values of nodes were obtained to calculate global diversity in terms of Shannon's entropy [29], which was then used to describe how many network layers are affected by a node. According to Eq. (4) definition, maximum entropy indicates a case in which a node is capable of connecting equally with all layers of a complex network, and a minimum entropy of 0 indicates a case in which all connections of a node are in the same layer of a complex network. As shown in Fig. 3, the k-shell entropy of node i, which ensures that its neighbors' k-shell values are much more diverse, is defined as

$$E_i\left(X_i\right) = -\sum_{j=1}^{ks_{\max}} p_i\left(x_j\right) \cdot \log_2 p_i\left(x_j\right) \tag{4}$$

$$p_i\left(x_j\right) = \frac{\left|x_j\right|}{\sum_{j=1}^{ks_{\max}} x_j} \tag{5}$$

$$\hat{E}_i\left(X_i\right) = \frac{E_i\left(X_i\right) - E_{\min}}{\log_2 ks_{\max} - E_{\min}} \tag{6}$$

where $X_i = \{1, 2, \ldots, ks_{\max}\}$ denotes the k-shell values of the neighbors of node i, $p_i\left(x_j\right)$ the probability of the x_j-core layer of neighbors, $\left|x_j\right|$ the number of nodes in the x_j-core layer of the complex network, and $\hat{E}_i\left(X_i\right)$ the normalized k-core entropy required for the case under consideration.

In step 2, the node's degree centrality is used to analyze the value of local features in the complex network; the degree centrality of neighbors is also considered. A high influence value indicates that a node and its neighbors have high degree centrality, in turn indicating that the node is capable of reaching the widest possible local range. The local feature of node i is defined as

$$L_i = \log_2\left(\sum_{j \in NB_{h=1}(i)} C_d\left(j\right)\right) \tag{7}$$

$$\hat{L}_i = \frac{L_i - L_{\min}}{L_{\max} - L_{\min}} \tag{8}$$

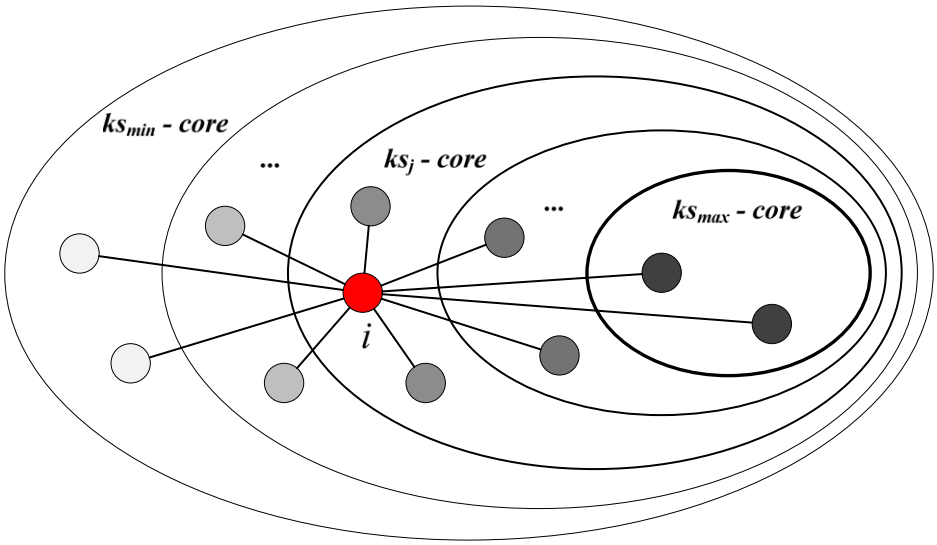

Fig. 3. An illustration of k-shell entropy to describe how neighbors are distributed in the network layers.

where $C_d(j)$ is the degree centrality of neighbor j and $NB_{h=1}(i)$ the neighbor set of node i at a h-hop distance. $L_i(i)$ can be extended to become a "neighbor's neighbor" version, meaning that all neighbors of node i with a 2-hop distance are considered, and that $\hat{L}_i(i)$ is the normalized local feature required for the case under consideration.

According to the definition of an influential spreader, E_i and L_i are considered simultaneously to maximize the spreading capability of node i in a complex network. The identified nodes are expected to connect to hub nodes in different network layers. Finally, global diversity E_i and local feature L_i are combined to denote IF_i, the final influence of node i, defined as

$$IF_i = E_i \cdot L_i. \tag{9}$$

Time and space complexities are important for some applications; those of the proposed two-step framework (including the combined step) are presented in Fig. 2. In step one, the time complexity associated with computing global diversity value E is $O(\langle k\rangle \cdot n)$, meaning that each node visits all of its neighbors to acquire k-shell values, where $\langle k\rangle$ is the average node degree centrality, and space complexity is denoted as the $O(n)$ of storing the E values of nodes. In step two, the time complexity of computing local feature value L is $O\left(\langle k\rangle^2 \cdot n\right)$, meaning that each node accumulates the sum of its neighbors' degree centrality values, and where space complexity is denoted as the $O(n)$ of storing the L values of nodes. In the combined step, the time complexity is $O(n)$, meaning that final influence IF values are E multiplied by L (or an adding operation if introducing the logarithm to equation [11]), and space complexity is denoted as the $O(n)$ of storing the IF values of nodes. Total time complexity is expressed as $O\left(\langle k\rangle^2 \cdot n + \langle k\rangle \cdot n + n\right) = O\left(\left(\langle k\rangle^2 + \langle k\rangle + 1\right) \cdot n\right)$ and total space complexity as $O(n + n + n) = O(3n)$, which can be reduced to $O(n)$ by storing E, L and IF values in the same memory space and ignoring the combined step to reduce time complexity to $O\left(\left(\langle k\rangle^2 + \langle k\rangle\right) \cdot n\right)$.

4. Experimental results and discussion

Basic complex network properties and results from an analysis of giant connected component (GCC) network structures are shown in Table 1 and Fig. 4, respectively. The correlation coefficient r in Fig. 4 is classified as high ($r \geqq 0.7$), medium ($0.4 \leqq r < 0.7$), or low ($r < 0.4$) [30]. Based on these classifications, Fig. 4(a) shows the correlation between E and the k-shell index. Communication network types identified as having high correlations include Email-Contacts ($r = 0.82$), Email-Enron ($r = 0.78$), and PolBlogs ($r = 0.74$)—that is, in these networks high global diversity nodes, located in inner core layers, communicate with other network core layers. In Fig. 4(b), regarding the correlation between E and closeness, the majority of network datasets (11 of 13) have at least medium correlations, meaning that high global diversity nodes tend toward network centers and have lower communication costs than other nodes. In Fig. 4(c), regarding the correlation between L and closeness, 12 of the 13 datasets had high correlations, indicating that high local feature nodes also tend toward network centers and can be used to both approximate closeness centrality and reduce computational costs. In Fig. 4(d), regarding the correlation between L and coreness [31], all network datasets had high correlations ($r \geqq 0.9$), meaning that L can also be used to approximate coreness—that is, the ability to reach the widest possible local range. In Fig. 4(e), regarding correlations between E and L values, the majority of datasets had at least medium correlations, meaning that higher correlations indicate a large number of core layers and neighbors being affected in a network.

In summary, high correlations were found in all results for the PolBlogs dataset. The network in the original PolBlogs dataset was a directed network of hyperlinks among weblogs about US politics [22], meaning that frequent interaction and

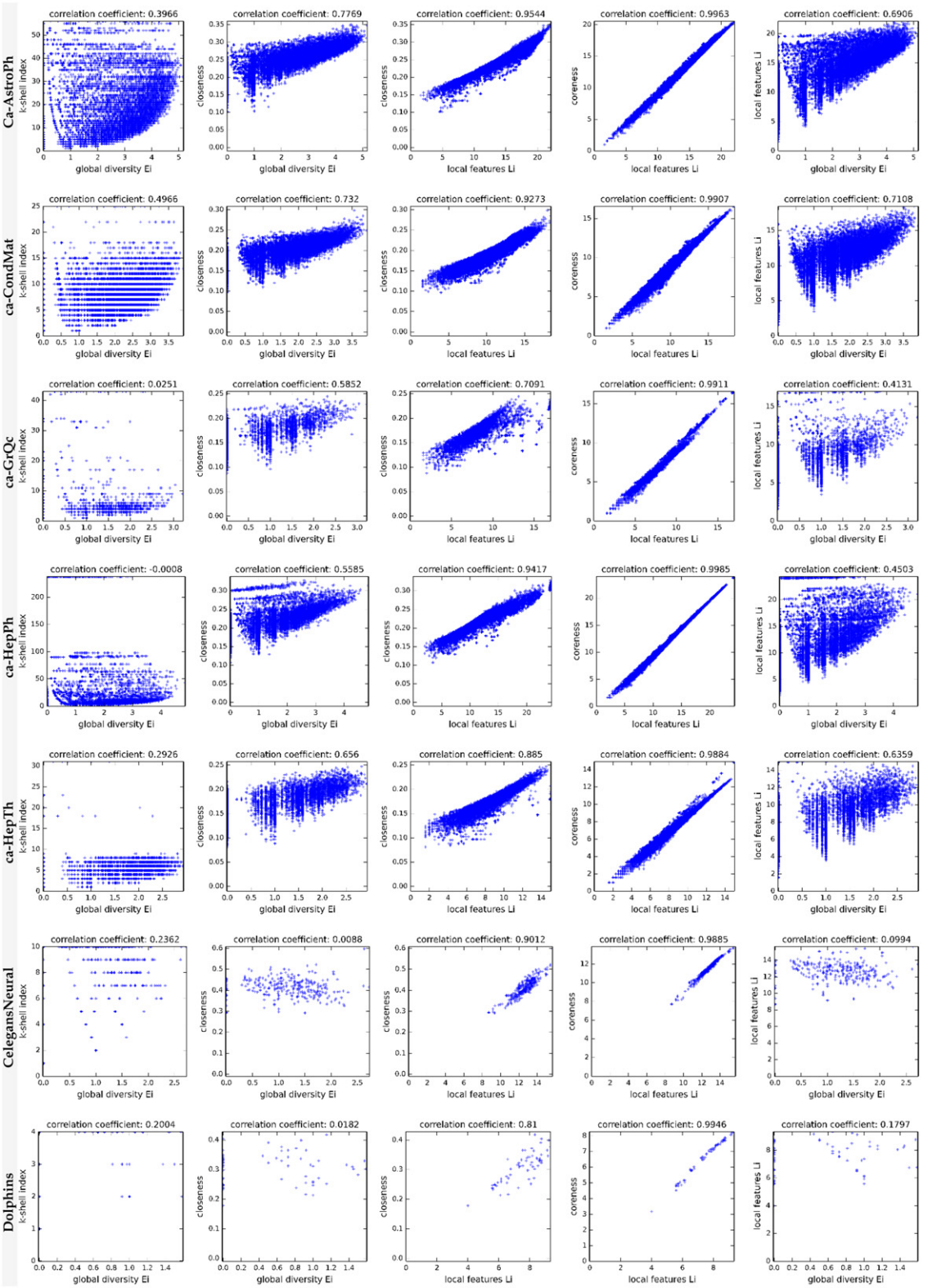

Fig. 4. Statistical results for the complex networks used in our spreading experiments. The scatter plot sub-figures are arranged from left to right in the following order: (a) global diversity and k-shell index, (b) global diversity and closeness, (c) local features and closeness, (d) local features and coreness, and (e) global diversity and local features. Correlation coefficients between x–y axis attributes are at the top of each sub-figure.

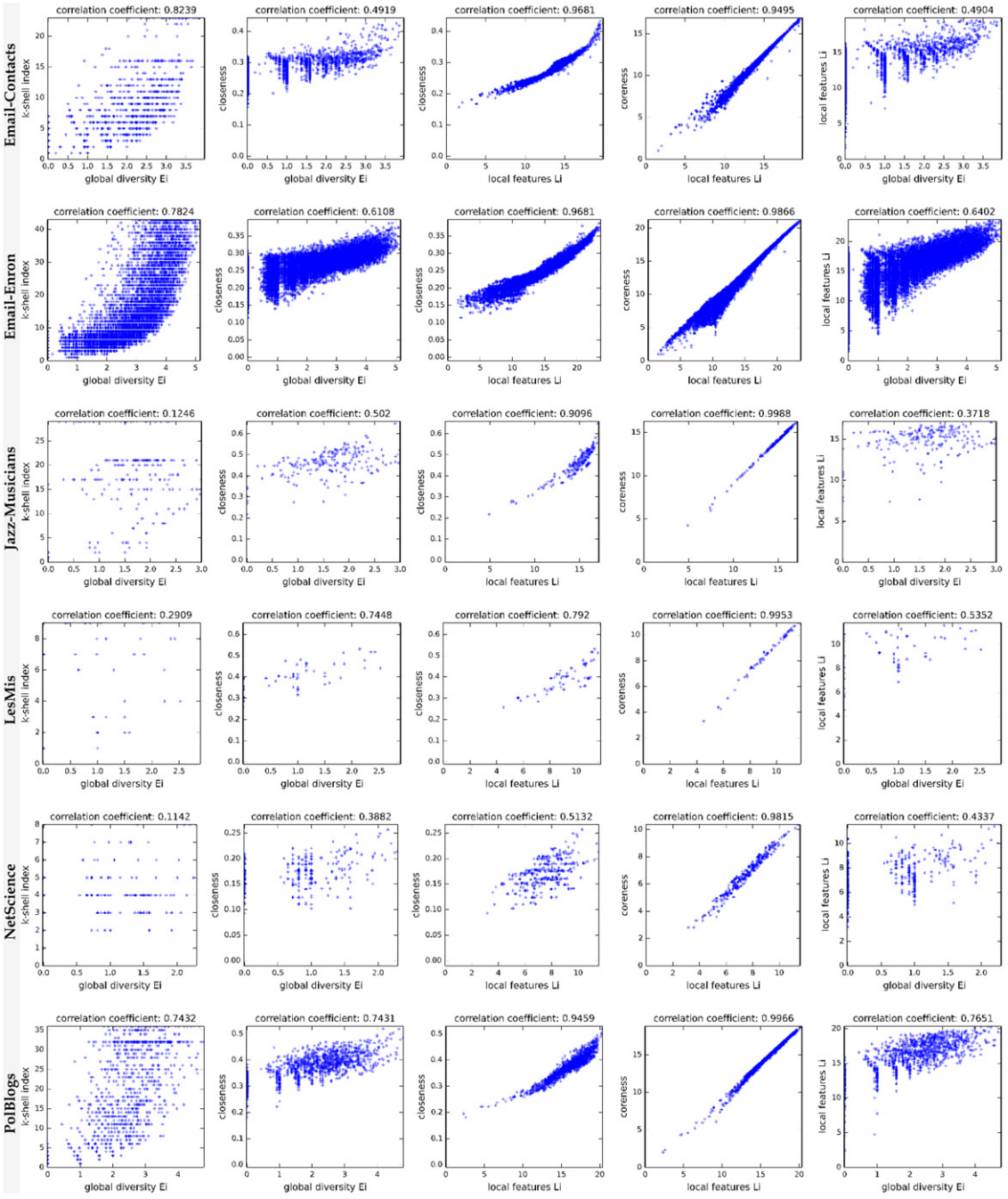

Fig. 4. (*continued*)

communication occurred inside the network. In this type of network, network spreaders can be the most influential initial nodes in the spreading of diseases, information, or rumors. However, according to the data shown in Table 1 and Fig. 4, we cannot offer clear evidence indicating that basic network properties or correlation coefficients can be used to categorize network structures in order to characterize complex network spreading dynamics.

For the spreading experiments we used three network dataset classifications: scientific collaboration, traditional social, and other. Measures were degree, betweenness, and closeness centralities; k-shell decomposition; coreness [31]; PageRank [34]; and our proposed method. Spreading experiment and SIR epidemic model parameters were as follows: 5000 simulations for each network dataset, with each simulation consisting of 50 time steps and with the top-1 node for each measure serving as the initial spreader. The β infection rates of the SIR epidemic model used in our experiments are

Table 1
Properties of the real-world networks used in this project. We considered only the largest connected network components when the original network was disconnected.

Network Type	Network	Description	N	E	$\langle c \rangle$	$k_{\max}$	$\langle k \rangle$	$ks_{\max}$	$\langle ks \rangle$	H	r	β_{thd}	β
Collaboration	ca-AstroPh	Co-authorship in astro-ph of arxiv.org.	17 903	196 972	0.63	504	22.00	56	13.11	2.99	0.20	0.02	0.03
	ca-CondMat	Co-authorship in cond-mat category.	21 363	91 286	0.64	279	8.55	25	5.12	2.63	0.13	0.04	0.05
	ca-GrQc	Co-authorship in gr-qc category.	4 158	13 422	0.56	81	6.46	43	4.58	2.79	0.64	0.06	0.15
	ca-HepPh	Co-authorship in hep-ph category.	11 204	117 619	0.62	491	21.00	238	15.93	6.23	0.63	0.01	0.05
	ca-HepTh	Co-authorship in hep-th category.	8 638	24 806	0.48	65	5.74	31	3.41	2.26	0.24	0.08	0.12
Social	Jazz-Musicians	Collaborations among 1920s jazz musicians.	198	2 742	0.62	100	27.70	29	17.27	1.40	0.02	0.03	0.04
	Email-Contacts	Email contacts in the Computer Science Department of University College, London.	12 625	20 362	0.11	576	3.23	23	1.65	34.25	−0.39	0.01	0.05
	Email-Enron	Enron email dataset.	33 696	180 811	0.51	1383	10.73	43	5.73	13.27	−0.12	0.01	0.05
Other	C. elegansNeural	Neural network of the C. elegans nematode.	297	2 148	0.29	134	14.46	10	7.98	1.80	−0.16	0.04	0.06
	Dolphins	Frequent associations among 62 dolphins.	62	159	0.26	12	5.13	4	3.16	1.33	−0.04	0.15	0.15
	LesMis	Les Miserables network.	77	254	0.57	36	6.60	9	4.73	1.83	−0.17	0.08	0.08
	NetScience	Network science collaborations.	379	914	0.74	34	4.82	8	3.47	1.66	−0.08	0.12	0.20
	PolBlogs	Political blogs.	1 222	16 714	0.32	351	27.36	36	14.82	2.97	−0.22	0.01	0.02

$H = \langle k \rangle / \langle k \rangle^2$, degree heterogeneity [32].
$\beta_{thd} = \langle k \rangle / \langle k \rangle$, theoretical epidemic threshold [33].

shown in Table 1. According to at least one previous study, a large infection rate makes no difference in terms of spreading measures [2]. To assign a suitable infection rate for each network dataset, infection rates were determined by comparing the theoretical epidemic threshold β_{thd} with the number used in referenced studies [31]. The recovery rate was always set at $\gamma = 1$ [2], meaning that every node in the infected set I entered recovered set R immediately after infecting its neighbors. Network-based simulation steps were as follows:

Step 1. During initialization, all nodes are in the S state except for the (top-1) initial spreader, which is in the I state.

Step 2. During each time step t, each node in the I state randomly infects its neighbors according to an infection rate β, and then enters the R state (i.e., $\gamma = 1$). The $\rho(t)$ cumulative incidence of contagion is the number of recovered nodes calculated at the end of each time step.

Step 3. Repeat step 2 until the maximum time step requirement is satisfied—or, if necessary, when the I state set is empty [2].

Experimental results and details regarding spreading dynamics (i.e., $\rho(50)$ results) are shown in Fig. 5 and Table 2. We found that the leading group could be defined as the spreading result of measures that are larger than the maximum result minus an inaccuracy factor of 1%, expressed as

$$LG = \{m \mid p_m(t) \geq (p_{\max}(t) - err * p_{\max}(t)), m \in M \text{ and } err \in [0,\ 1]\} \tag{10}$$

where M is the set of measures used in the experiment, $p_{\max}(t)$ the maximum result at time t, err the inaccuracy rate (0.01), and time step $t = 50$.

The number of recovered nodes $\rho(t)$ was used to measure and rank the spreading capability of various measures. The leading group can help determine the stability of a measure for identifying the influence of nodes in different networks. The measures inside the leading group had approximately the same spreading capability. The average rank in Table 2 was used to interpret the expected rank in different networks: a measure with a lower average rank was viewed as having better discrimination in terms of identifying good spreaders.

According to the inside leading group number (an indicator of measure stability), our proposed method performed well in terms of identifying the most influential nodes in different networks. Based on our experiment results, the proposed method is capable of identifying nodes that serve as good spreaders with global diversity in a complex network. In addition to being within the leading group, the method also had a better ranking compared to other measures within that group. The identified influence spreaders were capable of reaching large numbers of network nodes through their diverse global connections, of affecting network layers, and of exerting a maximum spreading effect on network layers. The degree centrality of a node

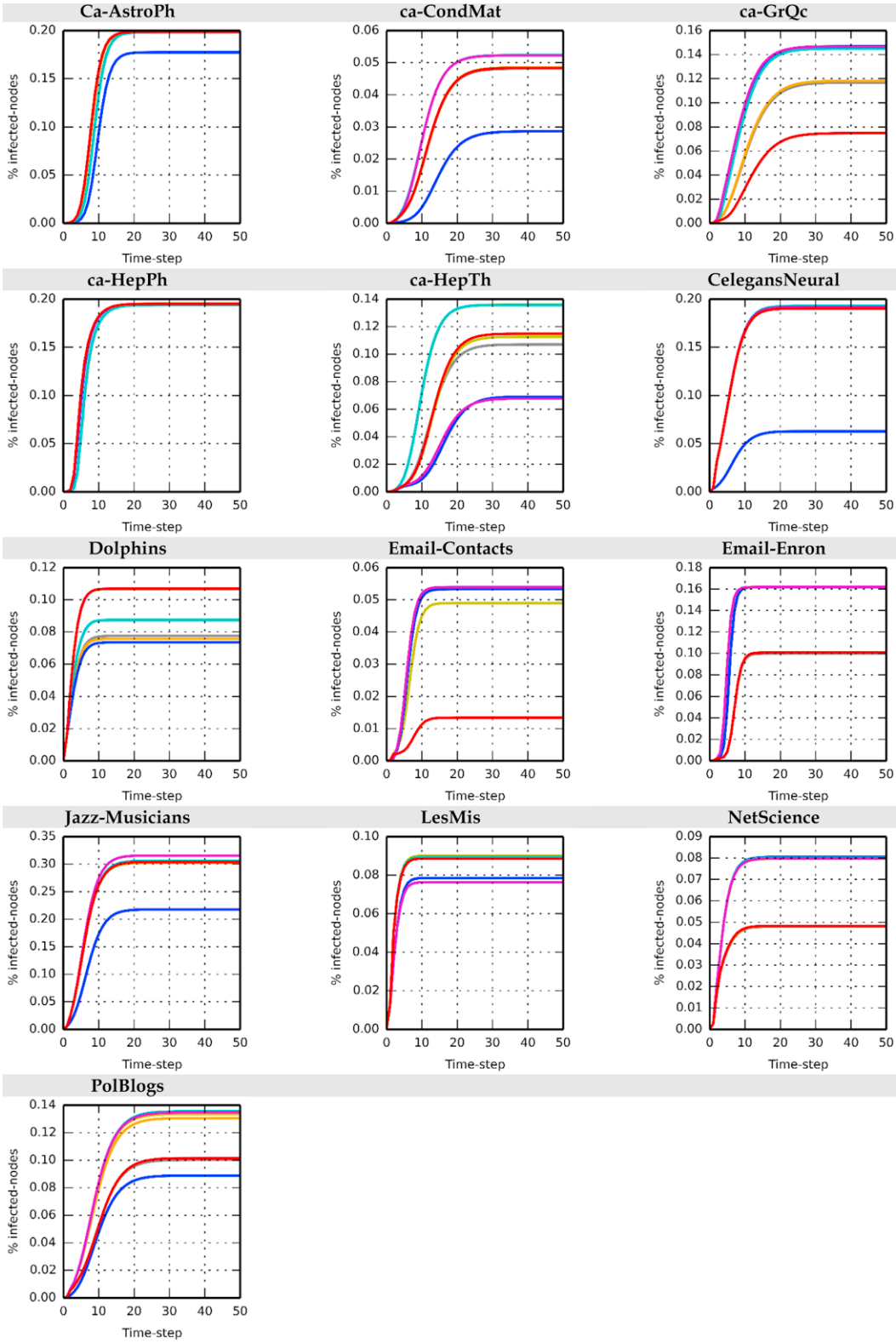

Fig. 5. Spreading dynamics results for different networks. Measurements shown are betweenness (gray), closeness (orange), degree (yellow), *k*-shell (blue), proposed method (cyan), coreness (magenta) and PageRank (red).

Table 2
Comparison of simulation results from different measures (including our proposed method) in experiments using the real-world networks shown in Table 1.

Network	$\rho(t=50)$						
Giant connected component (GCC)	Degree	Betweenness	Closeness	k-core	Coreness	PageRank	Proposed
ca-AstroPh	**0.1988**$_1$	**0.1983**$_6$	**0.1986**$_4$	0.1774$_7$	**0.1987**$_2$	**0.1987**$_2$	**0.1984**$_5$
ca-CondMat	**0.0486**$_3$	**0.0482**$_6$	**0.0485**$_4$	0.0286$_7$	**0.0523**$_2$	**0.0484**$_5$	**0.0523**$_1$
ca-GrQc	**0.1462**$_3$	0.1168$_6$	0.1182$_5$	**0.1469**$_1$	**0.1467**$_2$	0.0748$_7$	**0.1450**$_4$
ca-HepPh	**0.1952**$_1$	**0.1942**$_6$	**0.1952**$_1$	**0.1951**$_4$	**0.1950**$_5$	**0.1952**$_1$	**0.1941**$_7$
ca-HepTh	0.1126$_4$	0.1070$_5$	**0.1359**$_1$	0.0691$_6$	0.0678$_7$	0.1148$_3$	**0.1357**$_2$
Jazz-Musicians	0.3034$_5$	0.3036$_4$	0.3024$_6$	0.2177$_7$	**0.3156**$_1$	0.3040$_3$	**0.3060**$_2$
Email-Contacts	**0.0490**$_6$	**0.0533**$_5$	**0.0539**$_1$	**0.0534**$_4$	**0.0539**$_1$	0.0134$_7$	**0.0539**$_1$
Email-Enron	0.1008$_5$	0.0997$_7$	**0.1619**$_1$	**0.1618**$_4$	**0.1619**$_1$	0.1007$_6$	**0.1619**$_1$
C. elegansNeural	**0.1909**$_5$	**0.1922**$_2$	**0.1916**$_4$	0.0627$_7$	**0.1918**$_3$	**0.1902**$_6$	**0.1928**$_1$
Dolphins	**0.1067**$_3$	0.0775$_5$	0.0757$_6$	0.0734$_7$	**0.1069**$_1$	**0.1069**$_1$	0.0874$_4$
LesMis	**0.0901**$_1$	**0.0894**$_2$	**0.0893**$_4$	0.0785$_6$	0.0764$_7$	**0.0886**$_5$	**0.0894**$_2$
NetScience	**0.0797**$_4$	0.0479$_7$	0.0482$_6$	**0.0805**$_1$	**0.0799**$_3$	0.0483$_5$	**0.0802**$_2$
PolBlogs	**0.1336**$_3$	0.1003$_6$	**0.1306**$_4$	0.0888$_7$	**0.1347**$_2$	0.1014$_5$	**0.1354**$_1$
Inside leading group number:	10	6	9	5	11	6	12
Average rank:	3.3846	5.1538	3.6154	5.2308	2.8462	4.3077	2.5385

Bold, measurement result is inside the leading network group.
Subscript, rank of network in the measurement.

and its neighbors can be used to maintain the number of contact nodes in the local layer of a complex network. However, important differences were noted among measures. For example, the closeness measure performed well in the top-1 position of the ca-HepTh, ca-HepTh, Email-Contacts and Email-Enron networks (Table 2), but not in the Jazz-Musicians, Dolphins or NetScience networks. Since the characteristic the measure wanted to capture may not have been sufficiently strong in those networks, the most influential spreaders could not be identified.

Global diversity $\langle k\rangle$ and local feature $\langle k\rangle$ averages were used to determine the expected effect of certain combinations of factors, to classify network nodes into four categories, and to explain why the two definition conditions should be satisfied. For node type (a), $E_i \geq \langle k\rangle$ and $L_i \geq \langle k\rangle$—that is, those nodes identified as influential spreaders were capable of using the advantages of global diversity and local features to affect most network layers, and to infect a large number of neighbors (e.g., hubs) during the early stages of the spreading process. Accordingly, this node type is said to have maximum spreading ability. For node type (b), $E_i \geq \langle k\rangle$ and $L_i < \langle k\rangle$—that is, nodes affected many different network layers, but were only capable of infecting a small number of neighborhood nodes, meaning that the spreading range was dependent on whether neighbors were located in the inner core layer or throughout the high k-core layer; accordingly, this node type has high potential for spreading activity. For node type (c), $E_i < \langle k\rangle$ and $L_i \geq \langle k\rangle$—that is, nodes could only affect one or a small number of network layers, but they infected large numbers of neighbors. Here the spreading range was dependent on whether neighbors were located in a high k-core layer rather than a cluster; this type of node has the same potential as type (b) for spreading activity. For node type (d), $E_i < \langle k\rangle$ and $L_i < \langle k\rangle$—that is, nodes could only affect one or a small number of network layers and infect a small number of neighbors, indicating that the spreading range was constrained to periphery layers, and indicating minimum spreading capability.

Although the proposed method underscores the robustness and stability of identifying the most influential nodes of different networks, its limitations may be dependent on the type of node involved. For example, for node type (c), the maximum k-shell values of a network will be lower and network sizes considerably smaller in the absence of global diversity in a complex network. As shown in Table 2, nodes with high global diversity in the Dolphins network could not be identified. In that case, the spreading ability of nodes identified by our proposed method decreased to the degree centrality (ignoring the first term), and the influence of nodes was limited to local network layers. In the absence of global diversity, Eq. (9) becomes $IF_i \approx L_i$, which favors local network layers (i.e., degree centrality). In addition, the $\hat{\hat{E}}_i(X_i)$ normalized global diversity values produced by our proposed method are similar to the participation coefficients reported by Teitelbaum et al. [35], and the high global diversity of nodes that we observed are similar to those of connector hubs and kinless hubs, both of which have distinct participation coefficients.

Weighting schemes are usually added to equations to adjust the weights of different terms. However, we wanted to avoid problems associated with parameter optimization among different networks, therefore we made a purposeful decision to keep our equation simple and to not add a weighting scheme in this particular study. Another advantage of this decision is that the weights of different equation terms may be determined according to the network topology structure. For example, assume that the weights of global diversity E and local feature L are introduced into the IF equation. In the Dolphins network, global diversity E is absent and therefore can be ignored, making $IF \approx L$ and allowing the use of local feature L to determine node importance. This is similar to setting the assumed weights of E and L as $\alpha = 0.1$ and $(1-\alpha) = 0.9$. However, there are disadvantages involving the detailed adjustment of weights and parameter optimization, which cannot be applied to different network datasets for the specified purposes.

A growing number of complex network and social network researchers are using data-driven approaches to study the spreading phenomena of online social media and websites. For example, Pei et al. [36,37] have tracked information flow

involving actual diffusion processes using Facebook, LiveJournal, Twitter, and other online social network datasets. They used the sum of nearest neighbors' degrees to approximate a global measure (i.e., k-core decomposition), and found that the approximate measure performed almost as well as a k-shell index, and outperformed degree centrality and PageRank. Assuming that a network topology structure and actual information flow can be fully obtained, our proposed two-step framework can be applied to a data-driven approach as follows:

Step 1: Deploy the search algorithm (i.e., BFS) to extract the actual diffusion range of users and to calculate their global diversity E using a k-shell index or estimated values. Global diversity E in a data-driven approach is called actual-based diversity. In this paper global diversity is called topology-based diversity.

Step 2: Use local features to calculate how many neighbors may be infected during the spreading process.

Step 3: Combine actual-based global diversity E_{actual} and local feature L to acquire the network's node influence IF_{actual}.

A data-driven approach has great potential for directly analyzing actual user diffusion process data and for obtaining node influence values, as opposed to using a theoretical SIR or SIS model to analyze the spreading ability of nodes in a network. However, in cases involving a network topology structure and partial user data, a topology-based approach and theoretical epidemic model may still be useful in analyzing and identifying influential spreaders in a complex social network.

5. Conclusion

In this paper we described our proposal for a two-step framework for calculating the influence of network nodes. In step 1, a global algorithm was used to analyze global node information, with the entropy concept from information theory used to assist in measuring node global diversity. Affected global network layers could be identified using k-shell entropy. In step 2, the degree centralities of nodes and their neighbors were considered simultaneously in order to maintain the number of affected neighbors in the local layer of a complex network. In the final step, global diversity and local features were combined to determine the influence of nodes in the network. Our experimental results indicate that the proposed method performed well and maintained stability in leading groups of different network datasets. In other words, the proposed method is capable of identifying the most influential nodes as initial spreaders that disseminate information, ideas, or diseases in different networks.

Our plans are to add considerable details to our analysis, to introduce a sophisticated method for evaluating spreading ability, and to clarify how the proposed method is affected by network structure. For example, global algorithms such as community detection algorithms can be used to analyze and obtain global information on community network structures, and to determine how factors such as position and node role [35] affect the degree to which spreaders distribute information or diseases throughout a complex network. We also plan to study strategies associated with multiple initial spreaders in networks. Since overlapping infected areas for selected spreaders must be minimized [2], a multiple initial spreader scenario may either accelerate or hinder spreading within a complex network.

Acknowledgments

The work was supported in part by a Grant from the Republic of China National Science Council (MOST-103-2221-E-182-052). The work was supported in part by the High Speed Intelligent Communication (HSIC) Research Center, Chang Gung University, Taiwan.

References

[1] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, A model of Internet topology using k-shell decomposition, Proc. Natl. Acad. Sci. 104 (27) (2007) 11150–11154.
[2] 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 (11) (2010) 888–893.
[3] B. Hou, Y. Yao, D. Liao, Identifying all-around nodes for spreading dynamics in complex networks, Physica A 391 (15) (2012) 4012–4017.
[4] D. Chen, L. Lü, M.-S. Shang, Y.C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A 391 (4) (2012) 1777–1787.
[5] X. Zhang, J. Zhu, Q. Wang, H. Zhao, Identifying influential nodes in complex networks with community structure, Knowl.-Based Syst. 42 (2013) 74–84.
[6] A. Zeng, C.J. Zhang, Ranking spreaders by decomposing complex networks, Phys. Lett. A 377 (14) (2013) 1031–1035.
[7] S. Gao, J. Ma, Z. Chen, G. Wang, C. Xing, Ranking the spreading ability of nodes in complex networks based on local structure, Physica A 403 (2014) 130–147.
[8] J.G. Liu, Z.-M. Ren, Q. Guo, Ranking the spreading influence in complex networks, Physica A 392 (18) (2013) 4154–4159.
[9] B. Doerr, M. Fouz, T. Friedrich, Why rumors spread so quickly in social networks, Commun. ACM 55 (6) (2012) 70–75.
[10] N.A. Christakis, J.H. Fowler, Social network sensors for early detection of contagious outbreaks, PLoS One 5 (9) (2010) e12948.
[11] M. Newman, Networks: An Introduction, Oxford University Press, 2010.
[12] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1–2) (1938) 81–93.
[13] N. Eagle, M. Macy, R. Claxton, Network diversity and economic development, Science 328 (5981) (2010) 1029–1031.
[14] J. Cranshaw, E. Toch, J. Hong, A. Kittur, N. Sadeh, Bridging the gap between physical location and online social networks, in: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, 2010, pp. 119–128.
[15] E. Serin, S. Balcisoy, Entropy based sensitivity analysis and visualization of social networks, in: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM, 2012, pp. 1099–1104.
[16] S.J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 2010.
[17] M. Rosvall, C.T. Bergstrom, An information-theoretic framework for resolving community structure in complex networks, Proc. Natl. Acad. Sci. 104 (18) (2007) 7327–7331.
[18] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826.

[19] R. Pastor-Satorras, A. Vespignani, Epidemic dynamics and endemic states in complex networks, Phys. Rev. E 63 (6) (2001) 066117.
[20] C.Y. Huang, C.L. Lee, T.H. Wen, C.T. Sun, A computer virus spreading model based on resource limitations and interaction costs, J. Syst. Softw. 86 (3) (2012) 801–808.
[21] Stanford large network dataset collection. http://snap.stanford.edu.
[22] Network datasets collected by Mark Newman. http://www-personal.umich.edu/~mejn/netdata/.
[23] Network datasets collected by Hernán Alejandro Makse. http://lisgi1.engr.ccny.cuny.edu/~makse/soft_data.html.
[24] Jazz musicians network dataset. http://www.infochimps.com/datasets/jazz-musicians-network.
[25] P. Basaras, D. Katsaros, L. Tassiulas, Detecting influential spreaders in complex, dynamic networks, Computer 46 (4) (2013) 24–29.
[26] C.Y. Huang, T.H. Wen, Y.S. Tsai, FLUed: A novel four-layer model for simulating epidemic dynamics and assessing intervention policies, J. Appl. Math. 2013 (2013) e325816.
[27] L. Zhao, J. Wang, Y. Chen, Q. Wang, J. Cheng, H. Cui, SIHR rumor spreading model in social networks, Physica A 391 (7) (2012) 2444–2453.
[28] C.Y. Huang, C.T. Sun, H.C. Lin, Influence of local information on social simulations in small-world network models, J. Artif. Soc. Soc. Simul. 8 (4) (2005).
[29] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[30] C.P. Dancey, J. Reidy, Statistics Without Maths for Psychology, fifth ed., Prentice Hall, Harlow, England, New York, 2011.
[31] J. Bae, S. Kim, Identifying and ranking influential spreaders in complex networks by neighborhood coreness, Physica A 395 (2014) 549–559.
[32] H.B. Hu, X.F. Wang, Unified index to quantifying heterogeneity of complex networks, Physica A 387 (14) (2008) 3769–3780.
[33] C. Castellano, R. Pastor-Satorras, Thresholds for epidemic spreading in networks, Phys. Rev. Lett. 105 (21) (2010) 218701.
[34] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical Report, Stanford InfoLab, 1999.
[35] T. Teitelbaum, P. Balenzuela, P. Cano, J.M. Buldú, Community structures and role detection in music networks, Chaos 18 (4) (2008) 043105.
[36] S. Pei, H.A. Makse, Spreading dynamics in complex networks, J. Stat. Mech. 2013 (12) (2013) P12002.
[37] S. Pei, L. Muchnik, J.J.S. Andrade, Z. Zheng, H.A. Makse, Searching for superspreaders of information in real-world social media, Sci. Rep. 4 (2014).