IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Spreading dynamics in complex networks

This content has been downloaded from IOPscience. Please scroll down to see the full text.

J. Stat. Mech. (2013) P12002

(http://iopscience.iop.org/1742-5468/2013/12/P12002)

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 200.17.35.21
This content was downloaded on 20/02/2014 at 15:08

Please note that terms and conditions apply.

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Spreading dynamics in complex networks

Sen Pei[1,2] **and Hernán A Makse**[2]

[1] LMIB and School of Mathematics and Systems Science, Beihang University, Beijing, 100191, People's Republic of China
[2] Levich Institute and Physics Department, City College of New York, New York, NY 10031, USA
E-mail: peisen@smss.buaa.edu.cn and hmakse@lev.ccny.cuny.edu

Received 12 September 2013
Accepted 3 November 2013
Published 6 December 2013

Online at stacks.iop.org/JSTAT/2013/P12002
doi:10.1088/1742-5468/2013/12/P12002

Abstract. Searching for influential spreaders in complex networks is an issue of great significance for applications across various domains, ranging from epidemic control, innovation diffusion, viral marketing, and social movement to idea propagation. In this paper, we first display some of the most important theoretical models that describe spreading processes, and then discuss the problem of locating both the individual and multiple influential spreaders respectively. Recent approaches in these two topics are presented. For the identification of privileged single spreaders, we summarize several widely used centralities, such as degree, betweenness centrality, PageRank, k-shell, etc. We investigate the empirical diffusion data in a large scale online social community—LiveJournal. With this extensive dataset, we find that various measures can convey very distinct information of nodes. Of all the users in the LiveJournal social network, only a small fraction of them are involved in spreading. For the spreading processes in LiveJournal, while degree can locate nodes participating in information diffusion with higher probability, k-shell is more effective in finding nodes with a large influence. Our results should provide useful information for designing efficient spreading strategies in reality.

Keywords: network dynamics, random graphs, networks, online dynamics, communication, supply and information networks

J. Stat. Mech. (2013) P12002

© 2013 IOP Publishing Ltd and SISSA Medialab srl
1742-5468/13/P12002+21$33.00

Contents

1. Introduction

As a widespread process, spreading describes many important activities in the real world, ranging over the outbreak of epidemics [1]–[4], the spread of news and ideas [5]–[8], the diffusion of technique innovations [9]–[11], the promotion of commercial products [12]–[15], and the rise of political movements [16]–[19]. Understanding the mechanism behind the global spread of an epidemic or information is fundamental for applications in a variety of fields as diverse as epidemiology [3, 4], viral marketing [14, 15], collective dynamics [20]–[22] and the robustness of networks [23]–[25]. In real life, the diffusion of a contagious disease or a piece of information usually underlies individuals' contacts. Take influenza as an example. The transmission of the influenza virus mainly depends on the direct contact of infected persons and susceptible people. As people's interactions are responsible for these spreading processes, the position of a person in the complex network, which is formed according to individuals' social relations, usually determines the spreading capability of this person. As the empirical research has shown, the complex networks in reality, which describe various systems in the fields of social science [26]–[30], neuroscience [31]–[33], ecology [34, 35] and economics [36, 37], are by no means randomly connected. On the contrary, they display many nontrivial topological features such as heavy-tailed degree distributions [38], small world effects [39], high clustering coefficients [40], self-similarity [41, 42] and community structures [43, 44].

J. Stat. Mech. (2013) P12002

Therefore, people in social networks with such rich topological structures should vary considerably in their spreading abilities.

For the people in social networks, a few influential spreaders are able to start large-scale diffusion and thus become more important than the other persons in the spreading process [45]. How to identify these privileged nodes is of great significance for applications across various domains. For instance, the knowledge of influential spreaders is crucial for the design of efficient strategies to control the outbreak of epidemics [46, 47]. Also, targeting the vital people in information dissemination is helpful in conducting successful campaigns in commercial product promotions [14, 15]. Due to its utmost importance in practice, the problem of searching for influential spreaders in complex networks has attracted much attention.

In the issue of identifying super spreaders, there are two distinct subtopics: searching for individual influential spreaders [45] and a set of nodes that can maximize the influence [12]. In epidemic outbreaks or rumor diffusion, the spreading usually starts from a single source node. Therefore, monitoring the influential individual originators in social networks is important in preventing or decelerating the spreading. For locating individual spreaders, usually the topological or dynamical measures are utilized. We first rank the nodes according to specific measures and then select the nodes ranking top as influential spreaders. In the field of viral marketing, however, it is usually a set of nodes that are selected to start the propagation. The choice of the multiple originators should maximize the final influence since the goal of viral marketing is to persuade more people to buy the commercial products with less cost. A trivial way to find the best multiple spreaders is to select the nodes which are influential as individual spreaders. However this attempt usually fails for the reason that the top individual nodes tend to have large overlaps in their infected population [45]. In fact, the issue of finding a k-set nodes that can lead to maximal influence is a NP-hard optimization problem [48]. The solution mainly relies on heuristic algorithms based on dynamic models in complex networks.

While the overall topic of spreading is too wide even to introduce briefly, we will mainly focus on the two subtopics described above. In this paper, we first introduce some of the most important theoretical models for spreading, which are widely used in the research of influential spreaders. Then the measures for identifying individual influential spreaders are discussed. We discuss their features, calculation complexity and performance in locating the best single spreaders. After that, we present some progress in finding multiple influential spreaders with heuristic algorithms. Finally, we show the empirical results based on extensive data collected from large-scale online communities.

2. Theoretical models for spreading

The early approach of designing mathematical models that describe spreading processes started in the fields of sociology and epidemiology [49]–[52]. Formulating theoretical models that capture real diffusion is helpful for us to understand how a disease or information spreads across a population. More importantly, the spreading models can be used to predict the outcome of spreading, thus providing instructions to accelerate or decelerate the diffusion processes. After the initial works, many models are proposed in the applications of a series of social and biological phenomena.

doi:10.1088/1742-5468/2013/12/P12002

J. Stat. Mech. (2013) P12002

In the spreading models, there are usually two important elements to be specified. The first one is the underlying network, which describes how individuals interact with one another. The network is recorded by a graph $G(N, E)$ with N nodes and E edges. If node i and j have a chance to contact in reality, then there exists a social link between them. As has been shown, the topological structure of the social network can dramatically affect the outcome of spreading [46, 53]. Another element that is vital for models is the spreading rule by which information or disease diffuses from one person to another. According to the diffusion strategy, existing models of spreading typically fall into two categories: independent interaction models and threshold models. We will introduce these two types of models in detail in this section and explain some of their features.

2.1. Independent interaction models

In epidemic spreading, each time an infected node contacts with a susceptible node, there is a chance that the susceptible person gets infected. Based on this fact, independent interaction models assume that each interaction results in contagion with an independent probability. Specifically, whenever a susceptible person j is exposed to an infected person i, j will get infected with a probability p_{ij}, which is not affected by the contacts with other nodes. With such dynamics, independent interaction models mimic the contagious process directly and imply the fact that spreading underlies people's interactions. The more a susceptible individual contacts with infected people, the higher probability they will be infected. The susceptible–infected–recovered (SIR) and susceptible–infected–susceptible (SIS) models from epidemiology [1]–[3], and the Bass model from the innovation diffusion literature [54] are examples of independent interaction models. Here we mainly discuss the well-known SIR and SIS models. Different models and generalizations can be found in the references [1]–[4], [52].

As classical mathematical models for epidemic spreading, the SIR and SIS models were first proposed by epidemiologists [49]–[52]. Since the mechanisms of SIR and SIS models are suitable to describe various spreading processes, such as disease spreading, idea propagation, and innovation diffusion, they have long been used in the study of spreading. In SIR model, each person can be in one of three possible states [1], susceptible (S), infected (I), or recovered (R). Susceptible individuals are healthy persons that may catch the disease. Infected people stand for the persons who have got the disease and are able to spread it to susceptible individuals. A person is recovered if they have been cured of the disease and become immune to it. In the classical SIR model, there are two adjustable parameters: the transmission probability λ and the recovery probability μ. The states of individuals evolve as follows:

$$\mathrm{S}(i) + \mathrm{I}(j) \rightarrow_{\lambda} \mathrm{I}(i) + \mathrm{I}(j), \tag{1}$$

$$\mathrm{I}(i) \rightarrow_{\mu} \mathrm{R}(i), \tag{2}$$

where i and j are two neighbors in the social network. While the SIR model is suitable to describe the spreading of disease with immunity, there are many diseases that an individual can catch more than once. In this case, the SIS model, in which we only consider the susceptible (S) and infected (I) states, can better describe the contagion. The contagion process of the SIS model is the same as equation (1), while equation (2) is replaced by $\mathrm{I}(i) \rightarrow_{\mu} \mathrm{S}(i)$. Generalizations of SIR and SIS models can be implemented by imposing a distribution of the transmission rate λ and the recovery rate μ.

doi:10.1088/1742-5468/2013/12/P12002

J. Stat. Mech. (2013) P12002

According to the distinct dynamics, there are fundamental differences in the outcomes of SIR and SIS models. In the SIR model, the infection will eventually die out, because once an individual becomes immune to the disease, he/she will never get infected again. Whereas, for SIS models, people can be infected many times. So the disease can reach an endemic state, where a certain fraction of population are kept infected. Considering this difference, when we measure the result of SIR model, we are interested in the fraction of individuals who have ever caught the disease. While for SIS model, it is the fraction of infected nodes persisting in the endemic state that we are concerned with.

Besides the spreading strategy, it is meaningful to discuss the impact of the underlying social network structure on the spreading results. In the early research, both SIR and SIS models were considered within the homogeneous mixing hypothesis [1], where the infectious and susceptible people contact with each other randomly. In this case, the underlying social network is actually an Erdós–Rényi random graph. The most important observation under this condition is the emergence of an epidemic threshold [55]. Take the SIR model as an example (without loss of generality we set $\mu = 1$). When a single node becomes infectious among a susceptible population, the epidemic threshold is given by $\lambda_c = 1/\langle k \rangle$, where $\langle k \rangle$ is the average number of connections of nodes. If $\lambda > \lambda_c$, the disease will infect a finite fraction of the population. On the other hand, if $\lambda < \lambda_c$, the fraction of infected individuals will tend to zero in the limit of a very large population.

As we all know, transmission networks in real life are by no means totally random. For example, the degree distribution of the sexual contact network is found to be power-law [56]. To this end, several important works have been done to understand the effects of a nontrivial network structure on the spreading outcomes [46], [57]–[62]. For the SIR model on uncorrelated graphs with a generic degree distribution $P(k)$ and a finite average degree $\langle k \rangle$, the epidemic threshold is defined by $\lambda_c = \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$ [46, 59, 60, 63, 64]. Apparently, for networks with $\langle k^2 \rangle < \infty$, the threshold has a finite value. Whereas, for networks with a strongly fluctuating degree distribution, the infinity of $\langle k^2 \rangle$ results in a vanishing epidemic threshold for large-scale networks. Analogously, the absence of an epidemic threshold in scale-free networks with the power-law exponent $2 < \gamma \leq 3$ was reported as well [46, 60]. Apart from the topological structures mentioned above, other complex topologies have been considered, such as high clustering [65, 66], small world [58, 67, 68], degree correlation [53, 69], etc. These works provide more insights into the interplay between network structure and epidemic outcomes for SIR and SIS models.

2.2. Threshold models

Although the independent interaction models can describe epidemic spreading in reality, numerous phenomena in economics and sociology [50], [70]–[76] are better described by the threshold models. In these applications, people tend to adopt a new behavior or information only if a certain fraction of their neighbors have already done so [77, 78]. In this case, the effect of a single interaction is no longer independent, but strongly depends on other exposures.

The simplest threshold model is the Linear Threshold Model. In this model, each node is assigned a threshold value, which is the fraction of neighbors required for it to adopt the new behavior or information. On each link (i, j), we define a weight to reflect the influence that j exerts on i. In the spreading process, some initial node S starts out adopting the new behavior. A node is defined as active if it is following the new behavior. At a given time,

J. Stat. Mech. (2013) P12002

any inactive node becomes active if the sum of weights from its active neighbors exceeds its threshold. In subsequent times, the activation of some nodes may cause other nodes to adopt the new behavior, and such a process can be applied repeatedly. This phenomenon is usually used to explain cascading behavior in social science [77, 78]. More generally, based on the Linear Threshold Model, we can assign each node a threshold function instead of using the weighted sum in the state updating. Such a General Threshold Model is more general since it can reflect any type of threshold rule.

Compared with the independent interaction model, the threshold model actually incorporates the memory of past exposure history. Therefore, the result of a single interaction is determined by other interactions. This radical difference turns out to have significant impact on the spreading dynamics. With such a threshold model, global cascading which is triggered by a small number of initial originators is observed [77]. Also it has been shown in a model with memory of exposure history that the final state of the spreading is controlled by only two parameters: P_1 and P_2, which stand for the probability that a node becomes infected due to one and two contacts respectively [79]. This indicates that the interplay of single contacts in the threshold model can lead to different dynamics with independent interaction models.

Apart from the models mentioned above, there are also other variations describing spreading in complex networks, including the standard rumor model proposed by Daley and Kendal [80]–[84], the voter model [85]–[88], the strategic game models [89]–[92], etc. All these models help us understand the mechanism of spreading in various domains and many profound results have been applied in reality.

3. Searching for individual influential spreaders

In order to find effective predictors for individual influential spreaders, various measures are designed to rank the nodes according to their statuses in spreading. Most of the proposed measures are determined by nodes' topological features as well as an assuming spreading mechanism. Here we will introduce some of the most important predictors that are widely used to quantify nodes' spreading ability.

In the context of social science, the topology of a social network is represented by an adjacency matrix $A = \{a_{ij}\}_{N\times N}$, where the element $a_{ij} > 0$ if there exists a link from j to i and $a_{ij} = 0$ otherwise. For an undirected network, A is a symmetric matrix with $a_{ij} = a_{ji}$. If the network is weighted, the element a_{ij} represents the weight of the link from j to i. Actually, the adjacency matrix A fully describes the topological structure of the social network. Once one has the adjacency matrix, it is possible to calculate the following measures for each node. Some of the measures only need the local information, i.e. the properties of a node's neighbors, while some others require the complete structure of the social network. Due to the large size of modern social networks, the calculation of global measures imposes a great challenge in the research of social networks.

3.1. Degree

In real social networks, it is observed that while most of the people have a small or moderate number of connections to other individuals, there are a very few hubs that maintain extremely large numbers of social relations. Such a phenomenon is described

J. Stat. Mech. (2013) P12002

by a power-law (or heavy-tailed) degree distribution. For an unweighted graph, degree is the number of links connecting to a node. For a weighted graph, it is defined as the sum of weights from edges connecting to a node. In an undirected network, according to the adjacency matrix of a graph, the degree $k(i)$ for a node i can be computed as follows:

$$k(i) = \sum_{j=1}^{N} a_{ij}. \tag{3}$$

In the case of a directed network, we usually define two separate measures of degree centrality, namely indegree and outdegree. Concretely, indegree is the number of links directed to the node and outdegree is the number of links that the node directs to others. Indegree and outdegree are defined as:

$$k_{\text{in}}(i) = \sum_{j=1}^{N} a_{ij}. \tag{4}$$

$$k_{\text{out}}(i) = \sum_{j=1}^{N} a_{ji}. \tag{5}$$

Computing degree centrality for all the nodes in a dense network takes complexity $O(V^2)$. However, for a network with a sparse adjacency matrix, which we usually encounter in reality, its computational complexity is reduced to $O(E)$, making the degree centrality a feasible measure even for very large networks. Albeit it is a local measure, degree is efficient in finding important nodes in many situations. For example, for epidemic spreading in scale-free networks, hubs are more likely to be infected and can lead to large scale diffusion [46]. Also in complex networks with a broad degree distribution, such as the Internet, power grid or other infrastructure networks, intentional attacks on the hubs can result in a rapid breakdown of the whole structure [23, 25]. Such fragility of scale-free networks under intentional attacks indicates that hubs play a prominent role in the structure stability. However, not all hubs are guaranteed to be super spreaders. For instance, if a hub is located in the periphery of the network, its spreading ability would be limited [45]. After all, degree only captures the number of the nearest neighbors of a node. In fact, the spreading capability of the neighbors can also affect the nodes' importance in spreading significantly. Therefore, measures involved with more information are desired to improve the performance of degree. Even so, due to its easy accessibility and relatively satisfactory performance, degree is still used as an effective predictor of influential spreaders in many applications.

3.2. Betweenness and closeness centralities

Betweenness and closeness centrality are two well-known ranking measures in social science [93, 94]. Both of them are proposed based on the assumption that information tends to traverse the network from the originator to the destination through the shortest path.

In the social network theory, betweenness centrality is defined as a measure of how many shortest paths cross through this node [93, 94]. For a network $G = (V, E)$, the

J. Stat. Mech. (2013) P12002

betweenness centrality of node i, denoted by $C_B(i)$ is defined as

$$C_B(i) = \sum_{s \neq i \neq t \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}, \tag{6}$$

where σ_{st} is the number of shortest paths between nodes s and t, and $\sigma_{st}(i)$ is the number of shortest paths between s and t which pass through node i. With this definition, the nodes with large betweenness centrality usually hold the vital positions in the shortest pathways between large numbers of pairs of nodes.

Apart from the betweenness centrality, in a connected network, there exists a natural distance metric between all pairs of nodes. The farness [95] of a node s is defined as the sum of its distances to all other nodes, and its closeness centrality is defined as the inverse of the farness. Thus, the smaller closeness a node has, the lower its total distance to all other nodes. Precisely, closeness centrality of node i is defined as [95]:

$$C_C(i) = \frac{1}{\sum_{t \in V \setminus i} d_G(i,t)}, \tag{7}$$

where $d_G(i,t)$ is the shortest distance between i and t. In fact, closeness centrality can be viewed as a measure of how long it will take for a piece of information to spread from a given node to other reachable nodes through the shortest paths in the network. The smaller a node's closeness is, the faster the information diffuses from this node.

However, on an unconnected network, the closeness centrality is not well defined. Since the distance between any two unreachable nodes is infinity, the closeness centrality of all nodes in an unconnected graph would be 0. To solve this problem, a modified version of the classic closeness, residual closeness [96], is proposed. The residual closeness of node i is defined as

$$C_{\mathrm{R}}(i) = \sum_{t \in V \setminus i} 2^{-d_G(i,t)}. \tag{8}$$

In general, betweenness and closeness centralities can identify crucial nodes in transportation. Take the betweenness centrality as an example, the nodes with large betweenness usually hold the vital positions in the pathways between pairs of nodes. If such nodes are intentionally attacked, the overall efficiency of spreading will be heavily damaged, since the increase of path length would make it difficult for a piece of information to spread to other nodes. In the networks with heavy-tailed degree distribution, hubs usually serve as intermediate nodes in the shortest paths between nodes [97]. So hubs are inclined to have large betweenness centrality. Besides, the nodes connecting two separate communities also have large betweenness centrality. Such nodes, although not necessarily being well connected, play the role of connecting bridges in the transportation between the nodes in two communities. Applications of betweenness centrality include computer and social networks [98], biology [99]–[101], transport [102], scientific cooperation [103] and so forth.

Comparing with degree, betweenness and closeness centralities care more about the global structure. Therefore, the calculation of betweenness and closeness requires the complete network structure. Moreover, both betweenness and closeness centralities involve calculating the shortest paths between all pairs of vertices on a graph, which is a rather time-consuming task. The classic algorithm finding shortest paths between all pairs of

J. Stat. Mech. (2013) P12002

nodes is the Floyd–Warshall algorithm [104], and it will take a complexity $O(V^3)$. Later on, some more efficient algorithms were developed for specific types of networks. On sparse networks, the Johnson algorithm requires $O(V^2 \log V + VE)$ time to compute the betweenness centrality [105]. Brandes has proposed a more efficient algorithm. For unweighted sparse networks, it has a complexity of $O(VE)$ [106]. Even with the fastest algorithm, for large-scale online social networks with tens of millions of nodes, such as Twitter and Facebook, it is usually infeasible to get the betweenness or closeness centralities in a reasonable time.

3.3. Eigenvector and PageRank centralities

Eigenvector centrality was first introduced in the research of sociology [107], where it was used to measure the influence of a person in a social network. The main idea behind eigenvector centrality is that a node's importance is not only determined by itself, but also affected by its neighbors' importance. A node connecting to important nodes will make itself also important. With this idea, the eigenvector centrality of vertex i can be defined as:

$$e(i) = \frac{1}{\lambda} \sum_{j=1}^{N} a_{ij} e(j), \tag{9}$$

where λ is a constant and a_{ij} is the entry of the adjacency matrix $A = \{a_{ij}\}_{N \times N}$. Actually, this equation can be rewritten in vector notation as

$$\mathbf{A}\mathbf{e} = \lambda \mathbf{e}. \tag{10}$$

In the matrix theory, there will be many different eigenvalues λ for which an eigenvector solution exists. However, when we quantify the influence of a node, it is required that the measure should be positive. According to the Perron–Frobenius theorem, only the largest eigenvalue can lead to such a centrality measure [108]. Clearly, eigenvector centrality not only depends on the degree of the nodes, but also on their neighbors' eigenvector centrality. Due to this recursive property, eigenvector centrality can reflect the global features of the network.

As a generalization and variation of eigenvector centrality, PageRank was originally introduced to rank web pages in the World Wide Web (www) [109]. As a successful ranking algorithm, it is not only adopted by webpage search engines such as Google, but also used in ranking the importance of elements in a wide range of applications, such as scientific ranking [110]–[112], gene research [113, 114], traffic and transportation [115], ecological systems [116], and even lexical semantics [117]. Compared with eigenvector centrality, PageRank introduces a small probability of random jumping to handle walking traps on a graph. The PageRank of a node in a network can be calculated from

$$p_t(i) = \frac{1-\alpha}{N} + \alpha \sum_{j} \frac{a_{ij} p_{t-1}(j)}{k_{\text{out}}(j)}, \tag{11}$$

where $k_{\text{out}}(j)$ is the number of outgoing links from node j and α is the jumping probability. Equation (11) actually describes a random walk process: a random walker moves along the links of the network with probability α, and jumps to a randomly selected node with probability $1-\alpha$. $p_t(i)$ is the probability that node i is visited by the random walker at

J. Stat. Mech. (2013) P12002

time t. As time t increases, the probability $p_t(i)$ will converge to a stationary probability $p(i)$. This value is defined as the PageRank, which is used to determine its ranking relative to other nodes. In the calculation, the conventional choice of α is 0.85. Different choices of α can affect the ranking results.

From a calculation aspect, both eigenvector and PageRank centralities can be computed efficiently by power iteration [118, 119]. Initially assign each node with the same score, and then iterate according to the corresponding update equations. The result usually converges quickly in iterations. So eigenvector and PageRank centralities can be applied to large-scale networks.

Based on the classic PageRank algorithm, several variations are proposed. One is the so-called LeaderRank [120]. On the basis of PageRank, LeaderRank introduces a ground node g, which has two directed links to every node in the original network. In this way, the network will become strongly connected. More importantly, LeaderRank is a parameter-free algorithm, thus getting rid of the influence of parameters. Although LeaderRank stems from PageRank, it is reported to be more stable to noisy data containing spurious and missing links. In the condition where spammers create fake links to obtain a high rank, LeaderRank performs more reliably than PageRank in ranking users. Also an extension of the PageRank algorithm, TwitterRank [121], is proposed to measure the influence of users in Twitter. TwitterRank measures the influence taking both the topical similarity between users and the link structure into consideration. Experimental results show that TwitterRank outperforms the one Twitter currently uses and other related algorithms, including the classical PageRank and topic-sensitive PageRank [122].

3.4. k-shell index

In graph theory, the k-shell index describes the location of a person in the social network [123]–[125]. The k-shell index of a node is obtained by a procedure called k-shell decomposition, where we successively prune nodes in the network layer by layer. Concretely, the decomposition starts by removing nodes with degree $k = 1$. After that, some nodes may have only one link left. So we continue pruning the network iteratively until there are no nodes with $k = 1$. The removed nodes fall into a k-shell with index $k_S = 1$. With a similar method, we iteratively remove the next k-shell $k_S = 2$ and higher k-shells until all nodes are pruned. In the decomposition procedure, each node is assigned with a k-shell index. The periphery of the network corresponds to small k_S and the nodes with high k_S define the core of the network.

Compared with the degree k, the k_S index provides a different type of information. By definition, a given layer with index k_S can be occupied with nodes of degree $k \geq k_S$. For random model networks, a strong correlation between k and the k_S index of a node exists: the nodes with lower degree are inclined to stay in the periphery of the network while the core region is mainly occupied by hubs. Therefore, both the degree and k-shell provide similar information. In real networks, however, this relation is often not true. In real networks hubs may have very different k_S values and can be located both in the periphery or in the core of the network [45].

Recently, it was reported that for SIR and SIS modeling, the most influential spreaders, which can lead to large scale epidemics, are located in the inner core of the network (large k_S region) [45]. The authors performed SIR and SIS models on a series of real networks, and found that the spreading processes that originate from high k-shell nodes have

J. Stat. Mech. (2013) P12002

larger average infected population than those starting from nodes with high degree and betweenness centrality. These results indicate that the k-shell index of a node is a better predictor of spreading influence than the commonly adopted degree and betweenness centrality. When a spreading process starts in the core of the network, the epidemic or information can diffuse through many pathways to the remaining part of the network. Moreover, it has been shown that nodes with high k_S are more easily infected and will be infected earlier than other nodes. The nodes located in the core region of network tend to have well-connected neighbors, and the neighbors of their neighbors are also prone to have large degree. It is these well-connected neighbors that make the nodes in the core region more efficient in spreading. Clearly, k-shell decomposition requires the complete network structure. Once we have the adjacency matrix, k-shell decomposition can be performed with complexity $O(E)$ [126]. Therefore, the k-shell index can be applied to large-scale networks.

Despite its effectiveness, k-shell also have several defects. In some situations, k-shell is limited due to the lack of resolution. For example, in networks with a tree structure, or networks formed by growth models [38], the k-shell index only contains a few discrete values. Such degeneracy of the k-shell index limits its predictive accuracy. These limitations have been identified also in some empirical studies of spreading [127]–[129].

Based on k-shell decomposition, several improvements or alternatives are proposed. Zeng [130] proposed a mixed degree decomposition (MDD) procedure in which both the residual degree (number of links between the remaining nodes) and the exhausted degree (number of links between the removed nodes) are considered. In each step of the MDD procedure, the nodes are removed according to the mixed degree $k_i^{(\mathrm{m})} = k_i^{(\mathrm{r})} + \lambda \times k_i^{(\mathrm{e})}$, where $k_i^{(\mathrm{r})}$ is the residual degree, $k_i^{(\mathrm{e})}$ is the exhausted degree, and λ is a tunable parameter between 0 and 1. When $\lambda = 0$, the MDD method becomes the k-shell method, while when $\lambda = 1$, the MDD method is equivalent to the degree. By simulating the epidemic spreading process on real networks, it is shown that the MDD method can improve the performance of k-shell. Another method based on the k-shell index is the μ-power community index (μ-PCI) [131], which is a balance of coreness and betweenness centrality. The metric is computed as follows: the μ-PCI of a node i is equal to k, such that there are up to $\mu \times k$ nodes in the μ-hop neighborhood of i with degree greater than or equal to k, and the rest of the nodes in that neighborhood have a degree less than or equal to k. Due to its computational complexity, the authors only present the results for $\mu = 1$. By modeling on real networks, it is shown the 1-PCI exhibits steady and reliable behavior. As 1-PCI values increase, influence also continuously increases until maximum infection is reached.

3.5. Path counting

Path counting was first proposed as the accessibility metric [133, 134]. The basic idea is to count the number of all possible walks of arbitrarily length departing from the source node. Recently, the concept of path counting was exploited to quantify the importance of individual roles in collective dynamics [132]. A measure called dynamical influence (DI) is proposed to quantify a node's influence in various dynamical processes. The DI is calculated as the leading left eigenvector of a characteristic matrix that records the topology and dynamics. Specifically, in the characteristic matrix M, the entry M_{ij} stands for the influence that node j exerts on node i. The ith entry of the left

J. Stat. Mech. (2013) P12002

eigenvector of M corresponding to the largest eigenvalue is defined as the DI of node i. Computationally, DI can be easily calculated by a power method. Starting from a uniform vector $w^{(0)} = (1, 1, \ldots, 1)$, we multiply it with higher and higher powers of M (l is an integer):

$$w^{(l)} = (1, 1, \ldots, 1)M^l. \tag{12}$$

Notice that the ith entry of $w^{(l)}$ is the number of all possible walks of length l departing from node i. This explains the idea of path counting behind the definition of DI.

This framework applies to a variety of dynamical models, including epidemic spreading models, the Ising model [135, 136], and diffusive processes such as the voter model [137] or phase coupled oscillators [138]. In the SIR model, DI is shown to be a good predictor of spreading efficiency at the critical regime, outperforming the predictions made by degree, k-shell index and betweenness centrality.

Another approach based on path counting is a method to approximate the number of infections resulting from a given initially infected node in a network of susceptible individuals [139]. This method directly considers the spreading process and provides estimation of the actual number of infections with probability analysis. The derivation of the impact of vertex i, i.e. the estimated number of infections given that vertex i was infected first, is a little bit lengthy, so we will not introduce the details here. The computation of node i's impact requires the number of walks from i to all other nodes of arbitrary length k with l repeated vertices, which is the bottleneck of the computation complexity for this method. In reality, to avoid long computing time, the impact of vertices can be estimated by imposing a maximal walk length L. Simulations show that very good results can be obtained with short path lengths $L = 4$. We should note here that the exact definition of the impact of nodes in this method relies on the assumed spreading models. So different models may result in different definitions.

Apart from the progress mentioned above, a method measuring node spreading power by expected cluster degree is proposed [140]. The authors quantify the spreading ability of a source node by the expectation of the degree of a disease cluster starting from it. The degree of a cluster of nodes is defined as the number of edges that connect nodes within and outside the cluster. Define the expected reach of node i, $ER_X(i)$, as the expectation of the degree of the infected cluster after X contagions starting from i. In practice, the $ER_X(i)$ can be obtained by counting all possible clusters of infected nodes which could appear after X infections starting from i and then taking the average cluster degree. This method was applied to the recent Ebola outbreak in Uganda, and it predicts that Ebola is unlikely to spread globally. Similar to other path counting methods, computational expense is still a major concern of $ER_X(i)$. This problem can be solved by imposing a maximal X value: it is shown that $X = 3$ is sufficient to determine the outcome [140].

4. Finding the most influential sets of nodes

In real-world applications, there are many situations where we need to find a small set of nodes that can spread information to the largest number of nodes in the network. For example, in viral marketing, a company tries to promote a new commercial product using word-of-mouth effects. The best strategy is to persuade more people to buy the product with a limited advertising budget. In this case, how to find the most influential sets of

J. Stat. Mech. (2013) P12002

nodes that can lead to maximal influence in complex networks becomes a fundamental issue.

The problem of finding the most influential multiple spreaders is different from the one locating single influential spreaders. For spreading processes originating from a set of nodes simultaneously, the distance between these originators from each other should be taken into consideration. This is because the nodes influenced by the origins may have great overlap. Since the methods for single influential spreaders are not guaranteed to find nodes that are far enough [45], we cannot just select the top k nodes using the predictors designed for single super spreaders. Actually, this issue has been abstracted as a fundamental algorithmic problem in computer science [141]. For any spreading process, there exists a influence function $f(\cdot)$ defined on a set of nodes S. Assuming that S is the set of originators, $f(S)$ is the expected number of infected nodes at the end of the spreading process. With this interpretation, the problem becomes to maximize $f(S)$ for the set S with k nodes. In fact, this is a rather hard computational problem. Kempe *et al* [48] has proved that for a generic class of threshold models, it is NP-hard to find the optimal set S. Therefore, what we can do is to find suboptimal results with heuristic algorithms.

The basic idea behind the construction of heuristic algorithms lies in the fact that, when we add nodes to the originators' set S, usually the spreading will not increase significantly, but once the right nodes are added, the process suddenly spreads widely. Such a property is described as submodularity mathematically [48]. It has been shown that most instances of threshold models are submodular [48, 142]. For models with such a property, Kempe *et al* [48] developed a greedy search algorithm: starting from a randomly selected node, each time we add the node that can lead to maximal increase of spreading if we include it in S, until the desired size of S is reached. Making use of the classical theorem of Nemhauser *et al* [143], they proved analytically that the solution of this greedy strategy is within 63% of optimal for several classes of models. This is the first provable approximation guarantees for heuristic algorithms for this problem. However, the spread estimation procedures, which are usually implemented by Monte Carlo simulations, are quite time-consuming, thus limiting the efficiency of this algorithm.

Based on these fundamental results, several improvements have been developed. To reduce the Monte Carlo simulations to estimate spread, Leskovec *et al* [144] exploited submodularity and proposed a 'cost-effective lazy forward' (CELF) optimization to the simple greedy algorithm. The main idea is that the marginal gain of a node in the current iteration cannot exceed its marginal gain in previous iterations. With this idea, CELF optimization significantly reduces the number of calls made to the spread estimation procedure. Later on, Chen *et al* [145] proposed an improved version of the original greedy algorithm, NewGreedy. More recently, Goyal *et al* [146] developed the algorithm CELF++, which is an extension to CELF that further reduces the number of spread estimation calls. Other approaches can be found in the literature [147]–[152]. These heuristic algorithms improve the computational efficiency of the original greedy algorithms.

Besides the greedy strategies for submodular models, an efficient algorithm based on message passing in statistical physics was proposed recently [153]. A message-passing algorithm is an efficient method to deal with problems in statistical physics and combinatorial optimization [154]–[156]. To solve the spread optimization problem, the authors mapped it on a high-dimensional static constraint-satisfaction model. Then

J. Stat. Mech. (2013) P12002

they developed efficient message-passing algorithms to find a solution to the spread maximization problem. Compared with the greedy algorithm, this new approach takes into account the cooperative characteristics, which are fundamental in real systems. With analytic and algorithmic results on random graphs as well as a real-world network, it is shown for a wide range of irreversible dynamics, even without submodularity, the spreading optimization problem can be solved efficiently on large networks.

5. Empirical research on influential spreaders

With the rapid development of the Internet and online communities, a huge number of large-scale datasets have become available for researchers to conduct analysis on spreading processes. In the past decade, a huge number of research works have been performed on datasets from various types of online social networks, including email communication [157], online social networks—Facebook [158]–[160], microblogging services—Twitter [121], [161]–[165], blogs sharing communities—LiveJournal [166, 167], and other online communities [7], [168]–[175].

Directly examining the real diffusion data enables us to reveal the exact spreading mechanisms beneath the propagation and discover new principles dominating the diffusion process. In this section, we would like to focus on the issue of finding individual super spreaders and present our new results based on a complete network structure and the record of the diffusion instances. As we have shown, most of the previous research on individual super spreaders is based on the assumptions of specific spreading rules. However, how the information diffuses in reality is still not clear. A recent study [170] shows that the structure of online diffusion networks cannot be fully described by current theoretical models. Some empirical research has found that the predictions based on specific models are not always correct [176]–[179]. Consequently, we want to check the validity of these measures with real diffusion data.

To achieve this, we have gathered social network as well as diffusion data from a well-known online blog community—LiveJournal. This online community has been used to study spreading in previous research works [157, 166, 167]. The social network is constructed by the friend relations in LiveJournal, i.e. if user a is in user b's friend list, then there is a directed social link from b to a. In LJ, users can obtain updates of friends' posts. Therefore, the information would flow along the incoming links of a source node. With this method, we obtain a complete social network with 9 636 481 nodes and 197 368 009 links. In order to extract the diffusion instances, we collected 56 180 137 posts published by LiveJournal users and filtered 598 833 posts that contain links to other posts in LiveJournal. In this way, if user a has cited user b's posts at least once, we put a diffusion link from b to a. The resulting unweighted directed graph is called a diffusion graph, from which we can infer each node's influence. In our study, we only consider the measures of indegree, PageRank and k-shell. Notice that, according to the definition of the LJ social network, indegree describes the number of audiences of a user. So here we discuss indegree rather than outdegree. Eigenvalue centrality is similar to PageRank. And betweenness centrality is infeasible for such a large network due to its great computational complexity. All the definitions of these measures can be found in section 3.

Our first observation of this dataset is that, in the LiveJournal social network with rich topological structure, different measures can reflect very distinct information for each

J. Stat. Mech. (2013) P12002

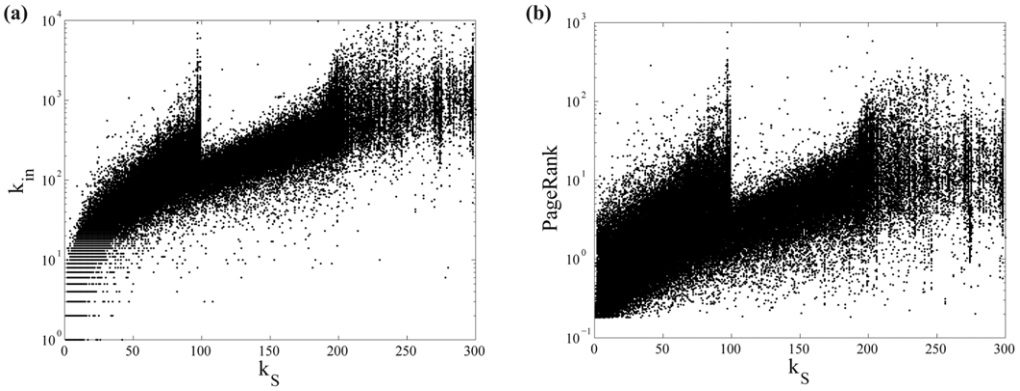

Figure 1. (a) Indegree k_{in} versus k-shell index k_S for nodes in LiveJournal. (b) Crossplot of PageRank and k_S for nodes in LiveJournal. The correlations between k_S and the other two measures are weak.

node. In figure 1(a), there exist lots of nodes with large indegree but small k_S. Meanwhile, not all the nodes with large k_S have large indegree. A similar result is also obtained for PageRank, which is presented in figure 1(b). Since the measures may convey different information for the same node, their abilities to reflect spreading power should be different as well.

Despite the large number of users in LiveJournal, there are only a small fraction of nodes participating in spreading. Precisely, only 246 423 users are involved in the information diffusion. To quantify a node's spreading ability, we infer its influence by the size of its outgoing component in the diffusion graph. Concretely, for each node in the diffusion graph, we first follow the diffusion links starting from it, finding the first layer nodes that adopt the information, and then track the links starting from these nodes and so on. This process applies recursively until no more diffusion links exist. We define the number of these reachable nodes from node i in the diffusion graph as the influence of node i, and denote it as M_i. Although the exact information diffusing from node i to these nodes may be different, node i has great potential to influence these nodes.

To compare the effect of different measures, in figure 2(a), we show the distribution of nodes for k_S, k_{in} and PageRank. To construct the data bins, we first take the logarithmic values of each measure, and then divide the range into ten intervals equally. Considering the different ranges of these measures, the intervals are normalized to $[0, 1]$, and from 0 to 1, the value increases from the minimum to maximum. We can see in the top region, k_S has far more nodes than k_{in} and PageRank. This is because while k_{in} and PageRank assign tens of thousands different values, k_S can only have several hundreds of discrete values. Figure 2(b) presents the number of users involved in spreading in each bin. The main difference between k_S and the other two measures is that the number of spreading users has a growing tendency as k_S increases. The larger number of spreading users in the top region of k_S can be explained by the larger number of users in this area, as shown in 2(a).

In order to check the distribution of spreading users according to different measures, we plot figure 2(c) to show the fraction of users participating in diffusion in each data bin for three predictors. The fraction of spreading users is defined as the ratio between the number of spreading users and the total number of users in each bin. Clearly, the

J. Stat. Mech. (2013) P12002

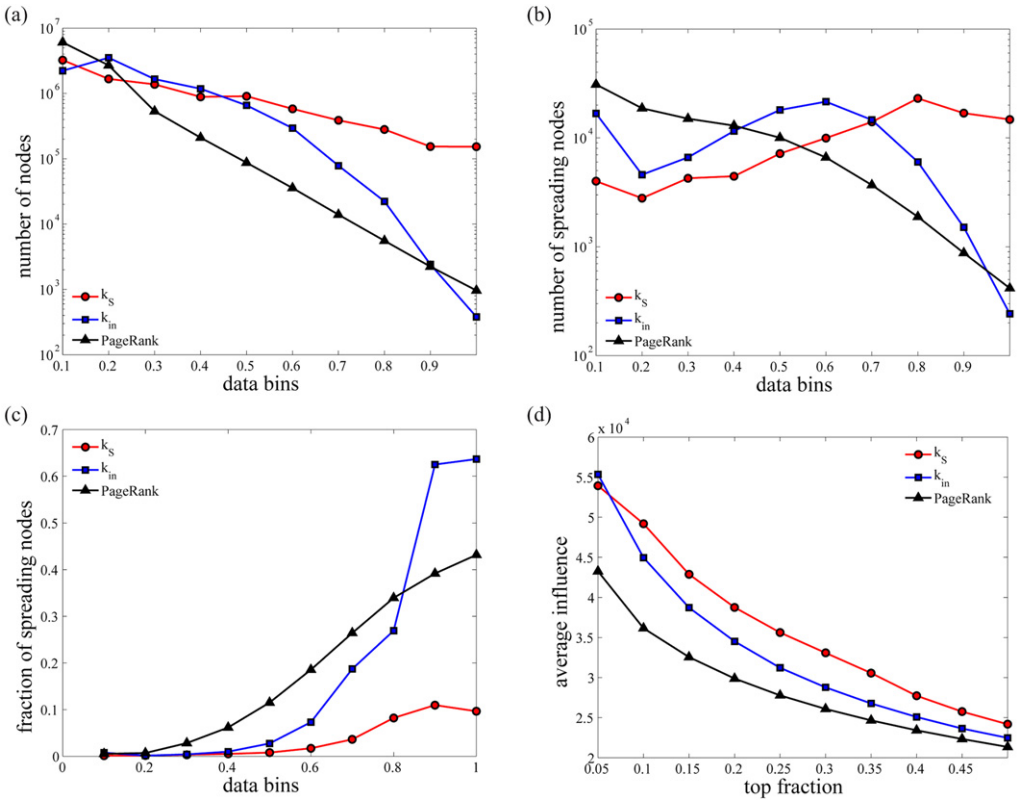

Figure 2. (a) The number of users in each data bin for k_S, $k_{\rm in}$ and PageRank respectively. The bins are created by dividing the range of each measure into ten parts equally according to the logarithmic value. From 0 to 1, the measure increases from minimum to maximum. (b) Number of nodes involved in spreading in each data bin for k_S, $k_{\rm in}$ and PageRank respectively. (c) Fraction of users in each bin that participate in information diffusion. (d) Average influence of users ranking top by different measures. We rank the spreading users by different measures, select the nodes ranking in the top f fraction, and then take the average of their influence.

fraction of all three measures increases as the value of each measure grows. This indicates that the users ranking top by these predictors have a larger probability of participating in information spreading. Particularly, in the top area, nodes with large $k_{\rm in}$ have much more chance to spread information than k_S and PageRank. The reason for that is there are very few hubs with extremely large indegree. Even though the number of spreading users is relatively small, when divided by the number of hubs, the fraction becomes larger than that of k_S and PageRank.

While indegree can find nodes participating in information diffusion with larger probability in the top region, it is desirable to check the influence of these identified users. For each predictor, we present the average influence of the nodes involved in diffusion that rank in the top f fraction in figure 2(d). The result show that the nodes identified by k_S in general have a larger influence than indegree and PageRank. This means that if a spreading process starts in the core region of the network, it will lead to larger diffusion. This result coincides with the recent report of SIR and SIS modeling on real-world networks [45].

J. Stat. Mech. (2013) P12002

Therefore, in practice identifying super individual spreaders with k_S is more reliable than indegree and PageRank.

6. Conclusion and discussion

Searching for influential spreaders in complex networks is a crucial issue for many applications. In this paper, we make a review of the most important theoretical models in describing spreading dynamics, and introduce the current methods to identify both the individual and multiple influential spreaders in various diffusion processes. Through empirical diffusion data from LiveJournal, we find that in practice different measures usually convey distinct information for nodes in social networks. Of all the users in the network, only a small fraction of users participate in spreading. Indegree can locate nodes that are involved in information diffusion with a higher probability than k-shell and PageRank. However, if we want to identify nodes with a large influence, it is preferred to use the k-shell index. Our results come from the direct analysis of empirical diffusion data, thus providing practical instructions in real-world applications.

Even though great improvements have been made in the research of finding influential spreaders, there are still many problems we need to investigate. For instance, our results of k-shell are only tested on a specific online community—LiveJournal. How the results apply in other systems needs to be further examined. For both single and multiple spreaders, most of the current algorithms require the topological structure of the underlying social network. In contrast, it is usually difficult to reconstruct the social network in practice. Consequently, some development of local algorithms would still be desirable. In future research works, these considerations would still attract attention from various domains and lead to further exploration.

Acknowledgments

We gratefully acknowledge funding by ARL under Cooperative Agreement Number W911NF-09-2-0053, NIH and NSFC (11290141, 11201018). We thank L Muchnik for useful discussions and providing the LiveJournal dataset.

References

[1] Anderson R M, May R M and Anderson B, 1992 *Infectious Diseases of Humans: Dynamics and Control* (Oxford: Oxford University Press)
[2] Hethcote H W, 2000 *SIAM Rev.* **42** 599
[3] Diekmann O and Heesterbeek J A P, 2000 *Mathematical Epidemiology of Infectious Diseases* (Chichester: Wiley)
[4] Keeling M J and Rohani P, 2008 *Modeling Infectious Diseases in Humans and Animals* (Princeton, NJ: Princeton University Press)
[5] Adar E, Zhang L, Adamic L A and Lukose R M, 2004 *Workshop on the Weblogging Ecosystem* vol 13
[6] Kumar R, Novak J, Raghavan P and Tomkins A, 2004 *Commun. ACM* **47** 35
[7] Adar E and Adamic L A, 2005 *Proc. 2005 IEEE/WIC/ACM Int. Conf. on Web Intelligence* pp 207–14
[8] Leskovec J, Backstrom L and Kleinberg J, 2009 *Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 497–506
[9] Metcalfe J S, 1987 *The Diffusion of Innovation: An Interpretive Survey* (Manchester: University of Manchester, Department of Economics)
[10] Strang D and Soule S A, 1998 *Annu. Rev. Sociol.* **24** 265

J. Stat. Mech. (2013) P12002

[11] Rogers E M, 2010 *Diffusion of Innovations* (New York: Free Press)
[12] Richardson M and Domingos P, 2002 *Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 61–70
[13] Domingos P M, 2005 *IEEE Intell. Syst.* **20** 80
[14] Watts D J, Peretti J and Frumin M, *Viral marketing for the real world*, 2007 *Harvard Business Rev. (May)* pp 22–3
[15] Leskovec J, Adamic L A and Huberman B A, 2007 *ACM Trans. Web (TWEB)* **1** 5
[16] Diani M and McAdam D, 2003 *Social Movements and Networks: Relational Approaches to Collective Action: Relational Approaches to Collective Action* (Oxford: Oxford University Press)
[17] McAdam D and Paulsen R, 1993 *Am. J. Sociol.* **99** 640
[18] Polletta F and Jasper J M, 2001 *Annu. Rev. Sociol.* **27** 283
[19] Snow D A, Zurcher L A Jr and Ekland-Olson S, 1980 *Am. Sociol. Rev.* **45** 787
[20] Albert R and Barabási A-L, 2002 *Rev. Mod. Phys.* **74** 47
[21] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U, 2006 *Phys. Rep.* **424** 175
[22] Barrat A, Barthelemy M and Vespignani A, 2008 *Dynamical Processes on Complex Networks* (Cambridge: Cambridge University Press)
[23] Albert R, Jeong H and Barabási A-L, 2000 *Nature* **406** 378
[24] Newman M E J, 2003 *SIAM Rev.* **45** 167
[25] Cohen R, Erez K, Ben-Avraham D and Havlin S, 2001 *Phys. Rev. Lett.* **86** 3682
[26] Wasserman S, 1994 *Social Network Analysis: Methods and Applications* (Cambridge: Cambridge University Rress)
[27] Freeman L C, 2004 *The Development of Social Network Analysis* (Vancouver: Empirical Press)
[28] Castellano C, Fortunato S and Loreto V, 2009 *Rev. Mod. Phys.* **81** 591
[29] Gallos L K, Havlin S, Liljeros F and Makse H A, 2012 *Phys. Rev.* X **2** 031014
[30] Yan S, Tang S, Pei S, Jiang S, Zhang X, Ding W and Zheng Z, 2013 *Physica* A **392** 3846
[31] Kinouchi O and Copelli M, 2008 *Nature Phys.* **2** 348
[32] Larremore D B, Shew W L and Restrepo J G, 2011 *Phys. Rev. Lett.* **106** 058101
[33] Pei S, Tang S, Yan S, Jiang S, Zhang X and Zheng Z, 2012 *Phys. Rev.* E **86** 021909
[34] Camacho J, Guimer à R and Amaral L A N, 2002 *Phys. Rev. Lett.* **88** 228102
[35] Dunne J A, Williams R J and Martinez N D, 2002 *Proc. Nat. Acad. Sci. USA* **99** 12917
[36] Yakovenko V M and Rosser J B Jr, 2009 *Rev. Mod. Phys.* **81** 1703
[37] Pei S, Tang S, Zhang X, Liu Z and Zheng Z, 2012 *Physica* A **391** 2023
[38] Barabási A-L and Albert R, 1999 *Science* **286** 509
[39] Watts D J and Strogatz S H, 1998 *Nature* **393** 440
[40] Strogatz S H, 2001 *Nature* **410** 268
[41] Song C, Havlin S and Makse H A, 2005 *Nature* **433** 392
[42] Gallos L K, Makse H A and Sigman M, 2012 *Proc. Nat. Acad. Sci. USA* **109** 2825
[43] Girvan M and Newman M E J, 2002 *Proc. Nat. Acad. Sci. USA* **99** 7821
[44] Galvao V, Miranda J G V, Andrade R F S, Andrade J S Jr, Gallos L K and Makse H A, 2010 *Proc. Nat. Acad. Sci. USA* **107** 5750
[45] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E and Makse H A, 2010 *Nature Phys.* **6** 888
[46] Pastor-Satorras R and Vespignani A, 2001 *Phys. Rev. Lett.* **86** 3200
[47] Boguná M and Pastor-Satorras R, 2002 *Phys. Rev.* E **66** 047104
[48] Kempe D, Kleinberg J and Tardosé É, 2003 *Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 137–46
[49] Schelling T, 1978 *Micromotives and Macrobehavior* (New York: Norton)
[50] Granovetter M, 1978 *Am. J. Sociol.* **83** 1420
[51] Granovetter M and Soong R, 1983 *J. Math. Sociol.* **9** 165
[52] Herrmann H J, 1986 *Phys. Rep.* **3** 153
[53] Boguná M, Pastor-Satorras R and Vespignani A, 2003 *Phys. Rev. Lett.* **90** 028701
[54] Bass F, 1969 *Manag. Sci.* **15** 215
[55] Murray J D, 1993 *Mathematical Biology* (Berlin: Springer)
[56] Liljeros F, Edling C R, Amaral L A N, Stanley H E and Aberg Y, 2001 *Nature* **411** 907
[57] Grassberger P, 1983 *Math. Biosci.* **63** 157
[58] Moore C and Newman M E J, 2000 *Phys. Rev.* E **61** 5678
[59] Newman M E J, 2002 *Phys. Rev.* E **66** 016128
[60] Pastor-Satorras R and Vespignani A, 2001 *Phys. Rev.* E **63** 066117
[61] Pastor-Satorras R and Vespignani A, 2002 *Phys. Rev.* E **65** 035108

J. Stat. Mech. (2013) P12002

[62] Pastor-Satorras R and Vespignani A, 2002 *Phys. Rev.* E **65** 036104
[63] Moreno Y, Pastor-Satorras R and Vespignani A, 2002 *Eur. Phys. J.* B **26** 521
[64] Dezso Z and Barabási A-L, 2002 *Phys. Rev.* E **65** 055103R
[65] Eguíluz V M and Klemm K, 2002 *Phys. Rev. Lett.* **89** 108701
[66] Serrano M Á and Boguná M, 2006 *Phys. Rev. Lett.* **97** 088701
[67] Kuperman M and Abramson G, 2001 *Phys. Rev. Lett.* **86** 2909
[68] Santos F C, Rodrigues J F and Pacheco J M, 2005 *Phys. Rev.* E **72** 056128
[69] Moreno Y, Gómez J B and Pacheco A F, 2003 *Phys. Rev.* E **68** 035103
[70] Gladwell M, 2000 *The Tipping Point: How Little Things Can Make a Big Difference* (New York: Little, Brown)
[71] Bikhchandani S, Hirshleifer D and Welch I, 1992 *J. Pol. Econ.* **100** 992
[72] Aguirre B E, Quarantelli E L and Mendoza J L, 1988 *Am. Sociol. Rev.* **53** 569
[73] Glaeser E L, Sacerdote B and Schheinkman J A, 1996 *Quart. J. Econ.* **111** 507
[74] Valente T W, 1995 *Network Models of the Diffusion of Innovations* (Cresskill, NJ: Hampton Press)
[75] Schelling T C, 1973 *J. Conflict Resolution* **17** 381
[76] Glance N S and Huberman B A, 1993 *J. Math. Sociol.* **17** 281
[77] Watts D J, 2002 *Proc. Nat. Acad. Sci. USA* **99** 5766
[78] Kleinberg J, 2007 *Algorithmic Game Theory* (*Cascading Behavior in Networks: Algorithmic and Economic Issues*) (Cambridge: Cambridge University Press) chapter 24, pp 613–32
[79] Dodds P S and Watts D J, 2004 *Phys. Rev. Lett.* **92** 218701
[80] Daley D J and Kendall D G, 1964 *Nature* **204** 1118
[81] Zanette D H, 2001 *Phys. Rev.* E **64** 050901(R)
[82] Liu Z, Lai Y C and Ye N, 2003 *Phys. Rev.* E **67** 031911
[83] Moreno Y, Nekovee M and Vespignani A, 2004 *Phys. Rev.* E **69** 055101(R)
[84] Moreno Y, Nekovee M and Pacheco A F, 2004 *Phys. Rev.* E **69** 066130
[85] Krapivsky P L and Redner S, 2003 *Phys. Rev. Lett.* **90** 238701
[86] Fontes L R, Schonmann R H and Sidoravicius V, 2002 *Commun. Math. Phys.* **228** 495
[87] Castellano C, Vilone D and Vespignani A, 2003 *Europhys. Lett.* **63** 153
[88] Sood V and Redner S, 2005 *Phys. Rev. Lett.* **94** 178701
[89] Morris S, 2000 *Rev. Econom. Stud.* **67** 57
[90] Blume L, 1993 *Games Econ. Behav.* **5** 387
[91] Ellison G, 1993 *Econometrica* **61** 1047
[92] Young H P, 1998 *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton, NJ: Princeton University Press)
[93] Freeman L C, 1979 *Soc. Netw.* **1** 215
[94] Friedkin N E, 1991 *Am. J. Sociol.* **96** 1478
[95] Sabidussi G, 1966 *Psychometrika* **31** 581
[96] Dangalchev C, 2006 *Physica* A **365** 556
[97] Boguná M, Krioukov D and Claffy K C, 2008 *Nature Phys.* **5** 74
[98] Brandes U, 2008 *Soc. Netw.* **30** 136
[99] Estrada E, 2007 *Ecol. Complexity* **4** 48
[100] Martín González A M, Dalsgaard B and Olesen J M, 2010 *Ecol. Complexity* **7** 36
[101] Zuo X N, Ehmke R, Mennes M, Imperati D, Castellanos F X, Sporns O and Milham M P, 2012 *Cerebral Cortex* **22** 1862
[102] Wang J, Mo H, Wang F and Jin F, 2011 *J. Transp. Geogr.* **19** 712
[103] Abbasi A, Hossain L and Leydesdorff L, 2012 *J. Informetrics* **6** 403
[104] Floyd R W, 1962 *Commun. ACM* **5** 345
[105] Johnson D B, 1977 *J. ACM (JACM)* **24** 1
[106] Brandes U, 2001 *J. Math. Sociol.* **25** 163
[107] Bonacich P, 1972 *J. Math. Sociol.* **2** 113
[108] Newman M E J, 2008 *The New Palgrave Encyclopedia of Economics* **2** 1
[109] Brin S and Page L, 1998 *Computer Networks and ISDN Systems* **30** 107
[110] Chen P, Xie H, Maslov S and Redner S, 2007 *J. Informetrics* **1** 8
[111] Walker D, Xie H, Yan K K and Maslov S, 2007 *J. Stat. Mech.* P06010
[112] Schmidt B M and Chingos M M, 2007 *PS-WASHINGTON* **40** 523
[113] Chen J, Aronow B J and Jegga A G, 2009 *BMC Bioinform.* **10** 73
[114] Iván G and Grolmusz V, 2011 *Bioinformatics* **27** 405
[115] Jiang B, Zhao S and Yin J, 2008 *J. Stat. Mech.* P07008
[116] Allesina S and Pascual M, 2009 *PLoS Comput. Biol.* **5** e1000494

J. Stat. Mech. (2013) P12002

[117] Navigli R and Lapata M, 2010 *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 678
[118] Arasu A, Novak J, Tomkins A and Tomlin J, 2002 *Proc. 11th Int. World Wide Web Conf., Poster Track* pp 107–17
[119] Franceschet M, 2011 *Commun. ACM* **54** 92
[120] Lü L, Zhang Y C, Yeung C H and Zhou T, 2011 *PLoS One* **6** e21202
[121] Weng J, Lim E P, Jiang J and He Q, 2010 *Proc. 3rd ACM Int. Conf. on Web Search and Data Mining* pp 261–70
[122] Haveliwala T H, 2002 *Proc. 11th Int. Conf. on World Wide Web* pp 517–26
[123] Bollobás B, *Graph Theory and Combinatorics*, 1984 *Proc. Cambridge Combinatorial Conference in Honor of P. Erdós* vol 35 (New York: Academic)
[124] Seidman S B, 1983 *Soc. Netw.* **5** 269
[125] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y and Shir E, 2007 *Proc. Nat. Acad. Sci. USA* **104** 11150
[126] Batagelj V and Zaversnik M, 2003 arXiv:cs/0310049
[127] Christakis N A and Fowler J H, 2010 *PLoS ONE* **5** e12948
[128] Castellano C and Pastor-Satorras R, 2012 *Sci. Rep.* **2** 371
[129] Borge-Holthoefer J and Moreno Y, 2012 *Phys. Rev.* E **85** 026116
[130] Zeng A and Zhang C J, 2013 *Phys. Lett.* A **377** 1031
[131] Basaras P, Katsaros D and Tassiulas L, 2013 *Computer* **46** 24
[132] Klemm K, Serrano M Á, Eguíluz V M and Miguel M S, 2012 *Sci. Rep.* **2** 292
[133] Travençolo B A N and da F Costa L, 2008 *Phys. Lett.* A **373** 89
[134] Viana M P, Batista J L B and Costa L F, 2012 *Phys. Rev.* E **85** 036105
[135] Brush S G, 1967 *Rev. Mod. Phys.* **39** 883
[136] Dorogovtsev S N, Goltsev A V and Mendes J F F, 2002 *Phys. Rev.* E **66** 016104
[137] Holley R A and Liggett T M, 1975 *Ann. Probab.* **3** 643
[138] Acebrón J A, Bonilla L L, Vicente C J P, Ritort F and Spigler R, 2005 *Rev. Mod. Phys.* **77** 137
[139] Bauer F and Lizier J T, 2012 *Europhys. Lett.* **99** 68007
[140] Lawyer G, 2012 arXiv:1209.6600
[141] Domingos P and Richardson M, 2001 *Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 57–66
[142] Kempe D, Kleinberg J and Tardos É, 2005 *Automata, Languages and Programming* (Berlin: Springer) pp 1127–38
[143] Nemhauser G L, Wolsey L A and Fisher M L, 1978 *Math. Program.* **14** 265
[144] Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J and Glance N, 2007 *Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 420–9
[145] Chen W, Wang Y and Yang S, 2009 *Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 199–208
[146] Goyal A, Lu W and Lakshmanan L V S, 2011 *Proc. 20th Int. Conf. Companion on World Wide Web* pp 47–8
[147] Goyal A, Lu W and Lakshmanan L V S, 2011 *ICDM: 2011 IEEE 11th Int. Conf. on Data Mining* pp 211–20
[148] Ben-Zwi O, Hermelin D, Lokshtanov D and Newman I, 2009 *Proc. 10th ACM Conf. on Electronic Commerce* pp 355–62
[149] Wang Y, Cong G, Song G and Xie K, 2010 *Proc. 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 1039–48
[150] Kimura M, Saito K and Nakano R, 2007 *AAAI* **7** 1371
[151] Kimura M, Saito K, Nakano R and Motoda H, 2009 *Social Computing and Behavioral Modeling* (New York: Springer) pp 1–8
[152] Chen W, Wang C and Wang Y, *Proc. 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 1029–38
[153] Altarelli F, Braunstein A, Dall'Asta L and Zecchina R, 2013 *J. Stat. Mech.* P09011
[154] Mézard M and Montanari A, 2009 *Information, Physics and Computation* (Oxford: Oxford University Press)
[155] Altarelli F, Braunstein A, Ramezanpour A and Zecchina R, 2011 *J. Stat. Mech.* P11009
[156] Altarelli F, Braunstein A, Ramezanpour A and Zecchina R, 2011 *Phys. Rev. Lett.* **106** 190601
[157] Liben-Nowell D and Kleinberg J, 2008 *Proc. Nat. Acad. Sci. USA* **105** 4633
[158] Viswanath B, Mislove A, Cha M and Gummadi K P, 2009 *Proc. 2nd ACM Workshop on Online Social Networks* pp 37–42
[159] Ahn Y Y, Han S, Kwak H, Moon S and Jeong H, 2007 *Proc. 16th Int. Conf. on World Wide Web* pp 835–44

J. Stat. Mech. (2013) P12002

[160] Mislove A, Marcon M, Gummadi K P, Druschel P and Bhattacharjee B, 2007 *Proc. 7th ACM SIGCOMM Conf. on Internet Measurement* pp 29–42
[161] Krishnamurthy B, Gill P and Arlitt M, 2008 *Proc. 1st Workshop on Online Social Networks* pp 19–24
[162] Cha M, Haddadi H, Benevenuto F and Gummadi P K, 2010 *ICWSM* 10 pp 10–7
[163] Bakshy E, Hofman J M, Mason W A and Watts D J, 2011 *Proc. 4th ACM Int. Conf. on Web Search and Data Mining* pp 65–74
[164] Kwak H, Lee C, Park H and Moon S, 2010 *Proc. 19th Int. Conf. on World Wide Web* pp 591–600
[165] González-Bailón S, Borge-Holthoefer J, Rivero A and Moreno Y, 2011 *Sci. Rep.* **1** 197
[166] Backstrom L, Huttenlocher D, Kleinberg J and Lan X, 2006 *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 44–54
[167] Liben-Nowell D, Novak J, Kumar R, Raghavan P and Tomkins A, 2005 *Proc. Nat. Acad. Sci. USA* **102** 11623
[168] Gruhl D, Guha R, Liben-Nowell D and Tomkins A, 2004 *Proc. 13th Int. Conf. on World Wide Web* pp 491–501
[169] Centola D, 2010 *Science* **329** 1194
[170] Goel S, Watts D J and Goldstein D G, 2012 *Proc. 13th ACM Conf. on Electronic Commerce* pp 623–38
[171] Aral S and Walker D, 2012 *Science* **337** 337
[172] Muchnik L, Pei S, Parra L C, Reis S D S, Andrade J S Jr, Havlin S and Makse H A, 2013 *Sci. Rep.* **3** 1783
[173] Sun E, Rosenn I, Marlow C and Lento T M, 2009 *ICWSM* pp 146–53
[174] Bakshy E, Karrer B and Adamic L A, 2009 *Proc. 10th ACM Conf. on Electronic commerce* pp 325–34
[175] Aral S, Muchnik L and Sundararajan A, 2009 *Proc. Nat. Acad. Sci. USA* **106** 21544
[176] Centola D and Macy M, 2007 *Am. J. Sociol.* **113** 702
[177] Goldenberg J, Libai B and Muller E, 2001 *Marketing Lett.* **12** 211
[178] Jackson M O and Lopez-Pintado D, 2011 arXiv:1111.0073
[179] Aral S, Muchnik L and Sundararajan A, 2013 *Netw. Sci. Publishing* **1** 125

J. Stat. Mech. (2013) P12002