Finding influential spreaders from human activity beyond network location

Byungjoon Min[1], Fredrik Liljeros[2,3], Hernán A. Makse[1,*]
1 Levich Institute and Physics Department, City College of New York, New York, NY, US
2 Department of Sociology, Stockholm University, Stockholm, Sweden
3 Institute for Futures Study, Stockholm, Sweden
*** E-mail: hmakse@lev.ccny.cuny.edu**

Abstract

Most centralities proposed for identifying influential spreaders on social networks to either spread a message or to stop an epidemic require the full topological information of the network on which spreading occurs. In practice, however, collecting all connections between agents in social networks can be hardly achieved. As a result, such metrics could be difficult to apply to real social networks. Consequently, a new approach for identifying influential people without the explicit network information is demanded in order to provide an efficient immunization or spreading strategy, in a practical sense. In this study, we seek a possible way for finding influential spreaders by using the social mechanisms of how social connections are formed in real networks. We find that a reliable immunization scheme can be achieved by asking people how they interacts with each other. From these surveys we find that the probabilistic tendency to connect to a hub has the strongest predictive power for influential spreaders among tested social mechanisms. Our observation also suggests that people who connect different communities is more likely to be an influential spreader when network has a strong modular structure. Our finding implies that not only the effect of network location but also the behavior of individuals is important to design optimal immunization or spreading schemes.

Author Summary

Design of efficient immunization strategies is important to lower the threshold of immunization against an infectious disease. For this, the most influential spreaders should be immunized with the highest priority in order to reduce epidemics. Finding influential spreaders is also important for viral marketing campaign to search efficient strategy of spreading via targeting superspreaders. Here, we propose an approach for identifying influential people based on social mechanisms of individuals, that is how connections are formed in real-world social networks. Since these properties can be obtained by using a survey conducted for a population, it is easy to apply for real spreading. We find that the social mechanism of connecting a hub can be reliable predictors for influential spreaders in future epidemics or viral marketing campaigns. From the analysis of the microscopic link formation, we find the effect of social mechanisms on spreading processes.

Introduction

Identifying influential spreaders on social networks is crucial for its practical application in real-world epidemic and information spreading [1–5]. For instance, superspreaders need to be immunized with the highest priority in order to prevent the pandemic of an infectious disease [6–10]. They are also important for spreading of information in viral marketing [4, 5, 11, 12]. To this end, several predictors for influential spreaders based on the topological property of complex networks, including high degree [6, 13] k-core [8, 14, 15], betweenness centrality [16], PageRank [17], and many others [18] were tested for identifying influential spreaders [8–10].

Most studies, however, have overlooked how to apply to the real-world social systems which is a serious problem in a practical sense. Most proposed centralities except the degree, which is a local centrality, require the information of the whole network structure. But collecting this information is nearly impracticable in real social systems. Specifically, gathering information of relationships among individuals is inevitably incomplete and erroneous [19], since it cannot but be conducted for a partial sample of a whole population. Thus, searching for the influential spreaders with these centralities may not be plausible for real-world spreading phenomena. On the other hand, if whole connections in a network are accessible, direct measuring for the influence of a single node is possible by using model simulation on the network, which obviates the need for predicting influential spreaders. Consequently, in reality most predictors proposed for an influential spreader are either inapplicable or unnecessary.

Thus more realistic approaches based on the characteristics of people such as their behaviors are demanded for predicting influential people without the explicit information of network structure. The benefit of this method is an easy applicability for any kinds of social networks since one can obtain the probabilistic actions of agents by using a survey conducted from a population. Through a survey, we can estimate the probabilistic tendency of how connections are established for each individual, for instance, how probable is to make a new friend by introduction from another friend or the frequency to make new friends from different groups. We find that these human actions have a large influence on the subsequent spreading of information and therefore can be a reliable predictor of the node's importance in a future epidemic or in a viral marketing campaign via targeting people identified by their probabilistic actions. In addition, such ranking obtained from surveys can also apply to the situations when the information for only some people is accessible.

The social mechanisms of link formation driving evolution of networks have been studied for a long time in order to explain and predict complex phenomena in society. A number of social mechanisms for connection establishments have been proposed in sociology [20, 21]. Thanks to the detailed records in online social networks that captures the action of every individual, it is now possible to quantify the frequency of occurrence of different types of mechanisms by directly observing social interactions [22]. Thus, recently, the frequencies of the social mechanisms for each person in a social network have been revealed from the full log of the activity in online social networks [22].

In this paper, we propose an approach to identify influential spreaders based on surveys on human behavior and social mechanisms that can be given to a population without the explicit information of networks. We decode the relation between people's characteristics that can be obtained by a survey and their influence in spreading using the real-world datasets that contain the full information of network evolution. Through the analysis of large-scale evolving networks, we identify the effect of the microscopic link formation on macroscopic consequences in spreading. We find that the interaction to connecting a hub can facilitate epidemic spreading and thus can be a reliable predictor of people's importance in future epidemics or viral marketing campaigns. We also find that people with high frequency to connect different communities are more

likely to be an influential spreader for the case when a network is composed of strongly connected modules. This research represents much to practical implication, since our finding can be adopted in reality requiring only the tendency of individuals' behaviors. Furthermore, our results provide a guideline for behavior to the public, about how to behave at the beginning stage of epidemic.

Materials and Methods

Social mechanism

In this paper, the social mechanisms are referred to as the probabilistic tendency of each kind of interaction among people in a given social network. The social mechanisms do not directly mean the motivation behind the link creation because several different mechanisms may result in the same type of link formation and link formation may not be motivated by only the structure [22]. In addition, these mechanisms are not complementary one another, because a link can be established by multiple different mechanisms. For instance, a newly created link can appear following balance and exchange interactions at the same time.

We use four classes of social mechanisms underlying the link creation on a network based on the multitheoretical multilevel formalism [20] proposed in sociology: (1) Exchange interaction corresponds to a newly form reciprocal link meaning that a new link is established in the opposite direction of an existing link. (2) Balance interaction corresponds to a newly form tie that closes a triangle by a directed edge. (3) Collective action (or preferential attachment [23]) corresponds to a link that connects with well-connected people. To be specific, in this study, we measure the extent of the collective action of each link as a continuous value using the cumulative probability $F(k_i)$ of the excess degree distribution for a newly connected neighbor i. Here, $F(k) = \sum_{k_j<k} k_j q(k_j)/\langle k\rangle$ where $q(k)$ is the degree distribution of a network and $\langle k\rangle$ represents the average degree of a network. (4) Structural hole interaction considers a newly created link that connects two different modules (communities). Community structure is identified by the local version of link community detection method [24] when a new link is established [see detailed in Text S2].

These social mechanisms are assigned on an evolving network at the moment when the link is newly added following the analysis developed in [22]. While constructing the evolving network by adding the new connection in sequential order, we characterize each connection to the corresponding social mechanisms based on a network configuration at the given moment. After all links are formed, the frequencies of social mechanisms of the origin node, i, a_i^{exc}, a_i^{bal}, a_i^{ca}, and a_i^{sh}, where i is node index, are defined as the number of neighbors that were connected by the corresponding mechanism, respectively, exchange, balance, collective action, and structural hole (the sum of the extent for the collective action of all connected nodes) normalized by the total number of neighbors. To be specific, the frequency a_i^α of social mechanism α for node i is defined as $a_i^\alpha = \frac{n_i^\alpha}{k_i^{out}}$, where n_i^α is the number of links formed corresponding to α social mechanism and k_i^{out} is the outdegree (the total number of new connections). Therefore, each variable ranges from zero to unity, and as a_i increases, the corresponding social interaction is more frequent.

We stress here that the extent of social mechanisms of link creation for each individual can be estimated in a real setting by the surveys given to the population. For instance, one first could ask people to list their contacts and then as a second stage ask questions about each contact [25]. For example, we could ask questions like, (exchange) did the person contact you first?, (balance) did the person have common friends with you when you contacted him/her?, (collective action) did the person have a lot of

Table 1. Properties of real-world networks used in this study.

Network	Name	Number of nodes	$\langle k \rangle$	Modularity [29]
QX.com favorite	QXF	80,407	13.07	0.4060
QX.com guestbook	QXG	59,854	7.10	0.3893
POK.com	POK	29,242	5.95	0.3992
Livejournal.com	LJ	315,936	3.56	0.6578
Prostitution	PRO	16,729	4.67	0.6294

$\langle k \rangle$ is the average degree of the network. We use the fast-greedy community detection algorithm [29] for measuring modularity.

contacts when you contacted him/her?, (structural hole) did person belong to another group than you when you contacted him/her? Therefore, an estimate of a_i for each individual can be obtained from the surveys conducted for the population. On the contrary, most centralities including k-shell index [8], betweenness centrality [16], and PageRank [17] cannot be obtained by this way since they require global network information.

Data sets

We examine two social networks of Internet dating services in Sweden [22, 27] and the forum of internet-mediated prostitution in Brazil [28]. These social networks represent potential pathways for epidemic spreading including sexually transmitted diseases. We use the data of the largest site *qx.se* for Nordic homosexual, bisexual, and transgender people in 2006 (QX). Actions of every individual in the community, including adding an individual to the favorite list and guestbook signing, were recorded for two months starting from Nov. 2005. We use adding favorite lists (QXF) and signing guestbook lists (QXG) among many activities. We also analyze pussokram *pussokram.com* dataset (POK) [27], which was a Swedish online dating site for friendship including flirting and non-romantic relations. The data contains a full log for 512 days starting from the day when the community was created in 2011. The POK network that we use in this study consists of message senders, receiver, and the timing of interactions in the community. Internet-mediated prostitution data (PRO) [28] comes from Brazilian online forum where sex-buyers evaluate prostitutes. We construct the PRO network by connecting sex-sellers with buyers. Since the PRO network is an undirected and bipartite graph, the exchange and balance interactions are not defined. In order to investigate the problem of identifying influential spreaders of information, we study the citation network in the posts of an online network service, *livejournal.com* (LJ), for information spreading on social networks [10]. One should note that the QX has already a large part of network (85 and 87 % for the QXF and QXG, respectively) whereas the others starts at time $t = 0$. Table 1 gives the basic information of the datasets.

We can reconstruct the evolving connection of networks, following the precise timing when a tie has been established, in contrast to the observation of static snapshots of networks. In our datasets, we can observe every evolution of social networks with the time stamp of link creations. We stress here that the precise information of temporal evolution is essential to identify the social mechanisms for each link. The social mechanisms should be defined at the moment when a new link established [22]. Accumulated static networks do not keep the order of time that links established and thus are misleading about the social interactions. In this regard, our datasets containing the full log of network evolution allow us to define social mechanisms properly.

Influential spreader

In order to assess the influence of people for epidemic spreading, we use the epidemic size M_i originating from a seed i in the susceptible-infected-recovered (SIR) model on the finally accumulated network [8]. The SIR model has been used to describe infectious disease for a long time [30]. At the same time, the SIR model is a plausible model of information spreading [8]. In the SIR model, each node can be in one of three states, susceptible, infected, or recovered (or removed). Initially, all nodes are in the susceptible state except for a single node in the infected state. At each time, the infected node spread a disease/information to a susceptible neighbor with infection probability β. At the steady state, we measure M_i as the fraction of finally infected nodes. We define a node with high M_i as highly influential.

We choose the infection probability β to be a value covering a small part of a network, $\beta \gtrsim \beta_c$ where β_c is the epidemic threshold for percolation [30, 31]. When $\beta \gg \beta_c$, all seed produces similar epidemic size because spreading can cover almost all network regardless of where it originated from [32].

Results

Predictor for influential spreaders based on human activity.

We recreate the entire network by adding all links in the order of time that they were established. In order to assess systematically the relation of the epidemic influence M_i with the social mechanisms as well as topological metrics, we use multilinear regression analysis [33] with the following model (Tables S1-S5):

$$M_i = c_0 + c_1 a_i^{\mathrm{exc}} + c_2 a_i^{\mathrm{bal}} + c_3 a_i^{\mathrm{ca}} + c_4 a_i^{\mathrm{sh}} + c_5 k_i + c_6 k_i^{\mathrm{sh}} + c_7 k_i^{\mathrm{2sum}} + c_8 k_i^{\mathrm{sum}} + \epsilon. \quad (1)$$

Here, k_i is the degree of node i, k_i^{sh} is the k-shell index [8] (Text S1). k_i^{sum} is the sum of degree of the nearest neighbors $k_i^{\mathrm{sum}} = \sum_{j \in V(i)} k_j$ where $V(i)$ is the set of node i's neighbors [10], k_i^{2sum} is the sum of degrees of the next-nearest neighbors, $k_i^{\mathrm{2sum}} = \sum_{j \in V_2(i)} k_j$ where $V_2(i)$ is the set of neighbors of node i's neighbors [10], and ϵ is the error term. We introduce the topological metrics, since we are interested in how much information we captured using the social mechanisms tendencies $\{a_i^{\mathrm{exc}}, a_i^{\mathrm{bal}}, a_i^{\mathrm{ca}}, a_i^{\mathrm{sh}}\}$ in comparison with the more common topological measurements, $\{k_i, k_i^{\mathrm{sh}}, k_i^{\mathrm{sum}}, k_i^{\mathrm{2sum}}\}$. In order to avoid biased observation due to the large fluctuation in the small degree region, we exclude the data of people with degree less than three from our analysis.

The k-shell index and its local proxy k^{sum} and k^{2sum} have been regarded as an efficient topological predictor for influential spreaders [8, 10]. In agreement with these previous studies, we find that k^{sh} can capture most of the fluctuation in the epidemic size for the datasets. To quantify the effect of each variable, we measure the difference $\Delta R^2(x)$ of the coefficient of determination when a variable x is excluded. In Fig. 1, the difference $\Delta R^2(k^{\mathrm{ks}})$ of the coefficient of determination is the largest when k^{ks} is excluded from Eq. (1), which confirms the importance of k^{ks}. In addition, more than 82.3 % of the fluctuations can be explained by solely the k-shell index for the QXF network (Table S1). For the QXG, POK, LJ, POK networks, we also find the similar trend as the QXF (Tables S2-S5). However, being a global quantity, the k-shell index can be difficult to obtain as discussed above. Therefore, k^{sh} has the limitation to apply for real social systems despite its strong correlation with the spreading influence. k^{sum} or k^{2sum} also captures a huge part of the variance in the data. While these are a local measurement, they still can be difficult to obtain because they require the exact number of friends of friends at the time when epidemic occurs [10].

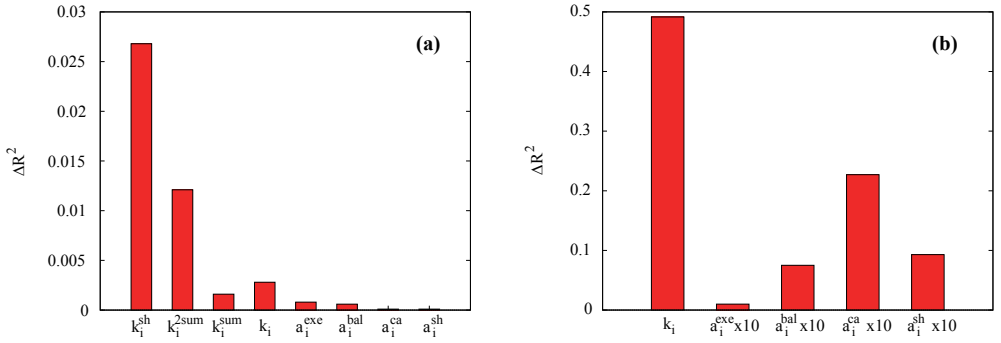

Figure 1. The difference $\Delta R^2(x)$ of the coefficient of determination when a variable x is excluded in regression analysis of (a) Eq. (1) and (b) Eq. (2) in QXF network. (a) k-shell index shows the largest drop of R^2, showing the strongest predictive power for influential spreaders. However, k-shell index is difficult to obtain since it requires the full topological information of the network. Although the degree and social mechanisms a^α show smaller predictive power than k-shell index, they can be easily obtained from surveys and have much implication in a real setting. (b) The degree shows the largest difference among the degree and social mechanisms that can be obtained from surveys. Next, among the social mechanisms, collective action shows the largest drop of R^2. Thus, collective action is a more reliable predictor than the others from the human behavioral point of view.

The degree k is not behavioral but in contrary to $k^{\rm sh}$, $k^{\rm 2sum}$, and $k^{\rm sum}$, the degree k can be estimated by a survey to individuals by a simple question: how many friends do you have? Therefore, even if we cannot conceive the structure of network, for many cases, we can access the information of the degree together with the other social mechanisms, a_i^α. Next we are interested in the case where the topological location such as k-shell cannot be obtained for the reasons explained above. Therefore, we regress the data of M_i with the variables which can be easily obtained by surveys using the following model, where $k^{\rm sh}$, $k^{\rm sum}$, and $k^{\rm 2sum}$ are excluded:

$$M_i = c_0 + c_1 a_i^{\rm exc} + c_2 a_i^{\rm bal} + c_3 a_i^{\rm ca} + c_4 a_i^{\rm sh} + c_5 k_i + \epsilon. \qquad (2)$$

When we consider Eq. (2), we can explain 63 % of the variance for the QXF network (Table S6), demonstrating that with only surveys we can capture extremely high amount of the variance. The all variables in Eq. (2) can be easily obtained from surveys, suggesting that we can rely on surveys for optimally immunization or viral marketing.

Using Eq. (2), we find that the degree is the most reliable predictor for the influential spreaders among the degree and a_i^α. When the degree is excluded from Eq. (2), the coefficient of determination R^2 drops 0.49 from 0.62, showing the largest difference (Fig. 1b). Since the degree represents the number of the transmission channels for a seed, the degree can play an important role in epidemic spreading on networks especially at the beginning stage of outbreak [8, 34]. When compared with the topological location of the people given by k-shell, we find that the degree alone can explain 58 % of the variance, which compared to the value of k-shell ($R^2 = 0.82$), indicating that the degree is a worse predictor than k-shell in agreement with [8]. In a real setting, however, the local degree can have more implication than k-shell because it can be easily obtained from surveys.

Next, we are interested in what social mechanisms a_i^α are more important for spreading besides the local degree. This is not only important for optimal immunization and information spreading but also for education of the population to avoid certain

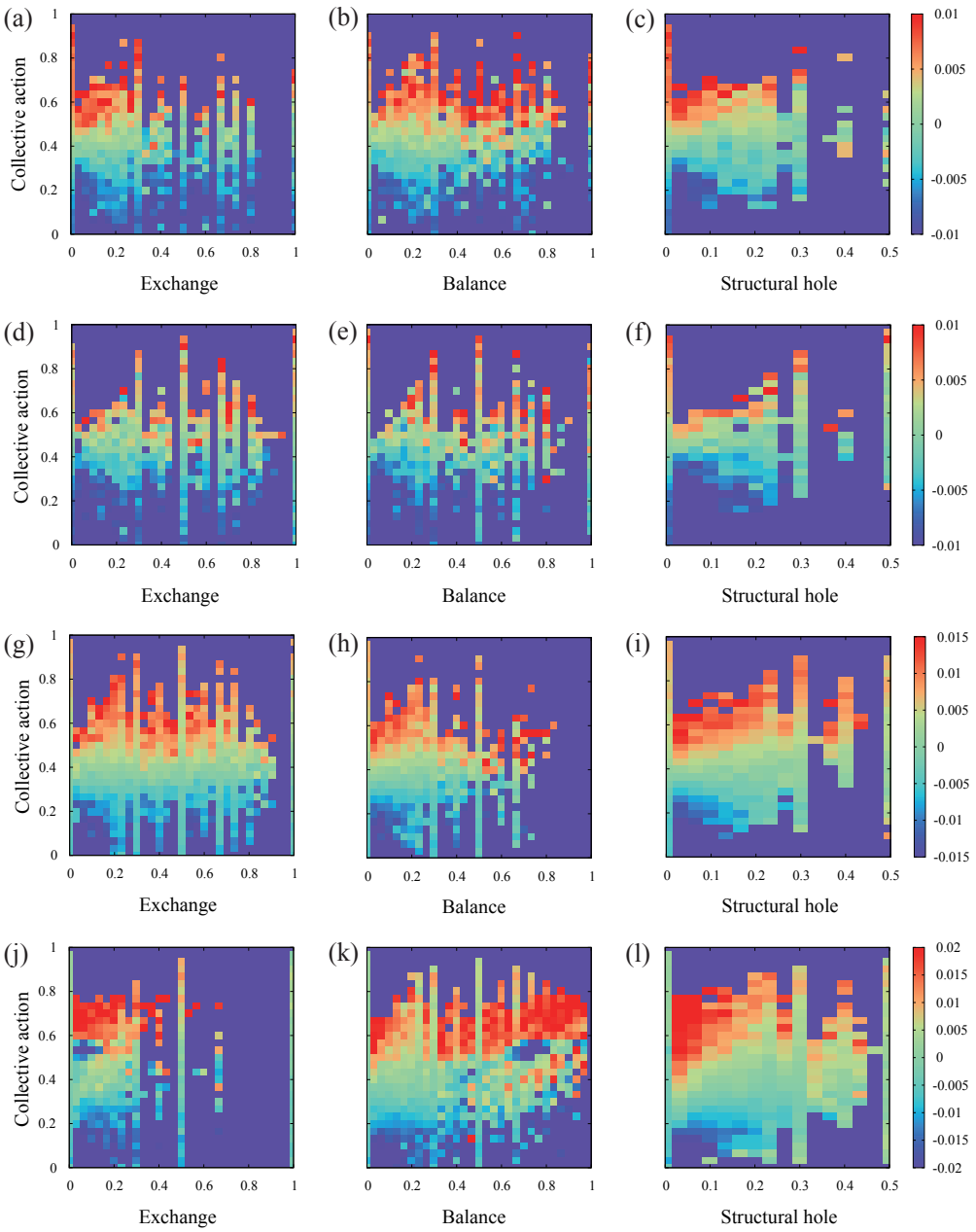

Figure 2. Collective action predicts influential spreaders more reliably than other social mechanisms. When spreading originates in people with $(a^{\alpha}, a^{\mathrm{ca}})$, the relative epidemic size $M(a^{\alpha}, a^{\mathrm{ca}})$ for the QXF with (a) a^{exc}, (b) a^{bal}, and (c) a^{sh}, (d-f) QXG, (g-i) POK, and (j-l) LJ networks. Collective action a^{ca} predicts the epidemic influence more reliably than the other social interactions when we compare for people with the same degree.

behaviors that could spread diseases to huge population. In order to examine the effect of the social mechanisms clearly, we study the deviation of the epidemic size ΔM_i from the average epidemic size for people with the same degree by following

$$\Delta M_i = M_i - \frac{\sum_j \delta_{k_i,k_j} M_j}{\sum_j \delta_{k_i,k_j}}, \tag{3}$$

where $\delta_{i,j}$ represents the Kronecker delta such that the function is 1 if the variables are equal and 0 otherwise. ΔM_i quantifies the impact of the social mechanisms after removing the effect induced by the degree, thus, more clearly identify the important social mechanism for spreading for people with the same degree.

To compare the influence of each social mechanisms in the spreading process, we study the average size ΔM infected in an epidemic originating at people i with a given $(a_i^{\rm exc}, a_i^{\rm bal}, a_i^{\rm ca}, a_i^{\rm sh})$. The average infected population over all the origins with the same pair of (a^α, a^β) is

$$\Delta M = \sum_{i \in W(a^\alpha, a^\beta)} \frac{\Delta M_i}{N(a^\alpha, a^\beta)}, \tag{4}$$

where $W(a^\alpha, a^\beta)$ is the union of all nodes with (a^α, a^β) and $N(a^\alpha, a^\beta)$ is the number of nodes with (a^α, a^β). In Fig. 2, we find that ΔM increases with increasing $a^{\rm ca}$ regardless with the other social mechanisms for all tested networks. This clear pattern suggests that $a^{\rm ca}$ predicts the epidemic influence more reliably than the other social interactions when we compare for people with the same degree.

The regression analysis of Eq. (2) also supports the importance of the collective action. When we remove $a_i^{\rm ca}$ from Eq. (2), the difference ΔR^2 of the coefficient of determination is the largest, which confirms the importance of collective action. Since people with high collective action are more likely to have many next nearest neighbors. they have high chance to develop larger epidemic outbreaks. On the contrary, people with less collective action, is likely to be located at the periphery of a network leading to a small impact in the spreading. Thus, the collective action is a reliable predictor from the human behavioral point of view when we factor out the popularity.

Strength of weak ties and community structure.

So far, we search the most influential spreaders based on social mechanisms, a_i^α which can be obtained by surveys. In sociology, a long-standing hypothesis for influential spreaders is the strength of weak ties [35]. According to the hypothesis, weak ties which bridge between two densely connected modules formed by strong ties play an important role especially in the job changing in labor market [35, 36], mobile communication networks [37], as well as brain [38]. While this hypothesis may seem counter-intuitive, for a perspective of information spreading, the weak ties is more likely to be a source of fresh information, so weak ties can have a stronger effect than strong ties.

In this section, we test the weak tie hypothesis by observing the evolution of link formation in a large scale real-world network. We define weak connection as a link bridging two different communities at the time when a new link is formed, called structural hole. If weak ties play an important role in spreading processes as the hypothesis of weak ties, people with high probability of structural hole interactions is more likely to have influence in spreading. In order to test the effect of weak ties (structural hole), we regress the data of M_i with the variables of social mechanisms a_i^α using the following model, where the network properties $k^{\rm sh}$, $k^{\rm sum}$, $k^{\rm 2sum}$, and k are excluded:

$$M_i = c_0 + c_1 a_i^{\rm exc} + c_2 a_i^{\rm bal} + c_3 a_i^{\rm ca} + c_4 a_i^{\rm sh} + \epsilon. \tag{5}$$

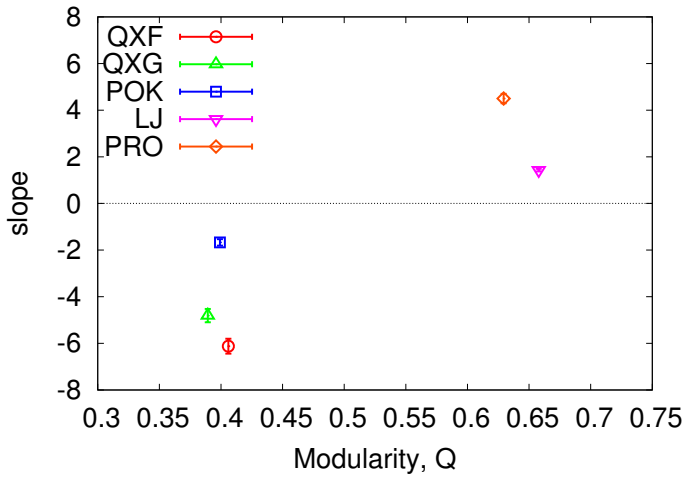

Figure 3. The effect of weak ties on spreading for different networks with diverse modularity. The panel shows the slope of the frequency of structural hole with respect to epidemic influence M_i in regression analysis as a function of modularity of a underlying network. For networks with highly modular structure such as LJ and PRO, the frequency of structural hole is positively correlated with M_i.

The degree k is also excluded in order to focus on the effect of behavioral factors on spreading.

From the regression analysis, we confirm that people with high frequency of structural hole interaction is more likely to be an influential spreaders on LJ and PRO networks as the weak tie hypothesis. In LJ and PRO networks, the frequency of structural hole a^{sh} is positively related with the spreading influence M_i with extremely small p-value (Fig. 3 and Tables S9 and S10). However, this pattern does not hold for all social networks that we tested. For QXF, QXG, and POK networks, a_i^{sh} is negatively correlated with M_i in contrary to the weak tie hypothesis (Fig. 3 and Tables S6-S8). This result suggests that the weak tie hypothesis may not be generically valid for all social networks.

The validity of the weak tie hypothesis can rely on the underlying network where spreading occurs. People with high frequency of structural hole interactions potentially spreads different communities all together. Therefore, if an underlying network of spreading has clear module structure, the effect of weak ties is significant [39]. However, when community structure is less clear the role of weak ties in spreading can be weakened. In order to check this prediction, we compare the modularity of networks [40] and the effect of weak ties (Fig. 3). When a network has strong community structure such as LJ and PRO whose modularity is 0.658 and 0.629, respectively, the frequency of structural hole is positively correlated with M_i. Therefore, the structural hole mechanisms can enhance the epidemic influence for networks with strong modular structure as the weak tie hypothesis. However, the weak tie hypothesis is not valid for networks with less clear module structure. For instance, the QXF, QXG, and POK networks showing less modularity around 0.4, a_i^{sh} play a minor role in spreading and negatively correlated with M_i (Fig. 3 and Tables S6-S8). If the modular structure is not significant, the weak ties are not clearly defined, leading to decrease of the effect of weak ties. Thus, the weak tie hypothesis is expected to be valid for strong module structure not universally for all social networks. In conclusion, people who connect different communities can be suspected as an influential people when an underlying

network is composed of strong modular structure,

Discussion

So far, most studies of spreading on complex networks have assumed that a network structure is known. This means that full information on any people on who is connected with whom is required, which may not be obtained in real settings. In agreement with the previous studies, we find that when the information of global structure of social networks is available, it is beneficial for identifying influential spreaders in an epidemic model capturing up to 90 % of the variance with simple variables with the k-shell [8,10]. In reality, however, it is difficult to gather the complete sets of interactions among people. Therefore, all the previous method for the influential spreaders based on the network topology could be impractical. Searching for influential spreaders without the information of a network is essential in order to prevent the global pandemic and minimize the cost for immunization.

Thus, we proposed a possible strategy for identifying the influential spreaders by using characteristics of people's behavior underlying the evolution of social networks. Our finding provides several pragmatic lessons for the efficient immunization strategy as well efficient information spreading campaigns. First, in the absence of k-shell, the degree is the first local quantity that can be used to predict the influential spreaders. From the behavioral variables quantifying the social mechanisms a_i^{α}, collective action gives a complementary information to the degree, so it is suitable for a strong indicator for influential spreaders when comparing among people with the same degree. Also, a person with a high tendency to connect two different groups via weak ties can also be suspected as a influential spreader when the network has a strong modular structure. Our analysis provide not only an applicable identifying scheme of influential spreader based on surveys but also a guideline for activity to the public, about how to behave when epidemic occurs. For instance, during the beginning stage of epidemic, one need to avoid meeting popular people or people belonging to a different group that could spread diseases to huge population.

Supporting Information

S1 Table

Multilinear regression for the QXF networks with Eq. (1)

$$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + c_6 k_i^{sh} + c_7 k_i^{2sum} + c_8 k_i^{sum} + \epsilon.$$

S2 Table

Multilinear regression for the QXG networks with Eq. (1).

$$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + c_6 k_i^{sh} + c_7 k_i^{2sum} + c_8 k_i^{sum} + \epsilon.$$

S3 Table

Multilinear regression for the POK networks with Eq. (1).

$$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + c_6 k_i^{sh} + c_7 k_i^{2sum} + c_8 k_i^{sum} + \epsilon.$$

S4 Table

Multilinear regression for the LJ networks with Eq. (1).

$$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + c_6 k_i^{sh} + c_7 k_i^{2sum} + c_8 k_i^{sum} + \epsilon.$$

S5 Table

Multilinear regression for the PRO networks with Eq. (1).
$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + c_6 k_i^{sh} + c_7 k_i^{2sum} + c_8 k_i^{sum} + \epsilon.$

S6 Table

Multilinear regression for the QXF networks with Eqs. (2) and (5).
$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + \epsilon.$

S7 Table

Multilinear regression for the QXG networks with Eqs. (2) and (5).
$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + \epsilon.$

S8 Table

Multilinear regression for the POK networks with Eqs. (2) and (5).
$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + \epsilon.$

S9 Table

Multilinear regression for the LJ networks with Eqs. (2) and (5).
$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + \epsilon.$

S10 Table

Multilinear regression for the PRO networks with Eqs. (2) and (5).
$M_i = c_0 + c_1 a_i^{exc} + c_2 a_i^{bal} + c_3 a_i^{ca} + c_4 a_i^{sh} + c_5 k_i + \epsilon.$

S1 Text

k-shell index. In order to assign k-shell index, we remove all nodes with degree less than k until no nodes less than k remains. Then the maximum subgraph whose remained degree is larger than k is called k-core. k-shell is the set of all nodes belonging to the k-core but not to the $(k+1)$-core. As a result, each node is assigned with a unique k-shell index.

S2 Text

Identifying structural hole in the link community. In order to identify the intercommunity links, called the structural hole, we use the link community detection method proposed in [24]. We adapt the method for local version only using local information of networks. When a new link e_{ik} is added, likewise the original link community algorithm [24], we define the similarity $S(e_{ik}, e_{jk})$ between two links e_{ik} and each of existed links e_{jk} by following,

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}, \tag{6}$$

where $n_+(i)$ is the set of neighbors of node i. If the similarity is less than a certain threshold meaning that two neighbors have only few fraction of common friends, we judge the newly added link as a structural hole.

Acknowledgments

This work was funded by NIH-NIGMS 1R21GM107641. FL was supported by Riksbankens Jubileumsfond (The Bank of Sweden Tercentenary Foundation) Grant nr. P12-0705:1.

References

1. Cohen R, Havlin S, ben-Avraham D, Efficient Immunization Strategies for computer networks and populations. Phys. Rev. Lett. 2003; 91: 247901.
2. Chen Y, Paul G, Havlin S, Liljeros F, Stanley HE, Finding a better immunization strategy. Phys. Rev. Lett. 2008; 101: 058701.
3. Holme P, Efficient local strategies for vaccination and network attack. Europhys. Lett. (EPL) 2004; 68: 908.
4. Domingos P, Richardson M, Mining the network value of customers. Seventh International Conference on Knowledge Discovery and Data Mining 2001.
5. Domingos P, Richardson M, Mining knowledge-sharing sites for viral marketing. Eighth Intl. Conf. on Knowledge Discovery and Data Mining 2002.
6. Albert R, Jeong H, Barabási AL, Error and attack tolerance of complex networks. Nature 2000; 406: 378.
7. Holme P, Kim BJ, Yoon CN, Han SK, Attack vulnerability of complex networks. Phys. Rev. E 2002; 65: 056109.
8. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA, Identification of influential spreaders in complex networks. Nat. Phys. 2010; 6: 888.
9. Castellano C, Pastor-Satorras R, Competing activation mechanisms in epidemics on networks. Sci. Rep. 2012; 2: 371.
10. Pei S, Muchnik L, Andrade, Jr. JS, Zheng Z, Makse HA, Searching for superspreaders of information in real-world social media. Sci. Rep. 2014; 4: 5547.
11. Kempe D, Kleinberg J, Tardos E, Maximizing the spread of influence through a social network. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining 2003; 137-143.
12. Kempe D, Kleinberg J, Tardos E, Influential nodes in a diffusion model for social networks. Proc. 32nd International Colloquium on Automata, Languages and Programming, ICALP 2005.
13. Cohen R, Erez K, Ben-Avraham D, Havlin S, Breakdown of the Internet under intentional attack, Phys. Rev. Lett. 2001; 86; 3682.
14. Dorogovtsev SN, Goltsev AV, Mendes JFF, K-core organization of complex networks. Phys. Rev. Lett. 2006; 96: 040601.
15. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E, A model of internet topology using k-shell decomposition. Proc. Nat. Acad. Sci. 2007; 104: 11150.
16. Freeman LC, Centrality in social networks: Conceptual clarification. Soc. Netw. 1979; 1: 215-239.

17. Brin S, Page L, The anatomy of a large-scale hypertextual web search engine. Comput. Networks ISDN 1998; 30: 107-117.

18. Pei S, Makse HA, Spreading dynamics in complex networks. J. Stat. Mech. 2013; P12002.

19. Lee S, Kim P, Jeong H, Statistical properties of sampled networks. Phys. Rev. E 2006; 73: 016102.

20. Contractor NS, Wasserman S, Faust K, Testing multitheoretical, multilevel hypotheses about organizational networks: An analytic framework and empirical example, Acad. Manag. Rev. 2006; 31: 681.

21. Monge PR, Contractor NS, Theories of communication networks. New York: Oxford University Press; 2003.

22. Gallos LK, Havlin S, Liljeros F, Makse HA, How people interact in evolving online affiliation networks. Phys. Rev. X 2012; 2: 031014.

23. Barabási AL, Albert R, Emergence of scaling in random networks. Science 1999; 286: 509.

24. Ahn Y, Bagrow JP, Lehmann S, Link communities reveal multiscale complexity in networks. Nature 2010; 466: 761-764.

25. Fridlund V, Stenqvist K, Nordvik MK, Condom use: The discrepancy between practice and behavioral expectation. Scandinavian Journal of Public Health. 2014; 42: 759-765.

26. Lu X, Bengtsson L, Britton T, Camitz M, Kim BJ, Thorson A, Liljeros F, The sensitivity of respondent-driven sampling. Journal of the Royal Statistical Society; Series A (Statistics in society) 2012; 175: 191-216.

27. Holme P, Edling CR, Liljeros F, Structure and time evolution of an Internet dating community. Soc. Netw. 2004; 26: 155.

28. Rocha LEC, Liljeros F, Holme P, Information dynamics shape the sexual networks of Internet-mediated prostitution. Proc. Natl. Acad. Sci. 2010; 107: 5706.

29. Clauset A, Newman MEJ, Moore C, Finding community structure in very large networks. Phys. Rev. E 2004; 70. 066111.

30. Anderson RM, May RMC, Infectious diseases of humans. Oxford: Oxford University Press; 1991.

31. Newman MEJ, The spread of epidemic disease on networks, Phys. Rev. E 2002; 66: 016128.

32. Pei S, Muchnik L, Tang S, Zheng Z, Makse HA, Exploring the complex pattern of information spreading in online blog communities. arXiv:1504.00495.

33. Gujarati D, Porter D, Basic Econometrics. 5th Ed. New York: McGraw-Hill; 1991.

34. Pastor-Satorras R, Vespignani A, Epidemic spreading in scale-free networks. Phys. Rev. Lett. 2006; 86: 3200.

35. Granovetter MS, The strength of weak ties. American journal of sociology. 1973; 1360-1380.

36. Montgomery JD, Job search and network composition: Implications of the strength of weak ties hypothesis. American Sociological Review 1992; 57: 586.

37. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási AL, Structure and tie strength in mobile communication networks. Proc. Nat. Acad. Sci. 2007; 104: 7332-7238.

38. Gallos LK, Makse HA, Sigman M, A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. Proc. Nat. Acad. Sci. 2012; 109: 2825.

39. Masuda N, Immunization of networks with community structure. New J. of Phys. 2009; 11: 123018.

40. Newman MEJ, Modularity and community structure in networks. Proc. Natl. Acad. Sci. 2006; 103; 8577-8696.