Searching for influencers in big-data complex networks

Byungjoon Min, Flaviano Morone, and Hernán A. Makse

Levich Institute and Physics Department, City College of New York, New York, NY 10031

Abstract. Localizing the most influential nodes in complex networks is a critical problem in an era of big data across different topics including information diffusion, epidemic outbreaks, and viral marketing. The common practice of identification of influencers is based on the use the number of connections called degree, which is defined locally. However, the topological location of individual would be more meaningful instead of the degree if hubs are located at the periphery of network. In addition, there is no guarantee of the performance for the widely used heuristic methods including the degree because of the lack of optimization of global influence. Here we present a strategy to search for influential individual in complex networks at single superspreader and collective influencers levels respectively. For single spreader, we find that superspreaders are located within the core identified by $k$-shell decomposition, insensitive to their degree. For collective influencers, we review theory and algorithm based on the optimization of global influence to find the optimal (minimal) set of influencers in linear time to break up the network.

1.1 Introduction

How to identify influential node in complex networks has been a crucial question for a long time due to its various applications in information dissemination [1], epidemic spreading [2], and viral marketing [3]. A set of influential nodes in a network which are much smaller than the size of the network can play a pivotal role [4–6]: (i) when activated in social networks like Twitter, they can spread information to the whole network, (ii) when removed due to breakdown, they efficiently dismantle the connectivity of the entire network, and (iii) when immunized in contact networks, they prevent the emergence of global epidemics. Finding influential nodes called influencers or superspreaders in a network has received much attention from physics, network science, and big-data science communities because of the practical importance [4–13].

The problem of identifying influencers has two distinct topics: finding individual superspreader [4] and a set of multiple influencers [6, 8] that can
optimize the global influence. For spreading processes originated from a single spreader, one of the most widely used strategies for identifying influencers in particular in networks with a scale-free degree distribution is the number of neighbors, degree, $k$ [7]. However, if a highly connected person (high degree) is located at the periphery of a network, the hub would not have much power in spreading processes. In contrast, a moderately connected person who is placed within the core of a network can have a huge impact in spreading, which would cause a large diffusion to the whole network. In this article, we illustrate that the superspreaders are consistently located in core of networks insensitive to their degree [4].

For the multiple influencers, a lot of heuristic algorithms [4,5,7,14–18] were introduced in order to estimate the influence of nodes. Since these rankings are not based on the optimization of global influence, however, they may fail to find optimal influencers. Therefore, a reliable algorithm to identify a set of influential nodes in complex networks is demanded with a rigorous ground of the optimization of global influence. The influence of nodes can be quantified by the extent of damage when nodes are removed from original network. For instance, the removal of a few influential nodes can fragment a complex network to disconnected small components. Therefore, the problem of identifying influential nodes can be mapped into optimal percolation, that is, finding the minimal set of nodes to destroy the largest connected component in a network. In other words, the most influential nodes are the minimal set of nodes so called “optimal influencers” that are responsible for global connectivity of a network. We develop an analytical framework to find a set of the optimal influencers in networks, with the mapping of optimal percolation [6]. Furthermore, we propose a scalable algorithm in linear time for the solution in massively large social media [19].

Here, we cover recent advance on the study of theory of finding influencers for the single spreader and collective influencers [4,6,19]. We first discuss the identification of superspreader originated from a single spreader and show that topological location plays a crucial role rather than the connections defined locally. Next, we review the theory of collective influencers with mapping of optimal percolation to search for optimal influencers that are the minimal set of nodes to break up the whole network.

1.2 Identifying superspreaders in complex networks

Spreading usually starts from a single origin in epidemics spreading and rumor diffusion [1, 2]. It is widely believed that a node with the highest degree would be the superspreader, among others [7]. However, the highly connected nodes would have a small impact on spreading processes when the high degree nodes are placed at the periphery of a network, surrounded by less connected nodes. Alternatively, nodes located within the core of a network can play an important role in spreading even though the nodes are less connected.


An example of the $k$-shell decomposition in a network. The two nodes with the same degree $k = 8$, blue and yellow nodes, are located respectively in the innermost core ($k_S = 3$) and the periphery ($k_S = 1$) in this network. The extent of the spreading efficiency in the SIR model originated from a single spreader for three different cases: $(k_S, k) = (63, 96)$ (node A), $(k_S, k) = (26, 96)$ (node B), and $(k_S, k) = (63, 65)$ (node C). On top of the contact network of inpatients (CNI), when origin has high $k$-shell index, i.e. $k_S = 63$ for nodes A and C, spreading covers a wide area of the entire network both for nodes A and C similarly, although the degree of C is much smaller than A. However, spreading remains largely localized for node B despite of its high degree, meaning that the degree cannot predict the spreading efficiency reliably. The results are obtained over over 10,000 different realizations with infection probability $\beta = 0.035$.

In order to validate this hypothesis, we study real-world complex networks as examples of complex networks of social interactions. We use four social networks: (i) the network of email contacts in the Computer Science Department of the University College London (Zhou, S., private communication), (ii) contact network of inpatients (CNI) from hospitals in Sweden [20], (iii) the network of actors who have starred in the same movie labeled by imdb.com as adult [21], and (iv) the friendship network in the LiveJournal.com community [22].
On top of the real-world networks, we study spreading process using a prototypical epidemic model, Susceptible-Infectious-Recovered (SIR) model [2,23]. In the SIR model, each node can be in one of three states, susceptible, infected, and recovered (or removed). Initially, all nodes are susceptible (S) except for one infected (I) node. The infected nodes infect their susceptible neighbors with the probability $\beta$ and then becomes recovered (R) where they cannot become infected again. The SIR model can describe not only epidemic spreading but also information and rumor spreading [24]. In this study, we use relatively small values for infection probability $\beta$, so that only limited fraction of nodes will be infected. Otherwise, a large fraction of the population will be infected and the effect of the initially infected node is not significant.

The core-periphery structure of complex networks are identified by using the $k$-shell (also known as $k$-core) decomposition [25–27]. We assign $k$-shell index, $k_S$, to each node after successive removals of nodes with degree smaller than $k_S$. Initially, the nodes with degree $k = 1$ are removed. After pruning all the nodes with degree $k = 1$, we recalculate the degree of each node and continue to remove the nodes with left degree $k = 1$. After removing all nodes with degree $k = 1$ iteratively, all the removed nodes become a $k$-shell with index $k_S = 1$. We iteratively remove and assign the next $k$-shell, $k_S = 2$. We continue to remove and assign higher $k$-shells until all nodes are removed from the original network. Then, each node is defined by a unique $k_S$ index according to their core-periphery structure determined by network topology. The periphery of networks is defined as small values of $k_S$ whereas the core shows large $k_S$ (Fig. 1.1a).

In order to quantify the spreading efficiency of a node $i$ in the SIR model, we measure the infected fraction in an epidemic originating at node $i$. Figures 1.1b-c show that the degree may not predict accurately the size of the infected nodes in a spreading process. The infected fraction can be far different even though the spreading starts from the same degree as shown in Figs. 1.1b-c. Instead, the size of infected population is predicted more accurately when we use the topological location of the origin defined by $k_S$ index. In Figs 1.1b and d, the nodes with the same $k$-shell index bring out similar infected populations even though they have very different $k$. This example implies that the coreness of the node has better predictive power than the degree.

For more systematic analysis, we quantify the influence of a node $i$ in the SIR model by using the average size of infected nodes $M_i$. Then, the infected nodes averaged over all the origins with respect to $(k_S,k)$ is given by

$$M(k_S,k) = \sum_{i \in \mathcal{T}(k_S,k)} \frac{M_i}{N(k_S,k)},$$

where $\mathcal{T}(k_S,k)$ is the union of all $N(k_S,k)$ nodes with $(k_S,k)$ values. We examine $M(k_S,k)$ in the four social networks to reveal the influence of nodes with $(k_S,k)$ (see Fig. 1.2). For the same $k$-shell index, $M(k_S,k)$ is insensitive to the degree as shown vertically in Fig. 1.2. However, $M(k_S,k)$ exhibits wide
Fig. 1.2. Average fraction of infected size $M(k_S,k)$ when spreading originates from nodes with $(k_S,k)$. We use for networks a, email contacts ($\beta = 8\%$), b, CNI network ($\beta = 4\%$), c, the actors network from IMDB data ($\beta = 1\%$), and d, the friendship network from Livejournal.com ($\beta = 1.5\%$). The $k$-shell index is more reliable predictor for superspreaders than the degree $k$. Average fraction is larger for nodes with high $k$-shell index $k_S$ independent to $k$ whereas origins of a given $k$ can cause either small or large infection, depending on $k_S$.

range for the same degree $k$, meaning that nodes with the same degree can have a different influence in the spreading process. This result implies that influential spreaders are consistently placed within the core identified by high $k$-shell index, regardless of the degree. Therefore, the $k$-shell index of a node is a better predictor of the spreading efficiency than the local property, like the degree.

We demonstrate that “superspreaders” are consistently located in the core of a network defined by the $k$-shell index insensitive to their degree. When spreading starts from the core of a network (high $k$-shell), the spreading origin is surrounded by other core nodes structurally. Thus, there exist many pathways to the rest of the network from the origin, so that the spreading can
easily reach a critical mass that maintains global outbreaks. However, when spreading originates from a high $k$ node but located at periphery, it is not guaranteed that the spreading will develop.

$k$-shell decomposition allows us to identify “superspreaders” in a general spreading process. For a wide range of infection probability $\beta$, the inner-core nodes identified by $k$-shell index shows higher spreading efficiency than outer shells [4]. In a different class of spreading model, susceptible-infected-susceptible model, we also find that the persistence of infected fraction is larger for nodes with high $k$-shell index than those with low [4]. Therefore, the $k$-shell index is a robust and reliable predictor to identify influential spreader on complex networks.

1.3 Identifying collective influencers in complex networks

In this section, we discuss the influence of multiple nodes as a general problem of node’s influence. Suppose that we conduct a viral marketing campaign with a limited resource and so we should choose a small set of nodes for the campaign. Then, we have to select “collective influencers” who can potentially spread the information to the larger part of the social network. Inversely, if the collective influencers are removed from the original network, it becomes fragment with small components.

Consider a network with $N$ nodes connected by $M$ edges, with an arbitrary degree distribution $P(k)$. When we remove a fraction $q$ of nodes in random, at a critical fraction the giant component of the network is destroyed, $G(q) = 0$ where $G$ the size of the giant component. Thus, the problem of optimal influencers can be mapped into finding the minimum fraction $q_c$ to dismantle a network, $G(q) = 0$.

To develop theory of collective influencers, we first introduce auxiliary variables $\nu_{i \rightarrow j}$ representing the probability that node $i$ belongs to the giant component when $j$ is absent. On a locally tree-like network, the variable can be expressed as

$$\nu_{i \rightarrow j} = n_i \left[ 1 - \prod_{k \in \partial i \setminus j} (1 - \nu_{k \rightarrow i}) \right],$$  

(1.2)

where $\partial i \setminus j$ is a set of nearest neighbors of $i$ excluding $j$ and $n_i$ represents whether the node $i$ is removed or not, i.e., $n_i = 0$ if node $i$ is removed and $n_i = 1$ if node $i$ is present. Once we calculate $\nu_{i \rightarrow j}$ for each connection, we can finally obtain the probability $\nu_i$ that node $i$ belongs to the giant connected component by following,

$$\nu_i = n_i \left[ 1 - \prod_{j \in \partial i} (1 - \nu_{j \rightarrow i}) \right].$$  

(1.3)
The value $\nu_{i \rightarrow j} = 0$ for all pairs is trivially a solution of Eq. (1.2), corresponding to the situation that no node belongs to the giant component. If the solution is stable, the iteration of Eq. (1.2) will converge into the trivial solution with no giant component. If it is not stable, we will obtain finally a different stable fixed solution and there should be a giant component. This fact allows us to apply stability analysis at the value $\nu_{i \rightarrow j} = 0$ in order to predict percolation threshold precisely.

By linearizing Eq. (1.2), we can determine the stability of the solution $\nu_{i \rightarrow j} = 0$ by the largest eigenvalue $\lambda(n; q)$ of the linear operator $\hat{M}$ defined on the $2M \times 2M$ directed edges as

$$\mathcal{M}_{k \rightarrow i, i \rightarrow j} = \frac{\partial \nu_{i \rightarrow j}}{\partial \nu_{k \rightarrow i}} |_{\{\nu_{i \rightarrow j} = 0\}}.$$  \tag{1.4}

For a locally-tree like network, $\hat{M}$ is given by

$$\mathcal{M}_{k \rightarrow i, i \rightarrow j} = n_i B_{k \rightarrow i, i \rightarrow j}$$  \tag{1.5}

where $B_{k \rightarrow i, i \rightarrow j}$ is the non-backtracking or Hashimoto matrix of the network \[28–30\]. If $\lambda(n; q) \leq 1$, the solution $\nu_{i \rightarrow j} = 0$ is stable. Hence, the optimal influence problem can be interpreted as identifying the minimal set of nodes to be removed that minimize the the largest eigenvalue $\lambda(n; q)$. The largest eigenvalue of $\hat{M}$ can be calculated by the Power Method:

$$\lambda(n) = \lim_{\ell \to \infty} \left[ \frac{|\hat{M}^\ell w_0|}{|w_0|} \right]^{1/\ell},$$  \tag{1.6}

where $w_0$ is an arbitrary non-zero vector with $2M$ entries.

For a finite $\ell$, we define the cost energy function $|\hat{M}^\ell w_0|$ from Equation (1.6) to be minimized. The solution of minimized cost function by perturbation series is

$$|\hat{M}^\ell w_0|^2 = \sum_{i=1}^{N} (k_i - 1) \sum_{j \in \partial \text{Ball}(i, 2\ell - 1)} \left( \prod_{k \in P_{2\ell-1}(i,j)} n_k \right) (k_j - 1),$$  \tag{1.7}

where $\text{Ball}(i, \ell)$ is the set of nodes inside a ball of radius $\ell$ defined as the shortest path around node $i$, $\partial \text{Ball}(i, \ell)$ is the frontier of the ball and $P_{\ell}(i, j)$ is the shortest path of length $\ell$ connecting $i$ and $j$.

From the cost energy function Eq. (1.7), we establish systematically the algorithm of node’s influence. Collective influence (CI) of node is defined as the extent of the drop in the energy function when we remove a node from the network. The algorithm for the collective influence is following: First, we identify the nodes belonging to the frontier of a given ball of radius $\ell$ of every node. Next, we assign to node $i$ the strength of collective influence by

$$\text{CI}_\ell(i) = (k_i - 1) \sum_{j \in \partial \text{Ball}(i, \ell)} (k_j - 1).$$  \tag{1.8}
Then, remove node $i^*$ with the highest $\text{CI}_i$ and set $n_{i^*} = 0$. As removing node $i^*$, the degree of its neighbors decreases by one. The above procedures are repeated to identify next highest CI node to remove until the giant component vanishes, $G(q) = 0$. Based on CI algorithm, we identify a set of collective influencers to destroy the giant component efficiently. In addition, this algorithm is able to find the minimal set of nodes in almost linear time with optimal implementation [19].

In order to validate the performance of CI algorithm, we compare the percolation threshold of CI and that of many heuristic centrality measures: high-degree (HD) [7], high-degree adaptive (HDA), PageRank (PR) [15], closeness centrality (CC) [14], eigenvector centrality (EC) [14], and k-core [4, 25] (Fig. 1.3). Figure 1.3a compares CI in Erdős-Rényi (ER) networks against other heuristic algorithms. For scale-free (SF) networks, we also compare CI with HDA and HD methods. In all networks, CI produces smaller percolation threshold than the other heuristic methods that we tested.
As an example of real-world network for information spreading, we construct a network of Twitter users. Figure 1.4 exhibits the size of giant component of Twitter users with respect to a removed fraction $q$. In the Twitter network, CI algorithm outperforms other heuristics in identification of optimal influencers. The CI algorithm allows us to find weak-nodes which are the low-degree nodes surrounded by hierarchically structured coronas of hubs as shown in Fig. 1.4b, which are not captured by other heuristics. A large number of weak-nodes identified by CI algorithm play an important role in global influence, so that the CI algorithm leads to nearly optimal solution for finding collective influencers.

### 1.4 Summary

We study the identification of the most influential individuals in complex networks for the individual spreader and collective influencers respectively. For a single spreader, we find that influential individuals in spreading processes are consistently placed within the core area of complex networks defined by $k$-core percolation, independent to their degree. For collective influencers, we propose an optimal method to search for influential nodes based on minimization of cost energy function of global influence, allowing us to find a new class of influencers called weak-nodes that are weakly connected nodes yet surrounded by hierarchical coronas of hubs.

Our theory for influential nodes in complex networks has a variety of applications. One of the most promising examples would be finding important area in the brain. For the brain networks, finding the most important regions...
allows us to design protocols to control brain functions by regulating these areas. In a broad sense, this research of identifying important areas in the brain can open a novel way to handle neurological disease with intervention to core areas prescribed by network theory, which would be a seminal advance [12].

1.5 Acknowledgement

This work was funded by NSF-PoLS PHY-1305476, NIH-NIGMS 1R21GM107641, NSF IIS-1515022, and Army Research Laboratory Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA).

References

21. The Internet Movie Database, www.imdb.com