Identifying Influential Nodes in Social Networks Based on Social Strength

Yanli Lee$^{1,*}$ and Yan Fu$^1$

$^1$CompleX Lab, University of Electronic Science and Technology of China, Chengdu 611731, China
Email: yanlicomplex@gmail.com; fuyan@uestc.edu.cn
$^*$Corresponding Author Email: liyanli@std.uestc.edu.cn

Abstract. Identifying influential nodes in social networks is significant in controlling the outbreak of epidemics, conducting advertisements for e-commercial products, predicting popular scientists or papers, and so on. Many methods have been proposed by paying much more attention to the role of nodes, while the social strength between nodes is rarely considered. Especially in unweighted networks, social strength is not explicit information. In this paper, we first characterize the social strength of unweighted networks, and then proposed a LOVital method based on the social strength. Numerical results on the relation network of Weibo dataset and the other datasets from various fields show that LOVital can much more accurately identify influential users.

1. Introduction

Network science brings a new perspective for many research fields, including social science, biology, economics, data science and so on [1-4]. Starting with revealing the macroscopic statistical regularities (e.g., scale-free [5], assortative mixing [6], small-world [7] and clustering [7]) and discovering the mesoscopic structural organization (communities [8, 9] and motifs [10]), till now, the network science mainly focuses on distinguishing roles played by individual nodes and links [11, 12]. In particular, identifying influential nodes has a wide range of practical applications [12, 13]. For example, deploying medical resources to influential spreaders or places can avoid the outbreak of epidemics effectively. Recommending new products to vital opinion leaders can successfully accelerate the spreading speed and the final adoption size of products (e.g., the scenario of influencer marketing).

The inhomogeneity of real networks implies the challenge to identify the influential nodes [14]. Considering the disease control scenario, compared with the dense homogeneity networks, the scale-free property emerged in many real sparse networks enhances the disease spreading ability, which makes the random immunization strategy failed to control the outbreak of epidemics even with high fraction of immunized individuals [15]. To tackle this problem, many topology-based methods are proposed to identify the effective influential nodes, including degree, betweenness [16], coreness [17-19], h-index [20] and PageRank [21]. All of these methods ignore the social strength. Even though the weight of each link can be employed as the social strength, it is still an open problem on unweighted networks.

The development of link prediction technique sheds light on this problem [11]. On one hand, we assume that two nodes with closer relationship are much more likely influenced by each other. Link prediction methods provide abundant ideas to characterize the social strength of two nodes, which can be used to design better influential nodes identification methods; On the other hand, link prediction and influential nodes identification problems always focus on some same scenarios. Employing the
results of each other to improve the performance of themselves is important and resource-saving. For example, predicting the latent relationship between users and ranking the influential users are all necessary for the company of Weibo. Thus, in this paper we design a new method LOVital based on the social strength characterized by the LO method [22]. LOVital performs better than other benchmarks on Weibo datasets. We also apply this method to other real networks from various fields, and it also performs best.

2. Methods
Given a network $G$, we formulate it as an adjacency matrix $A$, with $a_{ij} = 1$ if there is a link between nodes $i$ and $j$, and $a_{ij} = 0$, otherwise. Based on the basic assumption that one user is much more likely to be influenced by users with stronger social strength, we proposed the LOVital method. It consists of two steps:

- Quantifying the social strength of node pairs. A link prediction method linear optimization (short for LO) [22] is introduced here. LO is used to predict missing links in networks. The connection likelihood of nodes $i$ and $j$ can be denoted as

$$s_{ij} = \sum_k a_{ik} z_{kj}$$

Where $z_{kj}$ is an unknown value, and we can obtain it by minimizing the following loss function

$$\min_{z} \alpha ||A - AZ||_F^2 + ||Z||_F^2$$

Where $||X||_F^2 = \text{Tr}(X^TX)$, and $\alpha$ is a free parameter. Finally, we can obtain $S = AZ^* = \alpha A(\alpha A^TA + 1)^{-1} A^TA$, and $\alpha$ is set to 0.005 in this paper. Each element in $S$ is used as the social strength of the corresponding node pair.

- Quantifying the influence of the target node. We quantify the initial influence of the user $i$ as $l_i^0 = \sum_j s_{ij}$. And the final influence of user $i$ is defined as

$$l_i = l_i^0 + \sum_j S_{ij} l_j^0$$

3. Benchmarks
We compare the proposed method LOVital with five benchmarks, including degree, betweenness [16], coreness [17-19], $h$-index [20] and PageRank [21].

- **Degree.** The degree of the node $i$ is the number of the nearest neighbors of the node $i$, and it can be denoted as $k_i = \sum_j a_{ij}$.

- **Betweenness** [16]. The betweenness of the node $i$ counts all of the shortest paths passing through $i$, and it is defined as $BC_i = \sum_{s,t,s\neq t} g_{st}^i$, where $g_{st}$ is the number of shortest paths between nodes $s$ and $t$, and $g_{st}^i$ is the number of shortest paths that passing through $i$ among all the shortest paths between $s$ and $t$.

- **Coreness** [17-19]. It is proposed to mark the location of nodes in the topology of the network. The influence of one node in the core part is higher than the one that located in the periphery. It is calculated by the following steps: Step 1: removing all the nodes with smallest degree $k_c$; Step 2: updating the degree of each node in the remaining network, and removing all the nodes with degree not greater than $k_c$. Repeating step 2 until the smallest degree is greater than $k_c$. All the removed nodes in step 1 and step 2 form the $k_c$-shell, and the coreness values of the nodes in $k_c$-shell are $k_c$. Step 3: repeating step 1 and step 2 until all the nodes are removed, and the coreness of each node equals its corresponding shell layer.

- **$H$-index** [20]. The $h$-index of a node $i$ is defined as the largest $h$ that satisfies the node $i$ with at least $h$ neighbors with degrees no less than $h$. 

• PageRank [21]. The PageRank value of the node \( i \) is defined as

\[
Pr_i^t = \gamma \sum_j \left( Pr_j^{t-1} \frac{a_{ij}}{k_j} \right) + (1 - \gamma) \frac{1}{N},
\]

where \( \gamma \) is used to control the probability of a random jump, and \( \alpha \) is set to 0.85 in this paper. The initial value \( Pr_i^0 = \frac{1}{N} \), and the iteration process will stop if the PageRank values of all nodes reach the steady state.

4. Evaluation Criteria
The ground-truth of users’ influence is lacking, and it is still a challenge problem in influential nodes identification problem. Many spreading models are employed as alternatives to obtain the spreading influence of nodes, e.g., the susceptible-infected-recovered model (SIR for short) [23], the susceptible-infected-susceptible (SIS model) [24], etc. In this paper, we take the SIR model to evaluate the ranking results of methods. The influence \( Q \) of the target node in the SIR model is defined as the number of infected nodes in steady state averaged over 1000 independent runs. In particular, at the start-up of the SIR model, all the nodes are in the susceptible state except the target node, which is the unique infected node. At each step, the neighbours of every infected node are infected with probability \( \beta \), and each infected node can recover with probability \( \mu \). We set \( \mu = 1 \), \( \beta = 1.5 \beta_c = \frac{1.5(k)}{\langle k \rangle} \), where \( \beta_c \) is the epidemic threshold of the SIR model obtained by the heterogeneous mean-field theory, and \( k \) is the degree of each node. The Kendall’s tau correlation coefficient \( \tau \) [25] is used to evaluate the relevance of the rankings of the considered methods and the ranking obtained by the SIR model. The larger the \( \tau \), the better the related method in mining the influential users.

5. Results
In this section, we first take the Weibo dataset as an example to analyze the identification performance of LOVital; Furthermore, we further test the performance of LOVital on much more datasets from various fields.

5.1 The Analysis on Weibo Dataset
Among many social relationship scenarios, the follower-followee relationship in Weibo.com is typical. Many social marketing activities can be applied to the platform of Weibo.com by employing this type of relationship. In this work, we focus on a snapshot of a religion relationship network of Weibo.com [26], which is collected in April 2016. This network consists of 6875 believers in Christianity, Buddhism, Islam and Taoism, and 64712 undirected links between different users. The epidemic threshold of the SIR model \( \beta_c = 0.0066 \). Each node denotes one registered user on Weibo.com and each undirected link denotes the follower-followee relationship between two nodes.

The network structure can be found in figure 1 (a), and the color of the nodes shows the influence \( Q \) of each node which is obtained by the SIR model in section 4. The redder the color is, the more influential the corresponding node has, while the greener the color is, the less influence the corresponding node has. The tail of the distribution of \( [Q] \) follows \( p([Q]) \sim [Q]^{-r} \) with \( r = 2.6 \) (see figure 1 (b), which indicates that to identify the influential nodes is significant, as a large number of nodes can be infected by few influential nodes (values of the discrete variable \( Q \) are converted to integers here). As shown in figure 1 (c), compared with the other five benchmarks, LOVital performs overall best with larger Kendall’s tau coefficient \( \tau \) between the ranking obtained by LOVital and the ranking obtained by SIR model. Moreover, LOVital can much more accurately identify the top-N influential nodes. Most of the time, identifying the most influential users is much more important than getting an accurate ranking of users. As shown in figure 1 (d), the average influence \( Q \) of top-N nodes as ranked by LOVital is higher than the other five benchmarks. Meanwhile, we can find that among the nodes with large degree, there are some nodes with small \( Q \) (see the nodes pointed by red arrows in figure 1 (e), and the sizes of the circles are small), and these nodes are ranked behind by the LOVital method. Broadly speaking, LOVital can identify the influential nodes better in the Weibo dataset compared with the other five benchmarks.
4. Figure 1. (a) The religion network with the color showing the influence $Q$ of each node. (b) The distribution of the influence $[Q]$; (c) The Kendall’s tau coefficient $\tau$ between the rankings of six methods and the ranking obtained by SIR model. (d) The average influence $Q$ of top-N nodes as ranked by the six methods. (e) Degrees of nodes as ranked by the LOVital method, and the size of the circle is proportional to the influence $Q$ of the corresponding node.

5.2 The Applications on Datasets from Various Fields
In this subsection, we take much more datasets from various fields to test the performance of the proposed method. They include (i) DNC [27] – an email network, (ii) AB and BG [27] – two hyperlink networks, (iii) HS [27] – a friendship network, (iv) OPS [27] – a message sending network and (v) SmaG [28] – a citation network. The topology statistics of these networks are illustrated in table 1. The values of Kendall’s tau coefficient $\tau$ between the ranking of nodes obtained by the six methods and the ranking of nodes obtained by the SIR spreading mode are shown in table 2. Again, we find that LOVital performs overall best among the six methods on six real networks.

6. Conclusion
Social strength is an important factor to characterize the influence of a node, yet it is rarely considered on unweighted networks. In this paper, we proposed a new influential nodes identification method by considering the social strength of node pairs. It provided an idea to apply the link prediction technique to influential nodes identification problems. The SIR model is used to evaluate the performance of the related methods. Numerical experiments on the Weibo dataset and six real networks show that LOVital can better characterize the influence of each node, and it performs better compared with the other five benchmarks.

In the booming of information age, many data mining techniques that focusing on different problems are proposed. Employing the techniques to design influential nodes identification method not only can save computing resources, but also can improve the identification performance. Since there are many data mining problems occurring in the same platform or system, there are some potential relation among them, which can be better employed to promote their algorithmic performance. It is necessary for practical applications to develop a series of methods to combine them in the future.
Table 1. The structural statistics of 6 simple networks. \(|V|\) and \(|E|\) are the number of nodes and links of the network, respectively. \(\langle c \rangle\) is the clustering coefficient [6]. \(\rho = \frac{2|E|}{|V|(|V|-1)}\) is the density of the network. \(r\) is the network diameter.

| Networks | \(|V|\) | \(|E|\) | \(\langle c \rangle\) | \(\rho\) | \(r\) |
|----------|--------|--------|---------------|-------|------|
| DNC      | 1866   | 4384   | 0.22          | 0.0026| 8    |
| AB       | 146    | 180    | 0.05          | 0.0170| 6    |
| BG       | 1224   | 16715  | 0.32          | 0.0224| 8    |
| HS       | 70     | 274    | 0.46          | 0.1135| 6    |
| OPS      | 1899   | 13838  | 0.11          | 0.0077| 8    |
| SmaG     | 1024   | 4916   | 0.31          | 0.0094| 6    |

Table 2. The Kendall’s tau correlation coefficient \(\tau\) between the ranking scores obtained by the methods and the ranking of nodes obtained by the SIR spreading model. The approximate epidemic threshold \(\beta_c\) of DNC, AB, BG, HS, OPS and SmaG are 0.014, 0.072, 0.012, 0.120, 0.018 and 0.027, respectively.

| Networks | Degree | Betweenness | Coreness | H-index | PageRank | LOVital |
|----------|--------|-------------|----------|---------|----------|---------|
| DNC      | 0.490  | 0.343       | 0.484    | 0.663   | 0.624    | **0.881** |
| AB       | 0.358  | 0.305       | 0.337    | 0.678   | 0.546    | **0.850** |
| BG       | 0.914  | 0.826       | 0.913    | 0.573   | 0.899    | **0.959** |
| HS       | 0.826  | 0.751       | 0.716    | 0.573   | 0.824    | **0.944** |
| OPS      | 0.868  | 0.828       | 0.861    | 0.634   | 0.874    | **0.951** |
| SmaG     | 0.792  | 0.755       | 0.783    | 0.571   | 0.801    | **0.919** |

7. References
[1] Newman M. E. J. 2018 Networks. (Oxford: Oxford university press)
[2] Barabási A. L. 2016 Network science. (Cambridge: Cambridge university press)
[3] Fortunato S. et. al. 2018 Science of science. Science 359 eaa0185
[4] Gao J., Zhang Y. C. and Zhou T. 2019 Computational socioeconomics. Phys. Rep. 817 1-104
[5] Barabási A. L. and Albert R. 1999 Emergence of scaling in random networks. Science 286 509–512
[6] Newman M. E. J. 2002 Assortative mixing in networks. Phys. Rev. Lett. 89 208701
[7] Watts D. J. and Strogatz S. H. 1998 Collective dynamics of small-world networks. Nature 393 440–442
[8] Newman M. E. J. and Girvan M. 2004 Finding and evaluating community structure in networks. Phys. Rev. E 69 026113
[9] Xie W. B., Lee Y. L., Wang C., Chen D. B. and Zhou T. 2020 Hierarchical clustering supported by reciprocal nearest neighbors. Inf. Sci. 527 279-292
[10] Alon U. 2007 Network motifs: theory and experimental approaches. Nat. Rev. Genet. 8 450–461
[11] Lü L. and Zhou T. 2011 Link prediction in complex networks: A survey. Physica A 390 1150-1170
[12] Lü L., Chen D. B., Ren X. L., Zhang Q. M., Zhang Y. C. and Zhou T. 2016 Vital nodes identification in complex networks. Phys. Rep. 650 1–63
[13] Lee Y. L. and Fu Y. 2019 National Economic Status Inference from the World Trade Network. In 16th Int. Computer Conf. on Wavelet Active Media Technology and Information Processing (IEEE) pp. 96-99
[14] Albert, R., Jeong, H. and Barabási, A. L. 1999 Diameter of the world-wide web. Nature 401 130-131
[15] Cohen R, Havlin S and Ben-Avraham D. 2003 Efficient immunization strategies for computer networks and populations. Phys. Rev. Lett. 91 247901.
[16] Freeman L. C. 1977 A set of measures of centrality based on betweenness. 1977 Sociometry 40
35–41

[17] Dorogovtsev S. N., Goltsev A. V. and Mendes J. F. F. 2006 K-core organization of complex networks. *Phys. Rev. Lett.* **96** 40601

[18] Liu L., Zhou T., Zhang Q. M. and Stanley H. E. 2016 The H-index of a network node and its relation to degree and coreness. *Nat. Commun.* **7** 1-7

[19] Lee Y. L. and Zhou T. 2017 Fast asynchronous updating algorithms for k-shell indices. *Physica A* **482** 524–531

[20] Hirsch J. E. 2005 An index to quantify an individual’s scientific research output. *Proc. Natl. Acad. Sci. U. S. A.* **102** 16569–16572

[21] Brin S. and Page L. 1998 The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw.* **30** 107–117

[22] Pech R., Hao D., Lee Y. L., Yuan, Y. and Zhou, T. 2019 Link prediction via linear optimization. *Physica A* **528** 121319

[23] Anderson R. M., May R. M. and Anderson B. 1992 *Infectious diseases of humans: dynamics and control* (Oxford: Oxford University Press)

[24] Hethcote H. W. 2000 The mathematics of infectious diseases. *SIAM Rev.* **42** 599–653

[25] Nelsen, R. B. 2001 Kendall tau metric. (Berlin: Springer)

[26] Hu J., Zhang Q. M. and Zhou T. 2019 Segregation in religion networks. *EPJ Data Sci.* **8** 6

[27] Downloaded from the KONECT website: http://konect.uni-koblenz.de/networks/

[28] Batagelj V. and Mrvar A. 2006 Available at: http://vlado.fmf.uni-lj.si/pub/networks/data/