Comparison of Different Centrality Measures to Find Influential Nodes in Complex Networks

Fanpeng Meng (✉), Yijun Gu (✉), Shunshun Fu (✉), Mengdi Wang, and Yuchen Guo

College of Information Technology and Network Security, People’s Public Security University of China, Beijing 102600, China
guyijun@ppsuc.edu.cn, mengfanpeng0202@126.com

Abstract. In this paper, we compare the performance of representative centrality measures, classical and up-to-date, on more real networks in various fields. With the aid of SIR information diffusion model to simulate the vertices’ influence in real networks, we apply the kendall’s tau correlation coefficient, distinguishability and robustness to test different centrality measures at the same level., to show the best application scenarios for certain measure.

Keywords: Influential nodes · Comparison of centrality measures
Centrality methods · Complex networks · Social networks

1 Introduction

Studies have shown that influential nodes play an important role in all kinds of dynamic behavior in the complex network. Identifying influential nodes allow us to better control epidemic outbreaks, accelerate information propagation, conduct successful e-commerce advertisements, and so on. In this paper, we compare the performance of representative centrality measures, classical and up-to-date, on more real networks in various fields. With the aid of SIR information diffusion model to simulate the vertices’ influence in real networks, we apply the kendall’s tau correlation coefficient, distinguishability and robustness to test different centrality measures at the same level. All these work is aimed to provide a deep understanding of various characteristics and best application scenarios for certain measure.

The rest of the paper is organized as follows. In Sect. 2, we briefly overview of centrality measures mentioned above. In Sect. 3, we describe the dataset, the influence simulation model SIR model, and evaluation criteria. In Sect. 4, experiment results are illustrated to show the characteristics of centrality measures. Finally, some conclusions are presented in Sect. 5.

2 Review of Centrality Measures

Many centrality methods have been proposed to measure the estimated importance of nodes within the networks. In this paper, we divide these measures into four categories. They are local-based centrality measures, global-based centrality measures, semi-local-based centrality measures, and multi-centralities based measures.
2.1 Local-Based Centrality Measures

The local-based measures tend to capture the features of the node through the partial information around it in general, such as degree centrality (dc) and K-shell (ks) decomposition method [1].

2.2 Global-Based Centrality Measures

The global-based methods considering global information gives ranking results much better, such as betweenness centrality (bc) and closeness centrality (cc) and PageRank [2] (pg).

2.3 Semi-Local-Based Centrality Measures

Semi-local-based centralities are most widely studied nowadays which are tradeoff of the local structure and global structure, such as Local centrality (LC) [3], local structural centrality (LSC) [4], Local Weight (LW) [5], Sum of Edge Importance Coefficient (SEIC) [6], Local Triangle-based Centrality (LTC) [7], Coefficient of Local Centrality (CLC) [8], Two-Hop Connected Coreness Centrality (THCC) [9].

The vast majority of these semi-local-based centrality measures consider several hop of neighbors and have better performance both in computation complexity and effect than classical centrality measures, Multi-centralities-based Measures.

2.4 Multi-centralities-Based Measures

Some researchers think node importance is not affected by a single factor, but is affected by a number of factors. Hence a new evaluation method of node importance in social network is proposed, based on multi-centralities, i.e., dc, ks, bc, cc, pg, etc. The majority of the multi-centralities-based methods [10, 11, 12, 13] apply a multiple attribute decision making (MADM). However, some centrality measures in MADM are very time-consuming. Hence multi-centralities-based measures are out of discussion in this paper.

3 Data and Evaluation Criteria

3.1 Datasets

In this study, we focus on real social networks and the datasets used in the paper are listed: Karate [16], Jazz [17], Netscience [18], Facebook [19], Email [20], Blogs [21], CA-HepPH [22], PGP [23], Twitter [24], Epinions [25], Slashdot [26].

From Table 1, we can observe the computational complexity of the centrality measures above. An outline of some of the basic properties of these networks is shown in Table 2. In this table, n is the number of nodes, m is the number of edges, \( \bar{k} \) is the average degree in the network, \( \bar{d} \) is the average distance between reachable pairs of nodes in networks, \( \bar{cc} \) is the average clustering coefficient of nodes, \( \beta_{th} \) is the epidemic threshold calculated by
\[
\beta_{th} = \frac{k}{<k^2> - <k>} \quad [27, 28].
\]
3.2 SIR Model Simulation

To simulate a realistic spreading process and obtain the true spreading in fluence of nodes, we adopted the susceptible-infected-recovered (SIR) model [29]. The spreading ability of the original node $v$, $S_b(v)$, is defined as the number of nodes that were infected by the end of the spreading process that originated from node $v$. We assigned a small value to infection probability $b$, which was approximately the epidemic threshold $b_{th} = \frac{1}{\langle k^2 \rangle - \langle k \rangle}$ [27, 28], where $\langle k \rangle$ is the ith moment of the degree distribution [30]. We set the number of simulations to be 1000. The spreading in fluence of a node is defined as the average spreading ability of node $v$, except for ‘epinions’ and ‘slashdot’ whose size is too big to simulate the influence on the whole range of $b$ but just simulate on the $b_{th}$.

### Table 1. Summary of the ranking methods mentioned in this paper

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Complexity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree centrality (dc)</td>
<td>$O(n)$</td>
<td>[14]</td>
</tr>
<tr>
<td>k-shell decomposition (ks)</td>
<td>$O(m)$</td>
<td>[1]</td>
</tr>
<tr>
<td>PageRank (pg)</td>
<td>–</td>
<td>[2]</td>
</tr>
<tr>
<td>Closeness centrality (cc)</td>
<td>$O(n(n + m))$</td>
<td>[15]</td>
</tr>
<tr>
<td>Betweenness centrality (bc)</td>
<td>$O(nm)$</td>
<td>[14]</td>
</tr>
<tr>
<td>Local centrality index (LC)</td>
<td>$O(n\langle k \rangle^2)$</td>
<td>[3]</td>
</tr>
<tr>
<td>Local structure centrality (LSC)</td>
<td>$O(n\langle k \rangle^2)$</td>
<td>[4]</td>
</tr>
<tr>
<td>Local weight index (LW)</td>
<td>$O(nk)$</td>
<td>[5]</td>
</tr>
<tr>
<td>Sum of edges importance centrality (SEIC)</td>
<td>$O(n\langle k \rangle^2)$</td>
<td>[6]</td>
</tr>
<tr>
<td>Local triangle centrality (LTC)</td>
<td>$O(n\langle k \rangle^2)$</td>
<td>[7]</td>
</tr>
<tr>
<td>Local structure with a coefficient index (CLC)</td>
<td>$O(n\langle k \rangle^2)$</td>
<td>[8]</td>
</tr>
<tr>
<td>Two-Hop connected coreness index (THCC)</td>
<td>$O(n\langle k \rangle^2)$</td>
<td>[9]</td>
</tr>
</tbody>
</table>

### Table 2. The basic topological properties of the real networks studied in this work

<table>
<thead>
<tr>
<th>Network</th>
<th>$n$</th>
<th>$m$</th>
<th>$\tilde{k}$</th>
<th>$\tilde{d}$</th>
<th>$\bar{c}$</th>
<th>$\beta_{th}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karate</td>
<td>34</td>
<td>78</td>
<td>4.59</td>
<td>2.34</td>
<td>0.571</td>
<td>0.148</td>
</tr>
<tr>
<td>Jazz</td>
<td>198</td>
<td>2742</td>
<td>27.70</td>
<td>2.21</td>
<td>0.095</td>
<td>0.027</td>
</tr>
<tr>
<td>Netscience</td>
<td>379</td>
<td>914</td>
<td>4.82</td>
<td>1.98</td>
<td>0.113</td>
<td>0.142</td>
</tr>
<tr>
<td>Email</td>
<td>1133</td>
<td>5451</td>
<td>9.62</td>
<td>3.60</td>
<td>0.220</td>
<td>0.057</td>
</tr>
<tr>
<td>Facebook</td>
<td>4039</td>
<td>88234</td>
<td>43.69</td>
<td>3.69</td>
<td>0.606</td>
<td>0.010</td>
</tr>
<tr>
<td>Blogs</td>
<td>3982</td>
<td>6803</td>
<td>3.42</td>
<td>6.25</td>
<td>0.284</td>
<td>0.078</td>
</tr>
<tr>
<td>CA-HePh</td>
<td>12008</td>
<td>118521</td>
<td>19.74</td>
<td>5.21</td>
<td>0.611</td>
<td>0.008</td>
</tr>
<tr>
<td>PGP</td>
<td>10680</td>
<td>24316</td>
<td>4.55</td>
<td>7.48</td>
<td>0.266</td>
<td>0.056</td>
</tr>
<tr>
<td>Twitter</td>
<td>30173</td>
<td>137811</td>
<td>9.13</td>
<td>11.12</td>
<td>0.047</td>
<td>0.057</td>
</tr>
<tr>
<td>Epinions</td>
<td>75879</td>
<td>508837</td>
<td>13.41</td>
<td>4.75</td>
<td>0.138</td>
<td>0.005</td>
</tr>
<tr>
<td>Slashdot</td>
<td>77360</td>
<td>828161</td>
<td>23.41</td>
<td>4.11</td>
<td>0.056</td>
<td>0.004</td>
</tr>
</tbody>
</table>
3.3 Evaluation Criteria

All experiments conducted in this paper are on the PC with 8G of memory and Intel(R) Core(TM) i7-6500U CPU 2.50 GHz. The CPU time of 12 measures on 11 real networks is shown in Table 3.

Table 3. The CPU time (in seconds) of 12 measures on 11 real networks

<table>
<thead>
<tr>
<th>Network</th>
<th>pg</th>
<th>cc</th>
<th>bc</th>
<th>LC</th>
<th>LSC</th>
<th>LW</th>
<th>SEIC</th>
<th>LTC</th>
<th>CLC</th>
<th>THCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karate</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Jazz</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0</td>
<td>0.04</td>
<td>0.02</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>NetScience</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Email</td>
<td>0.2</td>
<td>2.4</td>
<td>7.3</td>
<td>0.1</td>
<td>0.17</td>
<td>0.1</td>
<td>0.24</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Facebook</td>
<td>3.3</td>
<td>94</td>
<td>198</td>
<td>2.4</td>
<td>6.64</td>
<td>3.2</td>
<td>13.78</td>
<td>2.56</td>
<td>6.4</td>
<td>6.3</td>
</tr>
<tr>
<td>Blogs</td>
<td>0.5</td>
<td>23.2</td>
<td>75.3</td>
<td>0.1</td>
<td>0.16</td>
<td>0.2</td>
<td>0.19</td>
<td>0.05</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>CA-HepPH</td>
<td>5</td>
<td>539</td>
<td>1675</td>
<td>4.7</td>
<td>11.82</td>
<td>5.1</td>
<td>33.5</td>
<td>5.7</td>
<td>11.0</td>
<td>11.3</td>
</tr>
<tr>
<td>PGP</td>
<td>1.4</td>
<td>205</td>
<td>739</td>
<td>0.2</td>
<td>0.72</td>
<td>0.3</td>
<td>0.97</td>
<td>0.2</td>
<td>0.6</td>
<td>0.9</td>
</tr>
<tr>
<td>Twitter</td>
<td>2.6</td>
<td>2046</td>
<td>8612</td>
<td>0.5</td>
<td>2.42</td>
<td>1.8</td>
<td>2.23</td>
<td>0.58</td>
<td>1.8</td>
<td>3.9</td>
</tr>
<tr>
<td>Epinions</td>
<td>8.4</td>
<td>8841</td>
<td>63883</td>
<td>8.8</td>
<td>39.66</td>
<td>16.2</td>
<td>175</td>
<td>7.94</td>
<td>16.8</td>
<td>29.3</td>
</tr>
<tr>
<td>Slashdot</td>
<td>12.8</td>
<td>26120</td>
<td>158430</td>
<td>22.8</td>
<td>69.25</td>
<td>36.2</td>
<td>1451</td>
<td>18.0</td>
<td>44.5</td>
<td>78.3</td>
</tr>
</tbody>
</table>

Kendall’s Tau Correlation Coefficient [31]

Kendall’s $\tau$ is defined as follows:

$$\tau(R_1, R_2) = \frac{N_c - N_d}{\frac{1}{2}N(N - 1)}$$

where $R_1$ and $R_2$ are two ranked lists that contain $N$ elements, respectively. $N_c$ and $N_d$ denote the amount of concordant and discordant pairs, respectively.

4 Experiments and Analyses

4.1 Computation Complexity

The CPU time of ten centrality measures on eleven networks is in Table 3. The CPU time of dc and ks is too little, all close to zero, so they are not included in the table.

4.2 Rank the Influence of Nodes

Rank the Influence of All Nodes

The kendall’a tau is to measure the consistency between the rank of nodes’ values calculated by certain centrality measure and the rank of nodes’ influence simulated by the SIR model.
We take the kendall’s tau, a method of quantitative analysis, to show the performance of different centrality measures on networks. As mentioned in the previous paragraph, the better a centrality is, the more correlative the values under certain centrality to the real influence. The kendall’s tau values results from the ranked values generated by twelve centrality measures and the ranked influence generated by SIR model on nine networks is show in Fig. 1.

The coefficient lines (hereafter called line) of local-based dc and ks usually have same trends, and perform mediocriely. Sometimes the coefficients of them are even below 0.5 (in jazz, netiscience and PGP).

In seven semi-local-based centrality measures, three of them (LW, LTC, THCC) consider the two-hop neighbors’ local structure and four of them (LTC, CLC, SEIC and LSC) consider the three-hop neighbors’ local structure. In eleven networks, only netiscience’s average distance is below 2, while others’ larger than 2. That’s why the semi-local-based centrality measures’ line in netiscience performs not well, except the LTC. LTC takes the number of mutual neighbors between node and its nearest

Fig. 1. Kendall’s τ results from the ranked values generated by twelve centrality measures and the ranked influence generated by SIR model on nine networks.
neighbors to adjust the nearest neighbors’ degree and sum up, which depend more on the nearest neighbors’ degree. In this case, LTC is less sensitive to the average distance of network than other six semi-local-based measures.

The lines of LSC perform very well in eight of nine networks, having high coefficients and low volatility. Only in the netscience, LSC performs bad, in which the coefficients are most below 0.6. The reason is that the netscience’s average distance is below 2 while LSC considers the second and third hop neighbors that nodes in netscience usually don’t have. Hence the average distance of network should be checked greater than 2 before applying the LSC.

CLC performs well, also better than LC, in eight of nine networks. CLC is actually the adaption of LC. Compared with LC, CLC takes the node’s clustering coefficient, which usually plays an negative impact on the information spreading, into consideration. Only in netscience, CLC is lower than LC.

Compared with LSC and CLC, LTC and THCC have more stable and close-to-best performance. In addition, LTC is less sensitive to the network’s average distance and can be applied to more networks.

The core ideas of SEIC and LW are similar, summing up the weights of a node’s edges as its centrality value. The difference is that the weights in LW is decided by the number of mutual neighbors of two neighboring nodes while the weights in SEIC is decided by the number of mutual two-hop neighbors of two neighboring nodes. SEIC considers the three-hop neighbors’ structure, but its line performs mediocly and even bad in facebook and blogs. LW performs even worse in most networks and the coefficients are all below 0.5 in these networks.

As global-based centrality measures, pg and bc perform almost the worst among all measures. The lines of pg and bc are usually under all other measures. To conclude, pg performs very well in hyperlink webpages, but is not practical in social networks.

**Rank the Most Influential Nodes**

In many practical applications, people are interested only in the most influential nodes in the networks. In this paper, we also focus on the performance of different centrality measures on the nodes whose influence ranks top 10% in the networks. The result is shown in Fig. 2. The topn_ratio list is [0.01, 0.02,…,0.09, 0.1].

Firstly, we analyze the large decrease of many measures in facebook when topn_ratio is larger than 0.5. We find that the average number of two-hop neighbors of the nodes, whose rank between topn_ratio[i] and topn_ratio[i + 1] of facebook, is [764.8, 756.25, 756.0, 756.65, 756.0, 941.601, 921.95, 946.37, 970.13, 965.95]. In this list, when topn_ratio is larger than 0.5, the number of two-hop neighbors of new nodes that taken into calculation grow abnormally. Thus centrality measures that have considered the impact of two-hop neighbors’ structure give a high value to these new nodes, resulting in the decrease in the line.

THCC performs the best in general. Its lines are the highest in six of eight networks and rank third in opinions and slashdot. In addition, THCC has low volatility and more stable than other measures.

LTC perform stably too, but its coefficients are usually 5%–10% lower than THCC.

CLC performs better than LC in most networks, but the line of CLC in facebook and CA-HepPH have large rise and falls and the lowest coefficient is even below 0.
Although the local clustering coefficient usually plays a negative role during information spreading, but it may not be suitable to the all kind of networks. Combining the result in facebook and CA-HepPH and the experimental result in [8], we find when the average clustering coefficient of the network in above 0.5, the kendall’s tau coefficient lines resulting from top 10% influential nodes’ ranked list generated by SIR model and their ranked list generated by CLC perform bad and have relative high volatility. SEIC and LW perform mediocly in most networks. The lines of SEIC have high volatility and the lines of LW in six of eight networks are under 0.4, which is bad.

As shown in the Fig. 2, classical centrality measures (dc, ks, pg, cc and bc) are not practical in these networks, pg and bc perform the worst.

We further analyze the average simulation influence of nodes that rank in the top 100 under twelve centrality measures on eight networks. Here we add three points. (1) THCC and LTC’s performance improve in epinions and slashdot, comparing favorably with LSC and even better in facebook and twitter. (2) Except facebook, CLC’s performance in other networks improve as well. (3) ks and dc perform best among classical centrality measures in identifying the 100 most influential nodes.

5 Conclusion and Future Work

In this paper, we apply the five classical and seven up-to-date centrality measures on eleven widely used networks. Three criteria (kendall’s tau correlation coefficient, distinguishability and robustness) have been taken into consideration to evaluate the performance of twelve measures. All these work is aimed to provide deeper understanding of various characteristics and best application scenarios of certain measure, and the conclusions.
For the purpose of influence maximization based on the influence ranking, we recommend to analyze the average degree, average distance and average clustering coefficient of networks before experiments. Combine the demands of efficiency and effectiveness, and determine the centrality measure suitable to your research.

For future work, we will simulate the influence of epinions and slashdot generated by SIR model on the whole range of epidemic probabilities and test the robustness of centrality measures on the epinions and slashdot networks. In future, we can test the measures on the diffusion data flows, the influence of which is based on the real information spreading instead of simulation by SIR. The application conditions of CLC and LSC in average clustering coefficient and average distance is based on the observation. For precise values more experiments need to conduct.

References

学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。

图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具