Big Data and Fans Economy Oriented Star Node Mining Improved Algorithm in Social Networks Based on Acquaintance Immunization

Jinbo Bai 1, Hongbo Li 2,3,* and Jianping Chen 2

1 Economics & Management College, Zhaoqing University, Zhaoqing, China
2 School of Computer Science and Software, Zhaoqing University, Zhaoqing, China
3 Department of Computer Science, Memorial University of Newfoundland, St. John’s, Canada
* Corresponding Author, Email: islhb@126.com

Abstract. With the booming and development of modern technologies in all kinds of fields, the scale of social networks is getting larger and larger, and it is more and more difficult to mine star nodes from this kind of network. On the basis of analyzing the relations among big data, fans economy and star nodes in social network in detail, aiming at the deficiencies of a star node mining algorithm proposed by us, a big data and fans economy oriented improved algorithm based on acquaintance immunization is put forward. Multiple contrast experiments on Karate club network and Dolphin social network demonstrate that the improved algorithm is superior to the old one from the quantity, degree and influence aspects of the mined star nodes.

1. Introduction

Big data is a terminology of voluminous or complex data sets that the traditional application softwares are inadequate to deal with or can’t process in a tolerable elapsed time [1-2]. It is generally believed that its characteristics, often simply named 5Vs, are as follows.

- Volume refers to the quantity of data is very big.
- Variety refers to the data are drew from multimedia, such as text, images, audio, video, etc.
- Velocity means that not only data but also results of analysis are available in a very short time.
- Value means the density of data value is comparatively extremely low.
- Veracity means the captured data are accuracy and reliable.

Unfortunately there are some controversies about the veracity characteristic. In recent years a lot of senior researchers, even some research communities, argue that there are many false data in reality. For example, it is found that there are voluminous faked transaction records on online e-commerce website. The purposes of some merchants to do so are simply to create a booming false impression to attract more customers to buy their products or services. According to the recent researchers’ views, the veracity characteristic should be excluded from the characteristics of big data.

With the booming of IT, the scales of social networks have become larger and larger, and the quantity of nodes has changed from several to millions. For example, there are only 34 and 62 nodes in the early classic social networks Karate club and Dolphin social respectively, and there are more than 300 million and 2 billion monthly active users in nowadays famous social networks twitter and Facebook respectively [3-4].

Besides the volume characteristic, data of nowadays social network are diverse. For example, Facebook user can send all kinds of multimedia messages, such as video, audio, picture, text, and so
on. In addition, nowadays social networks obviously have the value and velocity characteristics. So, nowadays social networks are big data sets in essence. There is a kind of networks in nowadays social networks, named fans networks. The main characteristics of this kind of networks are as follows.

- The nodes (vertexes) of the networks correspond to celebrities or stars, and their fans.
- The edges of the network are formed by interactions between fans and stars/celebrities.
- In general the node with bigger degree represents celebrities or stars, and has more influence on the neighbors (fans). We call them star nodes.
- A star node represents an opinion leader with Matthew effect, can be taken as a marketing object that can greatly reduce advertisement costs of many e-commerce enterprises [5].

The marketing operation by mining important nodes from fans networks is a novel economy form, named fans economy. Fans network has a great stickiness, and is a steady and huge profit resource. Therefore, the fans economy receives more and more attention. However, from the time cost aspect it is a severe challenge to mining star nodes in a large scale fans network.

We have put forward a star node mining algorithm based on acquaintance immunization in [5]. In this paper, orienting big data and fans economy an improved algorithm is proposed. Compared with the algorithm in [5], the contribution of the improved algorithm is that it not only inherits the high efficiency but also enhances the influence of the mined star nodes.

2. Related Works

2.1. Vital Nodes Mining

In the early developing phase of social network, researchers have proposed many measuring indices of vital nodes (sometimes named key nodes or important nodes), including degree centrality, closeness centrality, betweenness centrality [6], etc. However, the computing time cost of these indices is very high, reaches $O(n^3)$ in general, and it is unacceptable to apply them to a large scale social network.

In recent years researchers have proposed some new methods to measure vital nodes in social networks with large scales. For example, combining other works, Chen et al. proposed a measuring method through computing the numbers of a network’s spanning trees to redefine vital node concept [7-9]; based on the works of Ball, Page, Hsu and Wasserman; Tan et al. also proposed a measuring method through redefining the vital node concept, and believed that a vital node must can maximally increase a network’s agglomeration [10-13]; synthesizing the work of Zhao et al., Zhou et al. mined vital nodes by computing a contribution matrix [14-15]. Though most of the new methods can be applied to large scale social networks, liking the foregoing vital nodes mining methods by computing the centrality of degree, closeness and betweenness, they can’t mine star nodes in social networks.

In general, there are few efficient methods to mine star nodes in large scale social networks. Aiming at this problem we proposed an algorithm based on acquaintance immunization in [5]. The algorithm only uses local information of social networks, so it is very efficient and suitable to be applied to large scale networks.

2.2. Star Node Mining

2.2.1. Immunization strategy and its main methods in [5]. Immunization strategy is used to stop propagation of disease, rumor, virus, etc., in networks. Its fundamental is cutting off connections between infectious nodes and their neighbors, namely deleting the edges of infectious nodes and making them into isolated points. In essence, immunization means striving to destroy a network’s connectivity.

Immunization strategy derives from bionics, and has been applied to all kind of fields. In its development, many methods are put forward by researchers; there into random immunization, selected immunization and acquaintance immunization are more typical and more vigorous in the information technology field.

Random immunization firstly selects some nodes in a random way, then cuts off edges between them and their neighbours. For example, applying the method to Karate club network, as shown in
Figure 1 (a), and randomly selecting the nodes 25, 19, 12, 9 and 28, the result is as shown in Figure 1 (b).

Selected immunization firstly selects some nodes whose degrees are bigger than others, and then cuts off edges between them and their neighbours. For example, similarly applying it to Karate club network and selecting the nodes 34, 1, 33, 3 and 2, the result is as shown in Figure 1 (c).

Acquaintance immunization firstly selects some nodes in a random way, just liking random immunization, then finds the neighbours of these nodes, and cuts off edges between the neighbours and their neighbours. For example, applying the method to Karate club network, selecting the nodes 23, 21, 10, 8 and 19, finding their neighbours 33, 34, 3 and 1, the result is as shown in Figure 1 (d).

![Graphs showing different immunization methods applied to the Karate club network.](image)

**Figure 1.** Karate club network and results by applying immunization methods to it: (a) Karate club network, (b) effect of random immunization, (c) effect of selected immunization, (d) effect of acquaintance immunization [5].
In the foregoing examples, though the numbers of selected nodes are all the same, the destroyed degrees of networks are very different from each other. Apparently, selected immunization, acquaintance immunization and random immunization are in a decreasing order in terms of their destructive ability of network connectivity. Though the performance of acquaintance immunization is medium, the method only uses local information and is appropriate to large scale networks.

2.2.2. Star node mining algorithm in [5]. The node degrees of many networks conform to power law distribution [16]. According to this law, in a network a small degree node often has a big degree neighbor. Based on acquaintance immunization and taking advantage of this law, we proposed star node mining algorithm in [5].

There are two main steps in the star node mining algorithm in [5]. First, assigns a value to adjusting parameter \( r \) (the \( r \) is a percent), then randomly selects \( m = 1 + \text{floor}(n \times r) \) nodes into the initial set (the \( n \) is the node number of a network, and the floor is the top integral function). Second, finds the neighbors of the \( m \) nodes and puts them into the final set, namely star node set.

2.2.3. Deficiencies of algorithm in [5]. In the process of mining star nodes, the algorithm in [5] only take advantage of a social network’s local information. Comparing with degree centrality, closeness centrality, betweenness centrality algorithms, which have to use the global information, the time cost of the algorithm is low, only is no more than \( O(r \times n^2) \). So the algorithm is comparatively efficient and is appropriate for applying it to the large scale social networks. However, there are two main deficiencies in the algorithm.

First, the scope of influence maybe is comparatively small because that the algorithm does not prohibit the neighbors of the nodes being repeatedly selected in the initial set. For example, in Karate club network the initially selected nodes are 23, 21, 10, 8 and 19, and the nodes in the final set are 33, 34 and 3, as shown in Figure 2, where each of the final set nodes has a chance that be repeatedly selected, and definitely at least one node is selected twice according to the pigeonhole principle. In Figure 2, the background colors of the nodes in the initial set and the final set respectively are dark gray and light gray.

![Figure 2. Nodes of initial set and final set in Karate club network.](image)

Second, maybe there are non-star nodes in final set of the algorithm. According to the algorithm, the nodes in the final set are the neighbors of the nodes in the initial set. Obviously, some of the neighbors maybe non-star nodes, and their degrees are less than some nodes’ in the initial set. For example, under the assumption that we have put the nodes 57 of Dolphin social network into the initial set, and the neighbor node 62 is put into the final set, as shown in Figure 3, it is self-evidence that the node 62 is totally not a star node for its degree is even less than the degree of the initial node 57. In Figure 3, the back ground colors of nodes 57 and 62 are dark gray and light gray respectively.
The foregoing deficiencies restrict the application and development of the algorithm. So we propose an improved algorithm of it.

3. Improved Algorithm
There are still two main steps in the improved algorithm and the concrete contents are as follows.

Input: a network’s node set $G$, and adjusting parameter $r$ of initial selected node number.

Output: star node set $S$.

Step 1. Randomly selects $m=1+\lfloor n \times r \rfloor$ different nodes $v_1, v_2, \ldots, v_m$, and puts them into the initial set $I$.

\[
\text{for } i = 1 \text{ to } m \\
\quad \{ v = \text{randSelInitial}(G) ; \\
\quad \quad I = I + v_i ; \\
\quad \quad G = G - \{ v_i \} ; \\
\}\]

Step 2. Randomly selects a neighbor $v$ of each initial node $I_i$ coming from the initial set $I$, and according to certain conditions decides to push $I_i$ or $v$ into the star set $S$.

\[
\text{for } i = 1 \text{ to } m \\
\quad \{ v = \text{selNeighbor}(I) ; \\
\quad \quad I = I - \{ I_i \} ; \\
\quad \quad \text{if } (\text{isNew}(v) \&\& \text{deg}(v) > \text{deg}(I_i) ) \ S = S + \{ v \} ; \\
\quad \quad \text{else} \\
\quad \quad \quad \text{if } (\text{isNew}(I_i) ) S = S + \{ I_i \} ; \\
\}\]

In the step 1, the $n$ is the cardinal number of $G$; the function $\lfloor \rfloor$ takes a integer value which is no more than $n \times r$; the function $\text{randSelInitial}()$ randomly selects a new node for the initial set.

In the step 2, the function $\text{selNeighbor}()$ randomly selects a neighbor of the initial node $I_i$ coming from the initial set $I$, the function $\text{isNew}()$ judges a node whether is a new node of star set; the function $\text{deg}()$ computes the degree of a node. According to the improved algorithm, it is worth noting that a node maybe is not only an initial node but also a star node, namely a node maybe has a dual identity in some cases.
4. Contrast Experiments

4.1. About number of Star Nodes
We control the initial nodes are 23, 21, 10, 8 and 19, and apply the two algorithms to Karate club network and Dolphin social network. The results are shown in Figure 4-7. In the figures, the dark gray represents the initial nodes, the light gray represents the star node, and the grid dark gray represents dual identity node.

![Figure 4](image1.png)
**Figure 4.** About number of star nodes experiment applying the algorithm in [5] to Karate club network.

![Figure 5](image2.png)
**Figure 5.** About number of star nodes experiment applying the improved algorithm to Karate club network.

![Figure 6](image3.png)
**Figure 6.** About number of star nodes experiment applying the algorithm in [5] to Dolphin social network.
The experimental results demonstrate that the star nodes of the algorithm in [5] maybe are less than the initial nodes in number; nevertheless the star nodes of the improved algorithm in general are equal to the initial nodes in number. Though in some cases there are probabilities that the number of the star nodes of improved algorithm is less than the initial nodes, but they are very low. For example, in a network a node has only two leaf nodes, and they have been selected as initial nodes by lucky coincidence.

4.2. About Degrees of Initial Nodes and their Neighbor Nodes

We control the initial nodes are 3, 8, 19, 32 and 34, and apply the two algorithms to Karate club network. The results are shown in Figure 8-9. We take the nodes 2, 9, 25, 38, 39, 45 as the initial nodes, and apply the two algorithms to Dolphin social network. The results are shown in Figure 10-11. In the figures, the dark gray represents the initial nodes, the light gray represents the star node, and the grid dark gray represents dual identity node.

The experimental results demonstrate that the total degrees of the star nodes mining by the algorithm in [5] are obviously less than the total degrees of the initial nodes mining by the improved algorithm.
Figure 9. About degrees of initial nodes and star nodes experiment applying the improved algorithm to Karate club network.

Figure 10. About degrees of initial nodes and star nodes experiment applying the algorithm in [5] to Dolphin social network.

Figure 11. About degrees of initial nodes and star nodes experiment applying the improved algorithm to Dolphin social network.
4.3. About Influences of Mining Results of Two Algorithms

In this paper we take immunized edge ratio, simply IER, as an influence metrics of star node set. An IER can be computed through dividing the total edge number of a network by the immunized edge number. A bigger IER means that the nodes of a star node set connect more edges, and have more influence.

For comparison, we apply the two algorithms to Karate club and Dolphin social network 20 times, according to the Pareto principle [17] set the adjusting parameter $r$ is 0.2 (namely the number of initial nodes are 7 and 13 respectively), and compute the IER as seen in Table 1. It is worth noting that the numbers in Table 1 are all median excepting that the initial node numbers are fixed values.

| Network          | Algorithm applied       | Initial node number | Star node number | IER  |
|------------------|------------------------|---------------------|------------------|------|
| Karate club      | Algorithm in [5]        | 7                   | 5.8              | 0.51 |
| network          | Improved algorithm     | 7                   | 6.7              | 0.63 |
| Dolphin social   | Algorithm in [5]        | 13                  | 11.6             | 0.43 |
| network          | Improved algorithm     | 13                  | 12.4             | 0.54 |

Comparing with the algorithm in [5], the experimental results show that the star node sets mined by the improved algorithm get more influence: in Karate club network the IER gets a 12% increase from 0.51 to 0.63, and in Dolphin social network the IER gets an 11% increase from 0.43 to 0.54.

5. Conclusion

In the context of big data and fans economy, it is a severe challenge that mining star nodes from large scale networks. Aiming at the deficiencies of a star node mining algorithm, an improved algorithm based on acquaintance immunization is proposed. Compared with the old algorithm, under the same initial conditions, many times contrast experiments on Karate club network and Dolphin social network demonstrate that the star set mined by the new algorithm has bigger cardinal number and total degree, and possesses more influence.

6. Acknowledgments

The research is supported by Guangdong natural science foundation (2016A030313015), Zhaoqing Science and technology Innovation Guiding Category Project (201704030601), Innovation and School Developing Special Fund (504-20160171), and the corresponding author of this paper is Hongbo Li.

7. References

[1] https://en.wikipedia.org/wiki/Big_data
[2] https://en.wikipedia.org/wiki/Data_analysis
[3] https://en.wikipedia.org/wiki/Facebook
[4] https://en.wikipedia.org/wiki/Twitter
[5] Li H B, Bai J B, Zhang J P, Yang J and Zhao Y M 2017 Efficient Star Nodes Discovery Algorithm in Social Networks Based on Acquaintance Immunization ICIVC. 937
[6] Brandes U, Borgatti S P and Freeman L C 2016 Maintaining the quality of closeness and betweenness centrality Social Networks. 44 153
[7] Corley H and Sha D 1982 Most vital links and nodes in weighted network Operations Research Letters. 1 157
[8] Nardelli E, Proietti G and Widmayer P 2003 Finding the most vital node of a shortest path Theoretical Computer Science. 296 167
[9] Chen Y, Hu A Q and Hu X 2004 Evaluation method for node importance in communication networks Journal of China Institute of Communications. 25 129
[10] Ball M O, Golden B L and Vohra R V 1989 Finding the most vital arcs in a network Operations Research Letters. 8 73
[11] Page L B and Perry J E 1994 Reliability polynomials and link importance in networks IEEE Trans Reliability. 43 51
[12] Hsu L H, Jan R H, Lee Y C, Hung C N and Chern M S 1991 Finding the most vital edge with respect to minimum spanning tree in weighted graphs Information Processing Letters. 39 277
[13] Tan Y J, Wu J and Deng H Z 2006 Evaluation method for node importance based on node contraction in complex networks System Engineering-Theory & Practice. Nov. 79
[14] Zhao Y H, Wang Z L, Zheng J G and Xu J 2009 Finding the most vital node by node importance contribution matrix in communication networks Journal of Beijing University of Aeronautics and Astronautic. 35 1076
[15] Zhou X, Zhang F M, Li K W, Hui X B and Wu H S 2012 Finding vital node by node importance evaluation matrix in complex networks Acta Physica Sinica. 61 p050201
[16] Adamic L A and Huberman B A 2000 Power-law distribution of the world wide web Science. 287 2115
[17] Cato S 2016 Weak independence and the Pareto principle Social Choice and Welfare. 47 295