









Influence Maximization Algorithm in Social Networks Based on Three Degrees of Influence Rule

Hongbin Wang¹ , Guisheng Yin¹ , Lianke Zhou¹  ,
Xiaolong Chen¹ , and Dongjia Zhang² 

¹ College of Computer Science and Technology, Harbin Engineering University,
Harbin 150001, China

zhoulianke@hrbeu.edu.cn

² Aerospace Science and Industry Corp, Beijing 1100048, China

Abstract. Influence maximization algorithms in social networks are aimed at mining the most influential TOP-K nodes in the current social network, through which we will get the fastest spreading speed of information and the widest scope of influence by putting those nodes as initial active nodes and spreading them in a specific diffusion model. Nowadays, influence maximization algorithms in large-scale social networks are required to be of low time complexity and high accuracy, which are very hard to meet at the same time. The traditional Degree Centrality algorithm, despite of its simple structure and less complexity, has less satisfactory accuracy. The Closeness Centrality algorithm and the Betweenness Centrality algorithm are comparatively highly accurate having taken global metrics into consideration. However, their time complexity is also higher. Hence, a new algorithm based on Three Degrees of Influence Rule, namely, Linear-Decrescence Degree Centrality Algorithm, is proposed in this paper in order to meet the above two requirements for influence maximization algorithms in large-scale social networks. This algorithm, as a tradeoff between the low accuracy degree algorithm and other high time complexity algorithms, can meet the requirements of high accuracy and low time complexity at the same time.

Keywords: Social networks · Influence maximization
Three degrees of influence rule · Linear-Decrescence Degree Centrality

1 Research Background

The concept of “social network” [1] was first proposed by the British anthropologist Radcliffe Brown when he studied the social structure in the 1930s, which was used to describe all kinds of social relations among the organizations, individuals and between the organization and individuals. The main purpose of maximizing the influence of social network is to dig out the TOP-K nodes set with the greatest influence in the network through the existing social network relationship. And it has been widely applied in important fields such as marketing, disease prevention and rumor control.

For example, in the field of marketing, “viral marketing” [2–4] and “word of mouth effect” [5, 6] are the best applications for maximizing social network influence and social network communication models. Business companies always want to promote the newly developed products to the market at the lowest cost and to be accepted by the majority of the population. For this end, they will target a small number of “influential” users and present them the new product’s samples for its free trial first.

In the field of rumor control [8, 9], where people can freely talk and discuss about the national political and social hot events through various domestic social networks such as WeChat, Zhihu, microblogging and so on. However, while it allows people to express themselves freely, some unscrupulous criminals are trying to propagate rumors that violate the facts through these platforms, attempting to use these rumors to deceive and incite the masses to do something that endangers the country and society. Therefore, how to limit and control rumors into a small range in the early stages becomes also a major issue in social network analysis.

2 Research Status Quo

In recent years, the social network influence maximization algorithm has been widely focused and studied by researchers. In 2003, Kempe et al. [7] First demonstrated that the problem of maximizing influence was the NP-hard problem. Therefore, the heuristic algorithm and the greedy algorithm have become two major directions in solving the problem of influence maximization.

2.1 Heuristic Algorithms

The heuristic algorithm mainly considers the static structure characteristics of the social network such as the degree of nodes, the shortest path between nodes, network density, aggregation coefficients and betweenness.

Degree Centrality(DC) is the most direct and primitive measure of the centrality of nodes. It believes that there is a positive correlation between the degree and the influence of the node. It is shown in formula (1).

$$DC(u) = d(u) \quad (1)$$

Where $DC(n)$ represents the degree centrality of node u , and $d(u)$ represents its degree.

Betweenness Centrality (BC), is the betweenness of nodes. This algorithm believes that the larger the betweenness of nodes, that is, the larger of the number of nodes that are in the critical path (shortest path) between nodes in the social network, the greater its control over the communication between all nodes. It is shown in formula (2).

$$BC(u) = \sum_{i \in V} \sum_{j \neq i \in V} \frac{\delta_{ij}(u)}{\delta_{ij}} \quad (2)$$

Where $\delta(u)$ represents the betweenness of node u , δ_{ij} represents the number of shortest paths between nodes i and j , and $\delta_{ij}(u)$ indicates that there are $\delta_{ij}(u)$ shortest paths between the two nodes that passes node u .

Closeness Centrality (CC), which reflects the distance of nodes from the other nodes in the social network. The basic idea of this algorithm is that the smaller the cost of the node to communicate with the rest of the nodes, and the more important the node's position is in the network, that is, the greater the influence and is shown in formula (3).

$$CC(u) = \frac{n - 1}{\sum_{v \neq u, v \in V} dis(u, v)} \tag{3}$$

Where $CC(u)$ represents the Closeness Centrality of the node, n represents the number of nodes in the network, and $dis(u, v)$ represents the shortest path from node u to node v .

Since the DC algorithm is low in complexity but also low in accuracy and the BC and CC algorithms are of high precision but are highly complex, Professor Chen et al. [8] have put forward the Local Centrality algorithm (LC) as a compromise solution. The algorithm not only considers the nearest neighbor nodes of a node, but also takes the neighbor nodes of its neighbor nodes into account. It is shown in formula (4) and formula (5).

$$LC(u) = \sum_{v \in \Gamma(u)} Q(v) \tag{4}$$

$$Q(v) = \sum_{w \in \Gamma(v)} N(w) \tag{5}$$

Where $\Gamma(u)$ represents the nearest neighbor nodes set of node u , and $N(w)$ represents the number of its neighbor nodes and the number of neighbor nodes of its neighbor nodes. Thus, by definition, the LC algorithm, when estimating the influence, assumes that the influence can be propagated by the original activated node u through node v , node w , and node x to node z after four degrees, as shown in Fig. 1.

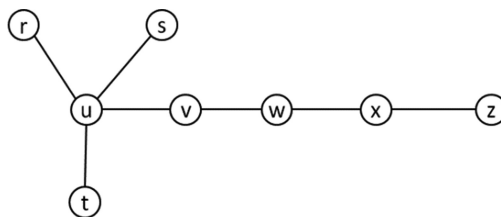


Fig. 1. Range of influence of LC algorithm

2.2 Greedy Algorithm

The basic research of the greedy algorithm is focused on the hill-climbing greedy algorithm, which selects the node with the greatest influence each time to calculate the local optimal solution. After several iterations, the global optimal solution is obtained. The biggest advantage of this algorithm is that it has high accuracy and can achieve the optimal approximation of $1 - 1 / e - \epsilon$. However, the disadvantage is that the algorithm is highly complex, needing long running time, which cannot meet the fast-computation requirement of large-scale social network.

Three Degrees of Influence Rule. Nicolas Kristakis, Harvard University Sociology Professor and James Fuller, the philosophy doctor of University of California, San Diego, have put forward three degrees of influence rule [9] in their work on the social network- “Large Connection: How social networks are formed and the impact on the behavior of the human reality” for the first time. Dr. James Fuller pointed out that our behaviors, attitudes and so on follow the so-called “three degrees of influence rule” in terms of its spread in the real life: these acts, attitudes will not only cause influence in the circle of our friends (first degree), but also in the second degree distance (friends of friends) and in the third degree distance (friends of friends’ friends).

The algorithm Ideas of Linear-Decrescence Degree Centrality. As stated by three degrees of influence rule above, the spread of influence in the social network is not boundless, but limited to within three degrees, and the influence is almost zero beyond the degree. In addition, even if the influence can spread within three degrees, the size is not fixed, but gradually decay. It is based on the above two theories that this paper proposes a new algorithm to maximize the influence of social network, that is, Linear-Decrescence Degree Centrality algorithm. The formula for Linear-Decrescence Degree Centrality algorithm is shown in formula (6) and formula (7).

$$LDDC(n) = |F(n)| + \alpha(|S(n)| + \beta|T(n)|) \quad (6)$$

$$F(n) \cap S(n) \cap T(n) = \emptyset \quad (7)$$

where α denotes the decrescence coefficient when the influence of node n propagates from the nodes within the first degree to the second degree (i.e., the neighbor node of n 's neighbor nodes), and β means the decrescence coefficient when the influence of node n propagates from the nodes within the second degree to the third degree (i.e., the neighbor node of the neighbor node of n 's neighbor nodes), and $0 \leq \alpha, \beta \leq 1$. The pseudo-code description of Linear-Decrescence Degree Centrality algorithm is shown in Algorithm 1.

The pseudo-code description of LDDC algorithm

Algorithm 1: LDDC Algorithm

```

Input( G(V, E) )
Output (the LDDC(i) value of each node I in the network )
1: for i = 1 to n
2:   F(i) = {the neighbor nodes of i }
3: end for
4: for i = 1 to n do
5:   for j in F(i)
6:     k = j's neighbor nodes
7:     if k ∉ F(i) && k ≠ i
8:       S(i) = S(i) ∪ {j}
9:   end for
10: end for
11: for i = 1 to n do
12:   for j in S(i)
13:     k = j's neighbor nodes
14:     if k ∉ F(i) && k ∉ S(i) && k ≠ i
15:       T(i) = T(i) ∪ {j}
16:   end for
17: end for
18: for i = 1 to n do
19:   LDDC(i) = |F(i)| + α(|S(i)| + β|T(i)|)
20: end for
    
```

For example, in Fig. 2, if we are going to solve the LDDC of node 10 and let's assume that $\alpha = 0.7$ and $\beta = 0.4$. In the neighbor node of node 10 the first degree are 6, 11, 23, namely, $F(n) = \{6, 11, 23\}$, $|F(n)| = 3$; the neighbor nodes of node 10 the second degree are 1, 12, 14, 20, 21, 22, that is, $S(n) = \{1, 12, 14, 20, 21, 22\}$, $|S(n)| = 6$; the neighbor nodes of node 10 the third degree are 2, 3, 4, 5, 7, 8, 9, 13, 15, 16, 17, 18, 19, i.e., $T(n) = \{2, 3, 4, 5, 7, 8, 9, 13, 15, 16, 17, 18, 19\}$, $|T(n)| = 13$. Although node 23 belongs to both the first degree node set and the second degree node set of node 10, it belongs to $F(n)$ only after processing.

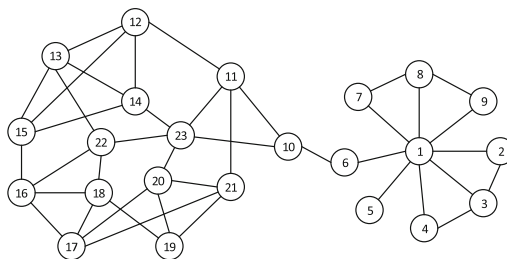


Fig. 2. Example graph of a social network

3 Experimental Results and Analysis

In this section, the true influence of the nodes in each data set is simulated and used as the standard set, the approximant influence estimated through the LDDC algorithm is used as the experimental set, and the approximation of influence through DC algorithm, BC algorithm, CC algorithm, LC algorithm is used as the contrast set. It should be noted that, in calculating the influence value through LDDC algorithm, it is found that when $\alpha = 0.7 \pm 0.05$ and $\beta = 0.4 \pm 0.05$, through a large number of experimental calculations and summary, there is less variation in the experimental results, and the experiment achieves its best effect. Therefore, in this paper, all calculation of influence involved in the LDDC algorithm, the unified value has been taken, that is, $\alpha = 0.7$, $\beta = 0.4$.

3.1 Selection of Data Sets

The three data sets are either non-directional networks or have been processed as non-directional networks. The static structural features of each data set are shown in Table 1.

Table 1. The static structural features of experimental data sets

Statistics items	Dolphin	Email	Blog
Number of nodes (n)	62	1133	3982
Number of edges (m)	159	5451	6803
Average degree ($\langle k \rangle$)	5.129	9.622	3.417
Maximum degree (kmax)	12	71	189
Average path length	3.111	3.606	6.252
Network density	0.084	0.009	0.001
Focusing factor	0.290	0.254	0.493

The dolphin dataset is a social network dataset that records the daily contact between a group of dolphins living in a community in Doubtful Sound, New Zealand. The email dataset is a social network dataset that records email communications between members of University of Rocía de Villegill, Spain. The blog dataset is a social network dataset of community relationships between bloggers on the MSN (Windows Live) Spaces website.

3.2 Simulation of True Value of Nodes' Influence

In this paper, the SIR epidemic model is used to conduct the influence spread. The true value of influence of the node is simulated and used as a standard to measure the effectiveness of the LDDC algorithm.

In the SIR model [10], the population is divided into three categories: the susceptible, the infected and the immunized. In the process of obtaining the true value of

influence of the node, only one node is set as the initial infected node. $F(t)$ is the total number of infected nodes and immune nodes at time t , and $F(t)$ is used to evaluate influence of the initial infected node at time t . Obviously, $F(t)$ increases as time t grows, and finally stabilizes, that is, there is no infected node in the final network, then $F(t)$ is expressed as $F(tc)$. Therefore, $F(tc)$, as a single initial infection node influence, the greater it is, the greater the impact of the node. Finally, we can use the $F(tc)$ value of the node to simulate the approximant true influence of the node. Considering that in practical applications, the propagation of influence is more important in the early stages, $F(t)$ at $t = 10$ is used instead of $F(tc)$ in the final steady state. If there is no special explanation, $F(t)$ refers to the total number of infected nodes and immuned nodes at $t = 10$.

3.3 Results and Analysis

Position Offset Method. The so-called position offset method refers to the method that first sorts $F(tc)$ value obtained by propagating the single node through the SIR model in descending order, using it as the measure standard, namely, the true value influence of the node, and the sorted results of $F(tc)$ represent the rank of true influence), then calculates position offset between the sorting positions derived from the algorithm proposed in this paper as well as other centrality algorithms and the measure standard. The smaller the position offset, the closer to the true sorting results, which means the more accurate the algorithm.

Table 2. The TOP-10 sorting centrality value of the centrality algorithms in the dolphin data set

TOP-K	DC	CC	BC	LC	LDDC	F(t)
1	12	26.77	422.88	2232	34.62	39.83
2	11	26.02	404.50	2100	33.4	39.64
3	11	25.49	218.46	2052	32.76	39.51
4	10	25.49	208.36	1847	32.64	39.18
5	9	24.97	184.90	1706	31.64	38.74
6	9	24.32	173.69	1675	30.84	38.72
7	9	23.85	171.24	1582	30.62	38.26
8	9	23.70	164.65	1541	28.98	38.06
9	9	23.55	159.71	1538	28.18	37.91
10	9	23.40	141.52	1521	27.94	37.67

Taking the dolphin data set as an example, the $F(t)$ value of the TOP-10 nodes and the corresponding node names sorted by the $F(t)$ value at $t = 10$ are shown in Tables 2 and 3 respectively. For example, the $F(t)$ value of the node 33 is the largest, ranking the first in the LDDC algorithm, and the offset position is $OP_{33} = |1 - 1| = 0$; in the LC algorithm it ranks the 4th, and the position offset is $OP_{33} = |1 - 4| = 3$. Then it can be concluded that the LDDC algorithm is more accurate for node 33. Furthermore, let's

Table 3. The TOP-10 nodes ranking of the centrality algorithms in the dolphin data set

TOP-K	DC	CC	BC	LC	LDDC	F(t)
1	14	33	33	37	33	33
2	37	37	13	45	37	37
3	45	40	36	14	36	40
4	51	36	1	33	40	36
5	13	50	17	50	50	14
6	17	20	40	40	20	20
7	20	43	37	16	14	50
8	29	14	14	21	43	45
9	33	13	51	18	13	16
10	57	21	43	20	45	38

Table 4. The calculation of position offset in the dolphin dataset

	TOP-10	TOP-20	TOP-30	TOP-40	TOP-50
DC	70	145	228	329	394
CC	16	45	84	102	147
BC	75	156	246	341	405
LC	37	101	167	210	244
LDDC	14	47	81	95	124

take a look at node 36, whose $F(t)$ value ranks the 4th. It ranks the 3rd in the LDDC algorithm, and the position offset is $OP_{36} = |4 - 3| = 1$; it ranks the 4th in CC algorithm, and its position offset is $OP_{36} = |4 - 4| = 0$, so the CC algorithm is more accurate for node 36. Then, for each centrality algorithm, the position offset is obtained for each node, and the total position offset of the algorithm is obtained. The offset position of the TOP-10 nodes is compared with the $F(t)$ result, shown in formula (8).

$$PO_{TOP-K} = \sum_{u \in TOP-K} PO_u \tag{8}$$

Table 4 is the calculated value of the position offset of the node set of the centrality algorithm TOP-10 ~ TOP50, using the ranking of results, $F(t)$, in the SIR propagation model in the dolphin social network as the measure standard. Figure 3 is the corresponding line graph.

The line graphs of position offset in the email data set, and blog data set are shown in Figs. 4 and 5.

An observation of Figs. 3, 4 and 5 shows that position offset values of LDDC algorithm and CC algorithm are smaller and closer, indicating that the influence ranking of LDDC algorithm and CC algorithm are closer to the real influence ranking; and position offset values of LC Algorithm, BC algorithm and DC algorithm are larger, indicating that there is a larger difference in the estimated influence rank and the rank of

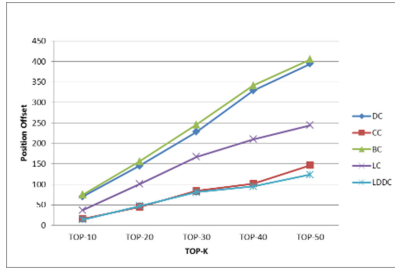


Fig. 3. The line graph of position offset in the dolphin dataset

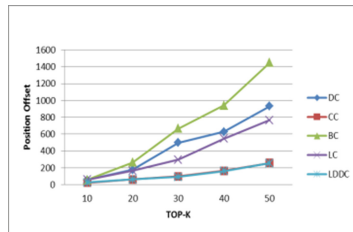


Fig. 4. The line graph of position offset in the email data set

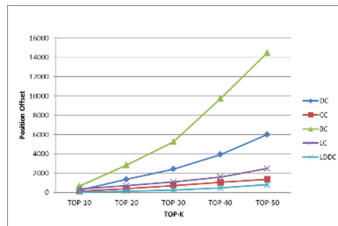


Fig. 5. The line graph of position offset in the blog data set

true influence. Therefore, according to the position offset method, the accuracy of these centrality algorithms can be ranked from high to low as: LDDC algorithm > CC algorithm > LC algorithm > DC algorithm \approx BC algorithm (Fig. 6).

It is shown that the $F(t)$ of the node increases gradually with the increase of the estimated influence of the LDDC algorithm and the CC algorithm, displaying a strong positive correlation in the dolphin data set. DC algorithm and LC algorithm also show a certain positive correlation, but it is slightly weaker; and BC algorithm’s correlation is the weakest.

An observation of Fig. 7: the scatterplot of correlation for the email dataset, shows that the LC algorithm has the strongest positive correlation, followed by the LDDC algorithm and the CC algorithm, and the positive correlation of the DC algorithm and the BC algorithm is relatively weak.

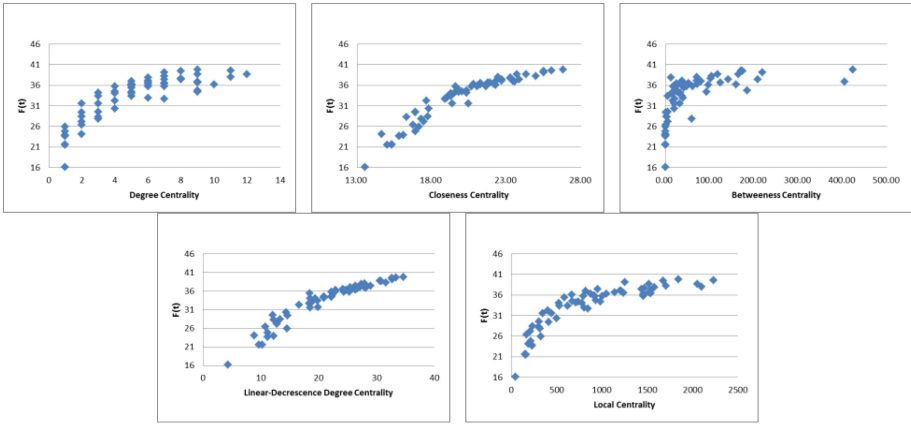


Fig. 6. The scatterplot of correlation for the dolphin dataset

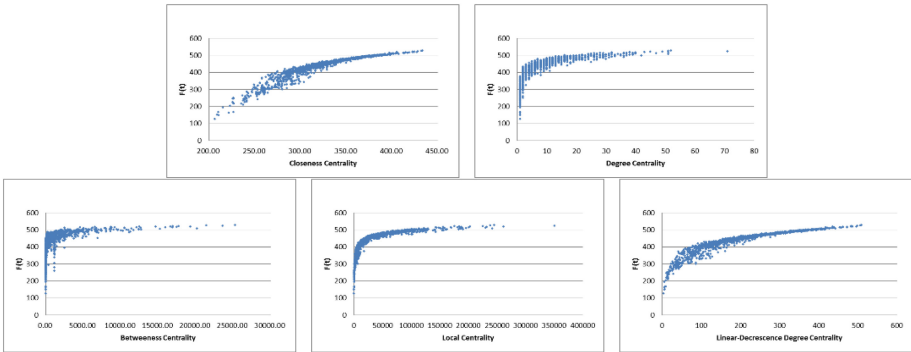


Fig. 7. The scatterplot of correlation for the email dataset

An observation of Fig. 8: the scatterplot of correlation for the blog dataset, shows that the positive correlation of the LDDC algorithm and the CC algorithm is strong, and that of the DC algorithm, LC algorithm and BC algorithm between $F(t)$ is weak.

TOP-K Difference Method. Taking the LDDC algorithm and the LC algorithm of the dolphin dataset as an example, the difference between the TOP-10 nodes is shown in Table 5, where nodes of the LC algorithm which are not included in TOP-10 nodes of the LDDC algorithm are 16, 21, 18, so we use them as the initial set of activated nodes, propagating through the SIR propagation model, and observe the final influence range $F(t)$. Similarly, the LDDC algorithm uses 36, 43, and 13 as the initial set of active nodes. Then the propagation results of LC algorithm and the LDDC algorithm in the dolphin data set within $t = 10$ are shown in Fig. 9.

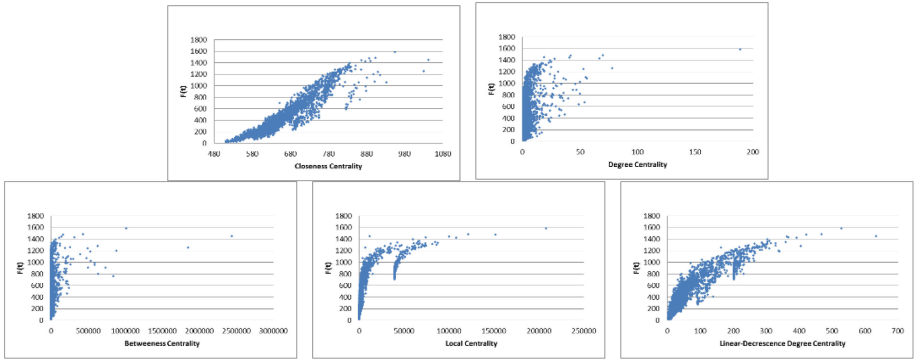


Fig. 8. The scatterplot of correlation for the blog dataset

Table 5. TOP-10 differences between the LDDC algorithm and LC algorithm in the dolphin data set

Algorithm name	LC	LDDC
TOP-10 node name	37	33
	45	37
	14	36
	33	40
	50	50
	40	20
	16	14
	21	43
	18	13
	20	45
The name of different nodes in TOP-10	16,21,18	36,43,13

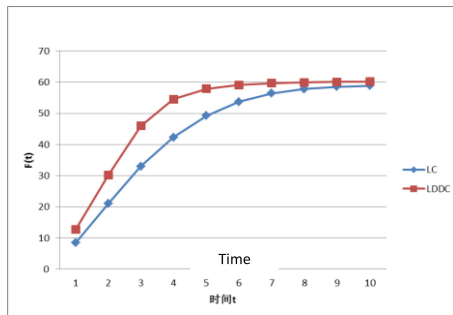


Fig. 9. TOP-10 differences between the LDDC algorithm and LC algorithm in the dolphin data set

4 Conclusion

This paper proposes a new algorithm of maximizing the influence based on Three Degrees of Influence Rule, that is, the linear-crescence degree centrality algorithm. This algorithm, as a tradeoff between the low accuracy degree algorithm and other high time complexity algorithms, can meet the requirements of high accuracy and low time complexity at the same time. The algorithm estimates the influence of the node in the global network by looking at the potential influence of the node within three degrees, taking into account the law that the influence of the node gradually decays when the influence spreads outwardly. The algorithm is simple in calculation and low in time complexity, with a short running time and high precision. It has good scalability even in the large-scale social network. Lastly, the validity of the linear attenuation algorithm is verified by the method of position offset method, correlation method and TOP-K difference method.

Acknowledgments. This work was funded by the National Natural Science Foundation of China under Grant (No. 61772152 and No. 61502037), the Basic Research Project (No. JCKY2016206B001, JCKY2014206C002 and JCKY2017604C010), and the Technical Foundation Project (No. JSQB2017206C002).

References

1. Guo, K.H.: Application of reliability assessment methods of small sample based on Bayes theory. *Sci. Tech. Innov. Prod.* (2014)
2. Liu, Q.: Study on the influence of prior reliability for bayesian estimation. In: *International Conference on Reliability, Maintainability and Safety*, pp. 407–410 (1999)
3. Maeda, T., Kimura, M.: A Study on Linearized Growth Curve Models for Software Reliability Data Analysis. *Ieice Technical Report 106*, pp. 21–26 (2006)
4. Yang, R., Tang, J., Sun, D.: Association rule data mining applications for Atlantic tropical cyclone intensity changes. *Weather Forecast.* **26**, 337–353 (2010)
5. Norman, A.T., Russell, C.A.: The pass-along effect: investigating word-of-mouth effects on online survey procedures. *J. Comput. Mediat. Commun.* **11**, 1085–1103 (2006)
6. File, K.M., Cermak, D.S.P., Prince, R.A.: Word-of-mouth effects in professional services buyer behaviour. *Serv. Ind. J.* **14**, 301–314 (1994)
7. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146 (2003)
8. Chen, D., Shang, M., Lv, Z., Fu, Y.: Detecting overlapping communities of weighted networks via a local algorithm. *Phys. A* **389**, 4177–4187 (2010)
9. Vanderweele, T.J.: Inference for influence over multiple degrees of separation on a social network. *Stat. Med.* **32**, 591 (2013)
10. Balkew, T., Model, S., Rate, R.R., Analysis, P.P.: The SIR Model When S(t) is a Multi-Exponential Function. *Dissertations & Theses - Gradworks 14*, 50-50 (2010)