Identify the Influential Spreaders in the Social Network

Xiaotong Diao, Ziyuan Dong

December 16, 2018

1 Introduction

Information Spreads (also known as Information diffusion, cascade) has attracted tremendous attention within the community of network science due to its potential applications in various disciplines[1]. Influential individuals usually have a main influence on a social network. And one of the crucial and fundamental issues is to efficiently identify multiple individuals that have the most influence in a large-scale complex networks, which can help us control the spreading process[2]. For example, we could stop the spreading of epidemics and rumors by controlling specific influential spreaders or deliver new products to a small group of initial users who have the maximal advertising effectiveness.Theoretically, it's an influence maximization problem.

In our project[1], we take the Gnutella peer to peer network for example and analyze the structure properties of social network. We find that social networks are scale-free networks. Intuitively, we think that a node with a higher degree should be more influential. However, we find that social networks conform to the core-periphery model, which means that nodes with high degrees are relatively close to each other. Therefore, the spreading ranges of high-degree nodes may overlap with each other, which leads to a waste of spreading effectiveness. So the states of neighbours should be consider when selecting a node as a new spreader. Based on the above considerations, We implement 4 heuristic methods to calculate the the potential influence of the nodes in Gnutella network. In these four methods, a node with too many infected neighbours won't be selected as a spreader even if its degree is high. To simulate the real situation, we also improve the SIR model. We specify that each infected node can only contact with one neighbour randomly in each iteration. And a node can't contact another node again if they have contacted before. We experimentally observe that the $GeneralDegreeDiscountAlgorithm$ method has the best effectiveness with different fractions of initial spreaders. Then we changed the transmission probability between each pair of nodes in term of different frequencies of connections. The result demonstrates that the effectiveness of $GeneralDegreeDiscount$ method is still the best in this case.

[1]Code link: https://github.com/DongZiyuan/Network-Analysis-Project

2 Dataset Analysis

Limited by the performance of our the computers, it is best to select a real social network with thousands of nodes for experiments. So we use the Gnutella Peer-to-peer Network [2]. It is a sequence of snapshots of the Gnutella peer-to-peer file sharing network from August 2002. There are total of 9 snapshots of Gnutella network collected in August 2002. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts. To simplify our task, we ignore the directions and the weights of its edges. Besides, the topological structures of these 9 snapshots are almost identical, although they have different numbers of nodes and edges. Taking the snapshot captured on August 8 for example, the Gnutella network of that time has 6301 nodes and 20777 edges. More basic statistics of is shown in Fig 1. The distribution of degree at that time is shown in Fig 2. As is shown, the degree distribution of a social network isn't uniform or normal. Some nodes have quite high degrees, but the others don't.

nodes	6301
edges	20777
number of triangles	2383
average clustering coefficient	0.0109
fraction of closed triangles	0.006983
diameter	9
number of components	1
is weighted	no
is directed	no

Figure 1: Network statistics

Moreover, we use the greedy modularity maximization to detect the communities in Gnutella network. As shown in Fig 3, nodes with different degrees are in different colors. Obviously, nodes with higher degrees gather in the center and upper left corner of the figure, which can be seen as a kind of core-periphery network. Fig 4 shows that there are 24 communities in total and the sizes of these communities very different.

3 Heuristic Methods

It is well known that a node with a higher degree can infected more nodes. However, some recent studies argued that node with the highest degree might not always be the most influential one. Though the effectiveness of the selecting spreaders with the highest degree

[2] http://snap.stanford.edu/data/p2p-Gnutella04.html

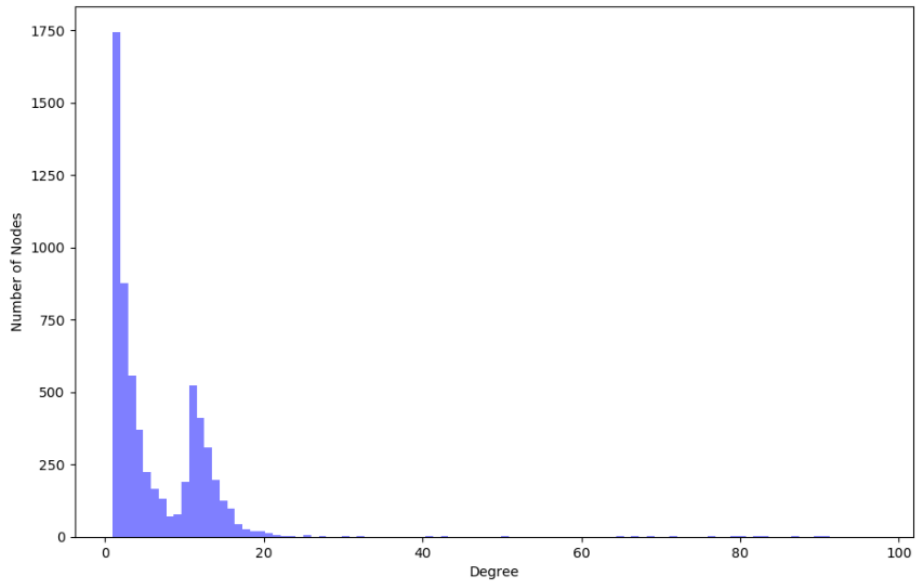

Figure 2: Degree distribution

is questionable, the low computational complexity of this strategy results in its widespread use in many fields. In this section, we implement 4 heuristic methods with good performance and better effectiveness.

At first, let us clarify what the wrong is with selecting spreaders with higher degrees. Let u and v be two adjacent nodes. Suppose u is already a spreader and we want to select v as a new spreader now. The existence of u weakens the potential influence of v in two ways: (i) it is no longer necessary for v to infect u. (ii) u may infect v with some probability later.

3.1 Degree Distance

Degree Distance[3] takes a naive approach to avoid the relative influence between adjacent nodes. Its idea is to try to make the spreaders disperse. So if a node is selected as a spreader, the other nodes within a certain distance to this node won't be considered any more. In particular, Degree Distance defines a candidate set C and a distance threshold d_t. At first, all the nodes are in the candidate set C. In each selecting round, a node with the maximum degree in C is selected as a spreader, and the other nodes within a distance d_t to it are removed from C. The procedure ends when all the spreaders are already selected. Empirically, we let $d_t = 3$ in our project.

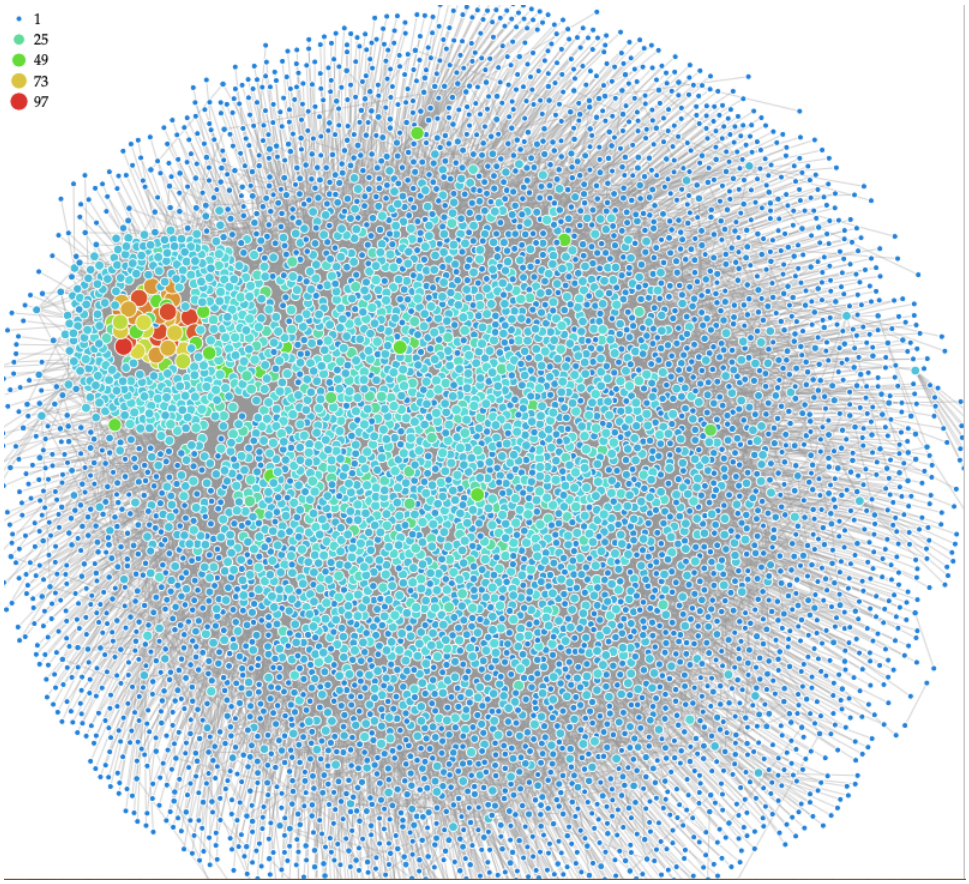

Figure 3: Graph of the Gnutella network

3.2 Single Discount

Since the potential influence of a node can be reduced by its infected neighbours, we should ignore the edges between this node with its infected neighbours when calculating the degree of this node. In other words, for a susceptible node, we should recalculate its degree by subtracting the number of its infected neighbours from its degree. Let sd_v be this "recalculated degree" of v, d_v be the original degree of v, t_v be the infected neighbours of v. So the Single Discount can be defined as:

$$sd_v = d_v - t_v \tag{1}$$

At first, all the nodes are susceptible and the sd of each node is equal to its degree. Then we select the node with the maximum sd(that is the node with the highest degree). Next, we update the sd of the other nodes. Repeat the preceding steps until the number of spreaders meet our requirement.

community	1	2	3	4	5	6	7
number of nodes	1280	1057	771	626	611	603	241
community	9	10	11	12	13	14	15
number of nodes	183	134	113	110	82	81	68
community	17	18	19	20	21	22	23
number of nodes	33	11	9	9	9	8	8

Figure 4: Sizes of 24 communities

3.3 Degree Discount

Degree Discount[4] is specifically designed for the independent cascade model. In this method, we calculate the expected number of neighbours that a node can infect when being selected as a new spreader. And we consider this expected number as the potential influence of this node. Let the transmission probability be p. Let t_v, d_v be the number of infected neighbours of v and the degree of v respectively. When p is small, we can ignore the states of the non-adjacent nodes of v and just consider the expected number of neighbors that can be infected directly by v. The number of susceptible neighbors of v is $d_v - t_v$. The probability that v is not influenced by its infected neighbors is $(1-p)^{t_v}$. In this situation, the expected number of nodes that may infected by v directly is $1 + (d_v - t_v)p$, including v itself. So the expected number of nodes that directly influenced by v is

$$(1-p)^{t_v}\left[1 + (d_v - t_v)\right] \tag{2}$$

Under the first order of Taylor expansion, when p is small, the left term can be approximated by $1 - t_v p + o(t_v p)$. After further simplification, the whole expression becomes

$$1 + \left[d_v - 2t_v - (d_v - t_v)t_v p\right] p + o(t_v p) \tag{3}$$

Then, the discount degree of v can be defined as

$$dd_v = d_v - 2t_v - (d_v - t_v)t_v p \tag{4}$$

Finally, do the same steps as the Single Discount does to select spreaders. Note that in the original Eq(1), dd_v is always non-negative. However, in the simplified form Eq(4), dd_v may be negative with some special parameters. In this situation, we manually set dd_v to be 0.

3.4 Generalized Degree Discount

Degree Discount ignores the differences of the neighbors of v. However the neighbors of v shouldn't be treated equally. For example, in Fig 5, the infected nodes are colored gray,

and s, t are both neighbors of v. Apparently, s have more infected neighbors than t. So the probability that s is influenced by its own neighbors is far larger than t, which weakens the expected contribution of s to the potential influence of v. So When calculating the contribution of s and t towards v, the latter one should be given more weight. In a word, Generalized Degree Discount takes the states of the neighbors of neighbors of v into consideration, not just the adjacent nodes of v. Similar to the analysis in Degree Discount,

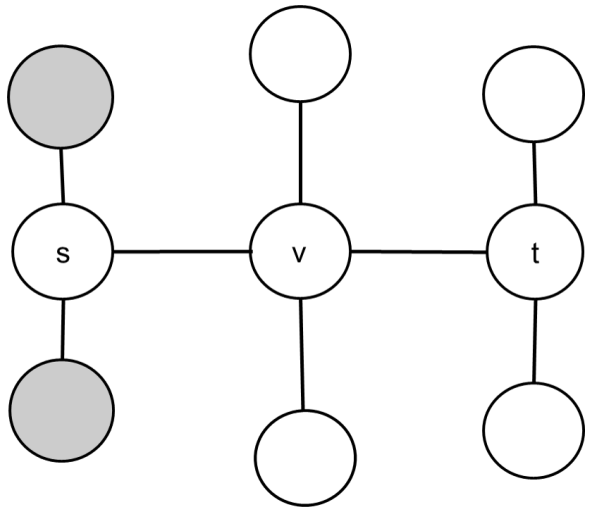

Figure 5: Generalized degree discount example

the probability that v is not influenced by its infected neighbors is $(1-p)^{t_v}$ and the number of nodes that may be infected directly by v is $1+(d_v-t_v)$, including v itself. For v' any susceptible neighbor w, the probability that w is not influenced by its own neighbors is $(1-p)^{t_w}$. So the probability that w is infected by v is $p(1-p)^{t_w}$. Together, the expected number of nodes that will be influenced by v is

$$(1-p)^{t_v}\left[1+\sum_{d_v-t_v}(1-p)^{t_w}p\right] \tag{5}$$

Departing from the conduction in Degree Discount, here we consider the second-order Taylor expansion for the left term and the first-order expansion for the right term:

$$\left\{1-t_vp+\frac{1}{2}t_v(t_v-1)p^2+o\left[(t_vp)^2\right]\right\}\left\{1+\sum_{d_v-t_v}\left[1-t_wp+o(t_wp)p\right]\right\} \tag{6}$$

After further simplification, the expression becomes

$$1+\left[d_v-2t_v-(d_v-t_v)t_vp+\frac{1}{2}t_v(t_v-1)p\sum_{d_v-t_v}t_sp\right]p+o(t_v^2p^2)+\sum_{d_v-t_v}o(t_vp^2) \tag{7}$$

Then, the generalized discounted degree[5] of v can be defined as

$$gdd_v = d_v - 2t_v - (d_v - t_v)t_v p + \frac{1}{2}t_v(t_v - 1)p \sum_{d_v - t_v} t_s \tag{8}$$

Similar to the situation in Degree Discount, the simplified equation of gdd_v may also be negative. In our real implementations, we set $gdd_v = 0$ in this situation. Finally, we select the node with the maximum ggd_v and update gdd_v of the others until the number of spreaders meet our requirement.

Compared to the formulation of Degree Discount(Eq(4)), the formulation of Generalized Degree Discount(Eq(8)) adds the last two terms. As the consideration of the latter one is deeper than the former one, Generalized Degree Discount should be more effective than Degree Discount. In reality, however, the difference of performance between Degree Discount and Generalized Degree Discount is not significant. Because the number of spreaders is not large, usually $t_v \ll d_v$ for all nodes. Thus, in Eq(8), the fourth and fifth terms are smaller than the third term, which makes Generalized Degree Discount just similar to Degree Discount.

3.5 Computational Complexity

If we want to select l spreaders, the algorithm must run for l rounds. Let N be the number of nodes and $\langle k \rangle$ be the mean degree. In each selecting round, the selection scheme costs $O(N)$, the neighbor finding scheme costs $O(\langle k \rangle^2)$, and for each of those neighbors, the updating process costs $O(\langle k \rangle)$. Then, the total time cost of the algorithm is $O(l(N + \langle k \rangle^2 + \langle k \rangle^3)) \approx O(l(N + \langle k \rangle^3))$. In many networks, the average degree is far less than the number of nodes: $\langle k \rangle \ll N$. Thus the time cost of Generalized Degree Discount will be nearly $O(lN)$, which is just linearly correlated with the scale of the network.

4 Testing Method

4.1 Improved SIR Model

We use SIR model to simulate the real spreading process and observe the performance of each algorithm. Because the epidemic transmission and the spread of information are similar. Specifically, susceptible state S denotes the individuals who are not aware of the information. Infected state I can be analogous to information carriers who are willing to spread information to their neighbors. And recovered state R represents the individuals who had previously received the information but later lost interest. But there is still some differences between the spreading process in real social network with epidemic spreading. Thus we improve some details of this model[6]. Firstly, in SIR model, an infected node can transmit virus to all its neighbors in one iteration. But in real social network, the number

of people that one can contact at the same time is very limited. So we assume that one can select only one of his neighbors to contact in each iteration. It's very important. Because if we don't limit the number of contacts of a person at the same time, there is no doubt the individual with the highest degree is the most influential, which not only agrees with the fact, but also makes our study meaningless. Secondly, we assume that one will not contact with the neighbors whom he has contact before. Because nobody tells the same information to one person twice. But it is reasonable that one tell some information to his neighbors who have heard about the same information from other people before. Below is specific implementation. At first, we have a initial spreaders list. And we create a adjacent list for each individual. In each iteration, each spreader selects a neighbor from his adjacent list, and transmit the information to this neighbor with a probability p. Next, he removes this neighbor from the adjacent list. Then he may become uninterested in the information with a probability q and will no longer spread this information. When there are no infected nodes, the spreading process stops. We use the fraction of recovered nodes to the total nodes as the measurement of spreading efficiency. Let R denotes the number of recovered nodes. Let N denotes the total number of nodes in the network. The spreading efficiency can be defined as:

$$Influence = \frac{R}{N} \tag{9}$$

4.2 Benchmark Methods

In complex networks, there are many centrality indexes that can show the importance of each node. Studies have shown that selecting the node with the highest centrality score is an effective strategy to maximize the spreading effectiveness. So we use 5 centrality indexes as our baseline methods to evaluate our heuristic methods.

4.2.1 Degree

Degree is a basic local centrality index for nodes. Generally, the node with higher degree is more influential.

4.2.2 Betweenness

Betweenness[7] is the ratio of the number of shortest paths passing through a node v to all shortest paths between all node pairs in a network. A high Betweenness score means that the node, for certain paths, is crucial to maintain node connections. The expression of Betweenness centrality is

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\delta_{st}(v)}{\delta_{st}}$$

4.2.3 Closeness

We use the variant Closeness centrality[8] instead of Freeman Closeness centrality. It is the sum of inverse distances to all other nodes. The expression is

$$C_C(v) = \sum_{t \neq v} \frac{1}{dist(t, v)}$$

4.2.4 Pagerank

PageRank[9] is an eigenvector-based algorithm. The Pagerank score of a node represents the fraction of time spent 'visiting' that vertex (measured over all time) in a random walk over the vertices (following outgoing edges from each vertex). The expression is

$$C_{PR}(v) = (1 - d) + d(C_{PR}(t_1)/C_{t_1} + ... + C_{PR}(t_n)/C(t_n))$$

where $t_i, i = 1, ...n$, are the Web pages that point to page v, $C(v)$ is the number of links originated at page v, and d is the damping factor. We set $d = 0.85$

4.2.5 Coreness

The Coreness centrality[10] is a simple but notably powerful indicator to assess the capability of information dissemination through the network. It considers both the degree and the Coreness of a node coincidently by counting the k-shell indices of its neighbor or neighbor of neighbors. The basic assumption is that a spreader with more connections to the neighbors located in the core of the network is more powerful. Based on this assumption, the neighborhood Coreness is defined as:

$$C_{nc}(v) = \sum_{w \in N(v)} ks(w)$$

where $N(v)$ is the set of the neighbors adjacent to v and $ks(w)$ is the k-shell index of its neighbor w. Recursively, the extended neighborhood Coreness is defined as follows:

$$C_{nc+}(v) = \sum_{w \in N(v)} C_{nc}(w)$$

where $C_{nc}(w)$is the neighborhood Coreness of neighbor w of v. We use the extended neighborhood Coreness as the Coreness centrality.

5 Results and Conclusions

5.1 Experimental results

We design two experiments to compare these 4 heuristic methods and 5 benchmark methods. In the first experiment, we test different numbers of initial spreaders. In the second experiment, we assign a transmission probability p for each edge. For each simulation under the SIR model, we repeat 100 times to get a more smooth result.

5.1.1 Effectiveness with different fractions of spreaders

We first set $q = 1/\langle k \rangle, p/q = 1.1$ and observe the effectiveness with different fractions of initial spreaders ranging from 0.2% to 2% for these 9 methods. We test all the 9 snapshots of the Gnutella network. Fig 6, 7, 8 are the results for snapshots on August 4, August 8 and August 25 respectively. The results show that Degree Distance, Degree Discount

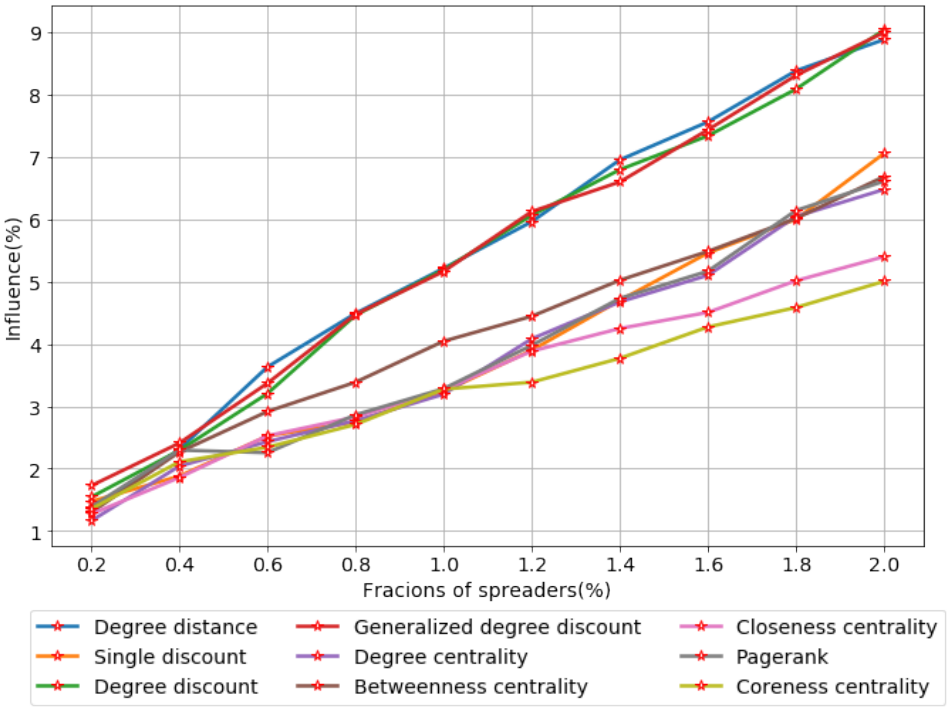

Figure 6: Effectiveness with different fractions of spreaders for snapshot on August 4

and Generalized Degree Discount perform significantly better than other methods for all the snapshots. And the results for different snapshots are similar, caused by the similar topological structures. Moreover, as we mentioned above, when the transmission probability

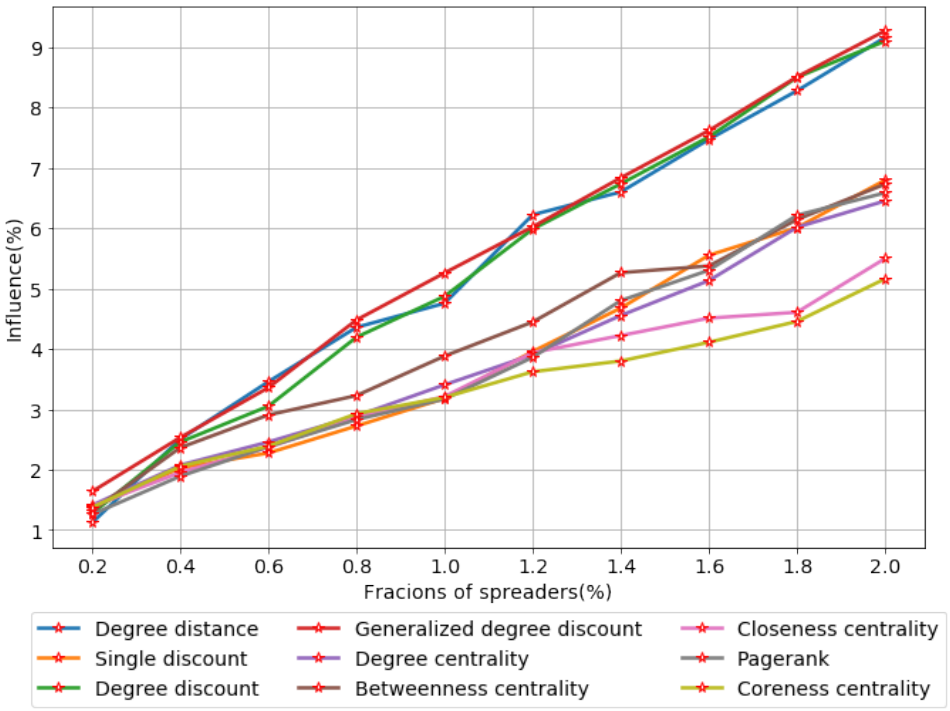

Figure 7: Effectiveness with different fractions of spreaders for snapshot on August 8

p is small, the difference between the Degree Discount and the Generalized Degree Discount is small.

5.1.2 Effectiveness with non-identical transmission probability

In the real social network, the transmission probability p reflects the level of trust between two people and p should be non-identical. So assigning a different p to each edge in term of the frequency of the connection between the two ends of this edge really makes sense. Unfortunately, we don't know the frequency of the connection. So we set $p = 0.2$ if the both ends of an edge are in the same community and set $p = 0.1$ when the two ends belong to different communities. Then we compare the spreading effectiveness of all the 9 methods. Because of the similarity of topological structure of all the snapshots, we use the result of the snapshot on August 8 to represent the results of other snapshots. Fig 9 shows that the Degree Discount and the Generalized Degree Discount perform much better than other methods. One possible reason is that these two methods consider the transmission probability and their selections of spreaders can adjust in term of specific p.

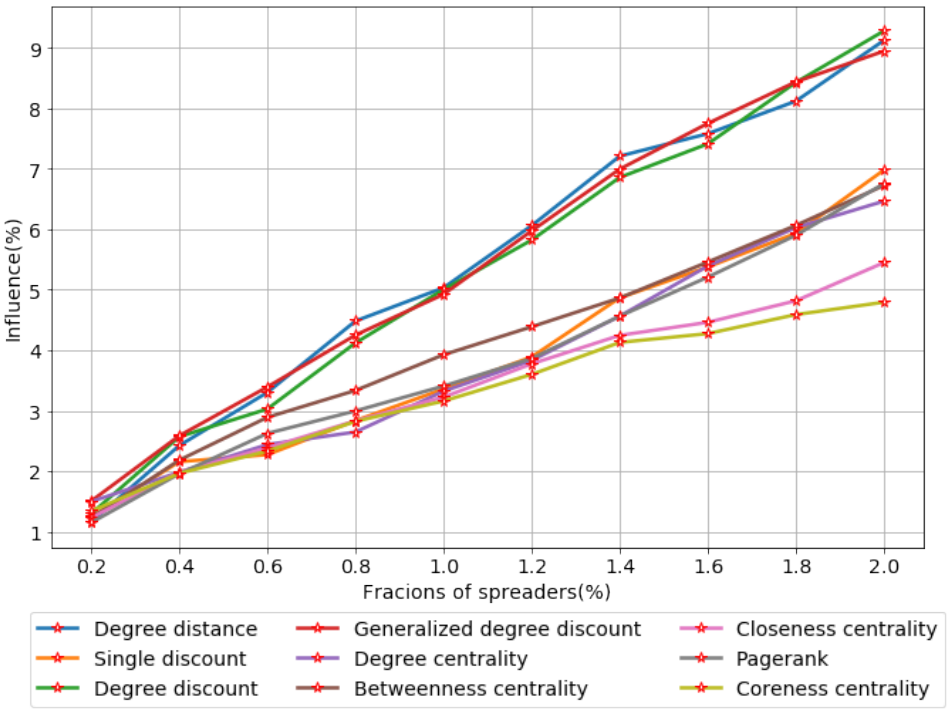

Figure 8: Effectiveness with different fractions of spreaders for snapshot on August 25

5.2 Conclusion

Besides the findings mentioned above, we also observe that Coreness and Closeness perform the worst among all methods. In Ref [11], Liu et al. found that nodes in high shells may not be influential because of the existence of core-like groups: groups of nodes that link very locally within themselves. For nodes in the core-like groups, the Coreness cannot reflect their location importance in the network, which reduces the accuracy of the k-shell decomposition process. Moreover, if nodes in the highest shell tend to links with one another, their influence areas may overlap significantly. Obviously, selecting those nodes as spreaders may cause a large fraction of the network to overlooked. The situation for Closeness is similar: nodes with high closeness values often distribute closely with one another.

In summary, dispersing the spreaders is the key idea of these 4 heuristic methods. In the real social network, there many factors such as transmission probability and the states of neighbors that can influence the final spreading effectiveness when selecting the initial spreaders. Therefore the more factors a method consider, the effectiveness of it is better. In addition, we still ignore many details that are important in reality. For example, if we want to maximize the advertising effectiveness, we should weight each individual in term of his purchasing power. A person with low degree may have a strong purchasing power.

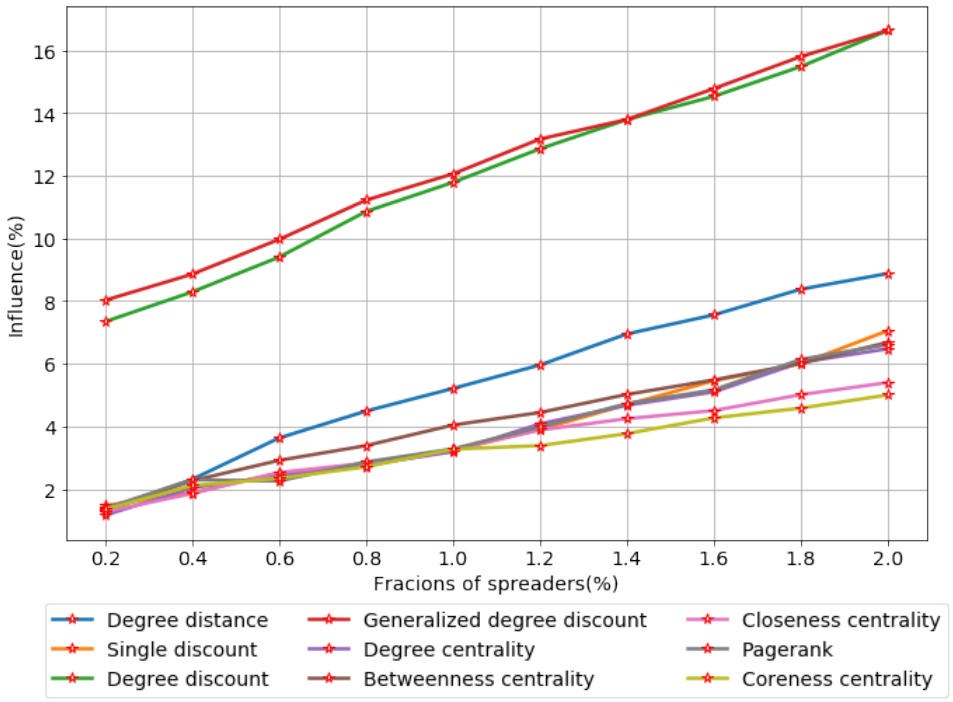

Figure 9: Effectiveness with non-identical transmission probability

So transmitting the advertising to him is more important than to someone who has a wide social cycle (high degree) and a weak purchasing power.

References

[1] Jalili, Mahdi and Perc, Matjaž, Information cascades in complex networks, Journal of Complex Networks, **5**, 665-693 (2017).

[2] Wang, Xiaojie and Zhang, Xue and Zhao, Chengli and Yi, Dongyun, Maximizing the spread of influence via generalized degree discount, PloS one, **11**, 10, e0164393 (2016).

[3] Sheikhahmadi A, Nematbakhsh MA, Shokrollahi A. Improving detection of influential nodes in complex networks. Physica A: Statistical Mechanics and its Applications. 2015; 436:833-845. doi: 10.1016/j.physa.2015.04.035

[4] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2009.p.199-208.

[5] Wang X, Zhang X, Zhao C, Yi D (2016) Maximizing the Spread of Influence via Generalized Degree Discount. PLOS ONE 11(10): e0164393. https://doi.org/10.1371/journal.pone.0164393

[6] Rui Yang, Bing-Hong Wang, Jie Ren, Wen-Jie Bai, Zhi-Wen Shi, Wen-Xu Wang, Tao Zhou, Epidemic spreading on heterogeneous networks with identical infectivity, Physics Letters A, Volume 364, Issues 3–4, 2007, Pages 189-193, ISSN 0375-9601, https://doi.org/10.1016/j.physleta.2006.12.021.

[7] Freeman LC. Centrality in social networks conceptual clarification. Social networks. 1979; 1(3):215-239. doi: 10.1016/0378-8733(78)90021-7

[8] Bavelas A. Communication patterns in task-oriented groups. Journal of the acoustical society of America. 1950; 22:725. doi: 10.1121/1.1906679

[9] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks. 2012; 56(18):3825-3833. doi: 10.1016/j.comnet.2012.10.007

[10] Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nature Physics. 2010; 6(11):888-893. doi: 10.1038/nphys1746

[11] Liu Y, Tang M, Zhou T, Do Y. Core-like groups result in invalidation of identifying super-spreader by kshell decomposition. Scientific Reports. 2015; 5:9602. doi: 10.1038/srep09602 PMID: 25946319