A New Centrality Measure for Influence Maximization in Social Networks

Suman Kundu, C.A. Murthy, and S.K. Pal

Center for Soft Computing Research
Indian Statistical Institute
Kolkata, India - 700108
sumankundu_r@isical.ac.in, murthy@isical.ac.in, sankar@isical.ac.in

Abstract. The paper addresses the problem of finding top k influential nodes in large scale directed social networks. We propose a centrality measure for independent cascade model, which is based on diffusion probability (or propagation probability) and degree centrality. We use (i) centrality based heuristics with the proposed centrality measure to get k influential individuals. We have also found the same using (ii) high degree heuristics and (iii) degree discount heuristics. A Monte-Carlo simulation has been conducted with top k-nodes found through different methods. The result of simulation indicates, k nodes obtained through (i) significantly outperform those obtain by (ii) and (iii). We further verify the differences statistically using T-Test and found the minimum significance level (p-value) when $k > 5$ is 0.022 compare with (ii) and 0.015 when comparing with (iii) for twitter data.

1 Introduction

Large scale online social networks became popular in recent years. Twitter, Facebook, Orkut, LinkedIn is few examples. These social networks have millions of users. People around the globe are connected with the purpose of common interests. These applications are becoming a huge marketing platform of products and services, specially spreading the information to a large number of people in a short amount of time. However, the most important question arises "How to select the influential individual quickly, to target for marketing?"

Domingos et al. were the first to study the problem as an algorithmic problem and proposed probabilistic methods in [3, 10]. In [5] Kempe et al. formulated the problem as a discrete optimization problem and showed that the problem is NP hard. They also proposed a greedy hill climbing approach, which provides $(1-1/e-\epsilon)$ approximation of the optimal solution. Finally, they showed through experiments that their approach provides significant improvement over the classical degree, and centrality based heuristic. However, for large scale graphs, the greedy approach may be time consuming. Chen et al. recently proposed few improvements of the model in [1]. They provided NewGreedy and further modified it to MixedGreedy. Even after the improvement, this approach would take days

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 242–247, 2011.

to run on large scale social networks. So, Chen et al. in [1] provided the degree discount heuristic model which runs much faster than the greedy model. In [4], authors provided another approach to solve the problem in less time. They called it set covering greedy algorithm. This algorithm, however, needs more time compared to the centrality based heuristic models.

In this paper, we propose a centrality measure, *diffusion degree*, and then we use it to rank influential individuals of large sample of directed social networks. We simulate the information spread with top k nodes from different algorithms and compare it with the simulation results of the proposed algorithm. We found proposed algorithm provides statistically significant improvements.

The paper is organized as follows, in Section 2 we provide the information diffusion model. Section 3 describes some related works. In the Section 4, centrality measure *diffusion degree* is described. Section 5 shows experimental results.

2 Information Diffusion Model

Independent Cascade Model: Independent cascade model of information diffusion is proposed by Lopez-Pinatado [6]. It is the most common model for information diffusion. In this model, nodes can have two states, either active or inactive. Nodes are allowed to switch from inactive to active but not in the other. The diffusion model starts with an initial set of active nodes. In time t, an active node u will get chance to activate its inactive neighbor v. v will become active with a probability λ called diffusion probability or propagation probability. u will not get any further chance to activate v. The diffusion probability is a user-defined parameter of the model. The process of diffusion stops when no further activation is possible. This method is called independent because the activation of a node does not depend on the history of active nodes.

3 Related Works

High Degree Heuristic Model: The most classic approach to solve the influence maximization problem is *High Degree Heuristics*. Here the influence is calculated based on the degree of a node, i.e. if k nodes are required to select as seed then the top k high degree node will be selected.

Degree Discount Heuristic Model: General idea of the *degree discount* algorithm of Chen et al. is that if one node is considered as seed then the links connecting with the node will not be counted as a degree of the other nodes, i.e. when considering the next node, the links connecting with the nodes already in the seed set will be discounted.

4 Proposed Diffusion Degree and Heuristic Model

Several attempts are made to improve efficiency of the greedy algorithm. However, for a large scale network its efficiency is far from the speed of centrality

based heuristics. Degree is commonly used for finding the seeds of the influence maximization problem. In [5], Kempe et al. showed through experimental results that high degree heuristics produces a large influence spread compared to other centrality based heuristics. In addition, some of the centrality measures like betweenness require huge computation load to calculate. In this section, we propose a centrality measure, *diffusion degree* based on the diffusion probability. The diffusion degree can be calculated quickly even for large scale networks. A heuristic model is then described for influence maximization problem.

Many of the available centrality measures considered only structural property of a node. However, when considering the diffusion process, diffusion probability plays a vital role in influence flow over the network. Additionally, the centrality based heuristic models did not consider the effect of neighborhood. Take an example of high degree heuristics, suppose a node (v_1) with the highest degree in the network is connected with some low degree nodes. Consider another node (v_2) with a less degree; and its neighbors are high degree nodes. Now, the obvious choice in the high degree model is v_1. In this case, the diffusion process propagate less level compared to v_2 because the neighborhood of v_2 can send the information to more nodes in the network than neighbors of v_1. Our contributed centrality measure considers the above mention properties of diffusion model and social networks.

The general degree centrality measure is proposed by Nieminen in [9]. The degree centrality of node v can be defined as

$$C_D(v) = \sum_{i=1}^{n} \sigma(u_i, v) \tag{1}$$

where function $\sigma(u_i, v)$ defined as,

$$\begin{aligned} \sigma(u_i, v) &= 1 \textit{ if and only if } u_i \textit{ and } v \textit{ are connected} \\ &= 0 \textit{ otherwise.} \end{aligned}$$

In a diffusion process, a node v with propagation probability λ_v, can activate its neighbor u with probability λ_v. So, considerable contribution of node v in the diffusion process is

$$C'_{DD}(v) = \lambda_v * C_D(v). \tag{2}$$

When the diffusion process propagates to the next level, active neighbors of v will try to activate their inactive neighbors.Thus the cumulative contribution in the diffusion process by neighbors of v will be maximized when all of its neighbors will be activated in the previous step. In this scenario, the total contribution of neighbors of v is

$$C''_{DD}(v) = \sum_{i \in neighbors(v)} C'_{DD}(i). \tag{3}$$

The diffusion degree of a node is defined as the cumulative contribution score of the node itself and its neighbors. So, from the equations 2 and 3 we can define the diffusion degree C_{DD} of node v as

$$C_{DD}(v) = C'_{DD}(v) + C''_{DD}(v) \tag{4}$$

$$= \lambda_v * C_D(v) + \sum_{i \in neighbors(v)} C'_{DD}(i) \tag{5}$$

$$= \lambda_v * C_D(v) + \sum_{i \in neighbors(v)} \lambda_i * C_D(i). \tag{6}$$

The diffusion degree measure depends upon the diffusion probability. However, this measure is independent of the nodes already selected. Thus calculating the diffusion degree for every node of the network could be determined in $O(N + E)$ time where N is the number of nodes and E is the number of edges in the network. In defining the diffusion degree, we consider the effect of immediate neighbors to a node because for a small diffusion probability, the effect of neighbor's neighbor of a node may be ignored[1].

Our heuristics model works similar to other centrality based heuristics for finding top k influence maximization problem. The only difference is that we use the diffusion degree instead of classical centrality measures. The algorithm is as follows

1. Find diffusion degree (C_{DD}) for all nodes of the network
2. Select top k nodes for k-top influence maximization problem.

5 Experiment and Results

In our experiment, we use directed social networks e.g. twitter. In case of directed networks like twitter, one person (or node) can influence its followers. It is unlike that one can influence a person he/she following. So, the out degree of nodes is ignored in our experiments. We use Monte-Carlo simulations of the independent cascade model for a sufficiently large number of times to get an accurate approximation of final influence spread. Reader may refer to [7] for additional information about Monte-Carlo methods.

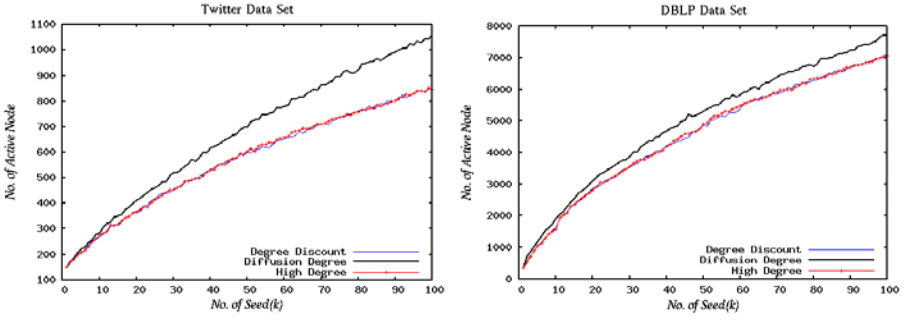
We compare our results with other centrality based heuristics. We avoid comparing results with the greedy approach because for a million node social networks and Monte-Carlo simulation, even the high end server takes days to compute results. Finally, we compare results statistically using T-Test. For more details about the T-Test and p-value readers may refer to [8].

5.1 Data Set

Our primary data set for experiment is twitter data used in [2]. It was obtained by a snowball sampling of the twitter site in late 2009. The data set contains over 400K nodes and more than 800K of relations. Unlike twitter, which is directly related to the problem domain, we use DBLP citation network [11] to verify our claim. The idea behind experimenting with a different data is to verify whether the improvements are only for a particular data set, or it has a similar impact on other real life data sets as well. The DBLP citation network contains over 447K nodes and over 2.3 million relations. Here also, we get better results compare to other centrality based model.

5.2 Results

Figure 1(a) clearly shows that our proposed algorithm outperforms high degree heuristics and degree discount heuristics in case of twitter data set. It is also clear that for directed network like twitter, the degree discount algorithm does not provide any significant improvement over the high degree heuristic. Figure 1(b) shows the results for DBLP citation network. Significant improvement is found in case of DBLP data set as well.



(a) Twitter Data Set with $\lambda = 0.05$ (b) DBLP Data Set with $\lambda = 0.05$

Fig. 1. Seed vs Influence Spread

In our experiment, we assumed that the diffusion probability for nodes is same and we simulated the information spread for $\lambda \in \{0.01, 0.02, ..., 0.07\}$; we found improvement when λ is more than or equal to 0.03. For smaller values of λ our model shows comparable results and for higher values, we are getting further improvements for both the data sets.

We have also performed experiments for different values of k. Specifically, we simulated the information spread for $0 \leq k \leq 100$. Additionally, we verified simulation results for each value of k using T-Test. In case of smaller value of $k (k \leq 10)$ the differences among three algorithms are not significant. However, as k increases, the results from T-Test show that the differences are significant. For twitter data, the minimum observed significant level (p-value) of our method compared to high degree heuristics is 0.022 when $k = 10$. For higher value of k, we found increasing significant differences. We got the highest significant level (p-value $1.46 * 10^{-138}$) when $k = 99$. Thus the results of the proposed method are statistically found to be significantly different from the results of the existing two methods.

6 Conclusion

In this paper, we proposed a centrality based heuristics model for influence maximization problem in social networks. We showed through experiment and

statistical tests that it has a significant improvement over other existing centrality based heuristics for directed networks. We believe our centrality measure, and the heuristic algorithm will provide comparable results for undirected social networks as well.

As a future work, we plan to test the algorithm with other samples of the twitter and to compare our results with a close optimal value produced by the greedy approach. In our model, we only considered the Independent Cascade Model. The work may be extended to see the outcomes in other cascade models as well.

References

[1] 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, pp. 199–208. ACM Press, New York (2009)

[2] Choudhury, M.D., Sundaram, H., John, A., Seligmann, D.D., Kelliher, A.: "birds of a feather": Does user homophily impact information diffusion in social media? CoRR abs/1006.1702 (2010)

[3] Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 57–66. ACM Press, New York (2001)

[4] Estevez, P.a., Vera, P., Saito, K.: Selecting the Most Influential Nodes in Social Networks. In: International Joint Conference on Neural Networks, August 2007, pp. 2397–2402 (2007)

[5] Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 2003, p. 137. ACM Press, New York (2003)

[6] López-Pintado, D.: Diffusion in complex social networks. Games and Economic Behavior 62(2), 573–590 (2008)

[7] MacKay, D.: Introduction to monte carlo methods. Learning in graphical models (1), 175–204 (1998)

[8] Montgomery, D., Runger, G.: Applied Statistics And Probability For Engineers. Wiley, India (2007)

[9] Nieminen, J.: On the Centrality in a Graph. Scandinavian Journal of Psychology 15, 332–336 (1974)

[10] Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 2002, p. 61 (2002)

[11] Tang, J., Yao, L., Zhang, D., Zhang, J.: A combination approach to web user profiling. ACM Transactions on Knowledge Discovery from Data V(March), 1–38 (2010)