Selecting Seed Nodes for Influence Maximization in Dynamic Networks

Shogo Osawa and Tsuyoshi Murata

Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, W8-59 2-12-1 Ookayama Meguro Tokyo 152-8552 Japan
{s_osawa,murata}@ai.cs.titech.ac.jp

Abstract. This paper proposes a method for solving influence maximization problem in a dynamic network. In our method, a node that increases its influence most will be searched and it is added to the seed nodes incrementally. Since exact computation of influence of a node is #P-Hard, we employ heuristics for approximate computation. The results of our experiments show that our method is more effective than the methods based on centralities for dynamic networks, especially when the networks exhibit community structures.

1 Introduction

Influence maximization problem is a problem of selecting the set of k nodes that is the most influential for propagating information (or diseases) to other nodes in a network. Solving this problem is important for minimizing disease propagation or maximizing the effect of advertisement in viral marketing. Since this problem is proved to be NP-Hard[KKT03], obtaining exact answer to the problem is intractable for large networks. Therefore, several methods such as Monte-Carlo simulation and heuristic-based methods have been proposed [CSH$^+$13] [CWW10] [JSC$^+$11] [JHC12]. These research are basically for static networks. Only few attempts have been made for influence maximization on dynamic networks whose edges are dynamically added or deleted.

Naive methods for solving influence maximization problem in dynamic networks are centrality-based methods, which select top k nodes of high centrality values. There are several definitions of centrality for dynamic networks, such as closeness centrality [HS12] and broadcast centrality [GPHE11]. One of the weaknesses of centrality-based methods is that nodes of high centrality might propagate information to adjacent nodes that overlap with each other.

Suppose we are going to select two nodes that are the most influential to the network shown in Figure 1. The number shown at the upper left of each node is its closeness centrality. For the sake of convenience, selected nodes will propagate information to all reachable nodes. Although two nodes of the largest closeness centralities in Figure 1 are nodes D and B, reachable nodes from them are exactly the same (A, B, C, D, and E). This means that selecting node B in addition to node D does not increase the power of influence of seed nodes. In this example, selecting node D and F will be a good choice because all other nodes in the network are reachable from these two. Therefore, just selecting nodes of high centrality values may not be a good method. This is also true in a dynamic network.

G. Mangioni et al. (eds.), *Complex Networks VI,*
Studies in Computational Intelligence 597, DOI: 10.1007/978-3-319-16112-9_9

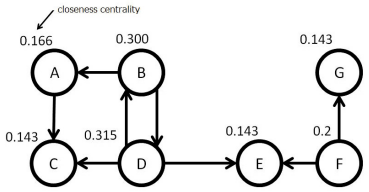


Fig. 1. An example of influence maximization

This paper proposes a method for solving influence maximization problem in dynamic networks. Our proposed method starts with an empty node set $\mathscr{S} = \emptyset$. Then a node n is added to $\mathscr{S}$ incrementally so that the influence of $\mathscr{S} \cup \{n\}$ will be maximized. Since the computation of exact influence is time consuming, the approximated power of influence of node set is computed. Experimental results show that our method is effective especially when a network exhibits community structures.

2 Related Works

2.1 Dynamic Networks

We focus on a network whose edges will appear or disappear dynamically and its nodes are static throughout its period. Such a dynamic network can be represented as a list of adjacency matrices: $G = (A_1, A_2, \ldots, A_T)$ where A_t is an adjacency matrix of a network at time t. T is the period of the dynamic network, and we assume that T is finite. An edge between node i and j at time t is represented as a triplet (t, i, j). A walk of length $k-1$ from node n_1 to node n_k is defined as a sequence of edges: $(t_1, n_1, n_2), (t_2, n_2, n_3), \ldots, (t_{k-1}, n_{k-1}, n_k)$, where $t_1 < t_2 < \cdots < t_{k-1}$ should be satisfied. A walk of no node revisit $(\forall i, j (i \neq j) \;\; n_i \neq n_j)$ is called as a path. The period of a path is the duration of time from the start to the end of the path, which is defined as $t_{k-1} - t_1 + 1$. A path of minimum period is the shortest path, and its period is the shortest period.

An aggregate network G_{agg} of a dynamic network $G = (A_1, A_2, \ldots, A_T)$ is a static network: $G_{agg} = \sum_{t=1}^{T} A_t$, in which times of all edges in G are ignored.

2.2 SI Model for Information Propagation

We focus on SI model [BZW07] as a model for information propagation. In SI model, state S (susceptible) or state I (infected) is assigned to each node. A node in state S does not have information, and a node in state I has information and is ready to propagate. At the initial stage of information propagation ($t = 1$), only seed nodes are assigned to state I and others are assigned to state S. At $t = 1, 2, \ldots, T$, information is propagated in the following steps:

1. For each edge (t, i, j) at time t, the following operation is done:
   a. If node i is in status I and if node j is in status S, then node j will be in status I with probability λ at time $t + 1$.

b. If the network is undirected, information is propagated to both directions. In other words, if node j is in status I and if node i is in status S, then node i will be in status I with probability λ at time $t+1$.

2. Information propagation is terminated at time $T+1$.

λ is a parameter for the ratio of infection. We assume that T is finite so the above steps will be terminated within finite time.

2.3 Formalization of an Influence Maximization Problem

For SI model, we define the power of influence of node set $\mathscr{S}$ as the expected number of nodes in status I at time $T+1$ when seed nodes are given as $\mathscr{S}$, and express it as $\sigma(\mathscr{S})$. Influence maximization problem is a problem of selecting the node set of size k that maximize $\sigma(\mathscr{S})$. In SIR model, which is a generalization of SI model, exact computation of σ for static networks is proved to be #P-Hard[PS12]. Based on this result, we can assume that exact computation of σ for dynamic networks is also #P-Hard.

2.4 Selecting Seed Nodes of the Maximum Influence

2.4.1 Centrality-Based Method

As a naive method for influence maximization, we can compute centralities of all nodes and select k biggest nodes. Closeness centrality in a dynamic network is defined based on an assumption that a node is central if the shortest periods from the node to all other nodes are small, which is expressed as follows[HS12]: $C_i^C = \frac{N-1}{\sum_j d_{ij}}$, where N is the number of nodes, d_{ij} is the shortest period from node i to node j, respectively. In the process of information propagation, not only the shortest path but also other longer paths will play important roles. Since closeness centrality focuses on the shortest path only, it may not be a good metric for information propagation.

Grindrod et al. extend Katz centrality[Kat53] to dynamic networks, and propose broadcast centrality[GPHE11]. Broadcast centrality takes all walks between two nodes into consideration, which is defined as follows: $C_i^B = \sum_{k=1}^{N} Q_{ik}$,where $Q_{ik} = \left[(I-aA_1)^{-1}(I-aA_2)^{-1}\cdots(I-aA_T)^{-1}\right]_{ik}$ and a is an attenuation parameter for discounting longer walks. If the maximum value of the largest eigenvalue of all adjacency matrices is $\lambda_{\max}$, parameter a has to satisfy $a < \frac{1}{\lambda_{\max}}$. The definition of walks by Grindrod et al. is a little bit different from the definition in the last section. In the last section, a walk $(t_1,n_1,n_2),(t_2,n_2,n_3),\ldots,(t_{k-1},n_{k-1},n_k)$ should satisfy $t_1 < t_2 < \cdots < t_{k-1}$, whereas a walk by Grindrod's definition should satisfy $t_1 \leq t_2 \leq \cdots \leq t_{k-1}$ only. In other words, the number of move at each time step in a walk in the last section is limited up to one, whereas there is no such limitation to a walk by Grindrod's definition. Grindrod's definition allows walks that cannot be the paths for information propagation of SI model, so it may not be a good metric for information propagation, either.

2.4.2 A Method Based on Monte-Carlo Simulation

Berger-wolf et al. propose a greedy method for solving influence maximization problem which approximates the power of influence of node set in SI model by Monte-Carlo

simulation[BW07]. However, the method needs much computational time for better approximation. Our proposed method uses fast heuristic instead of Monte-Carlo simulation to approximate the power of influence of node set.

3 Proposed Method for Selecting Seed Nodes

This section proposes a method for selecting seed nodes that starts from empty node set $\mathscr{S} = \emptyset$. In our method, node n that maximizes $\hat{\sigma}(\mathscr{S} \cup \{n\})$, where $\hat{\sigma}(\cdot)$ is approximated power of influence of node set, is added to $\mathscr{S}$ incrementally. $\hat{\sigma}(\mathscr{S})$ is calculated in the following way.

1. Let $\hat{p}_i(t)$ the approximated probability that node i is in status I at time t. $\hat{p}_i(1)$ is initialized as follows:
$$\hat{p}_i(1) = \begin{cases} 1 & i \in \mathscr{S} \\ 0 & i \notin \mathscr{S}. \end{cases}$$
2. At time $t = 2,3,\ldots,T+1$, $\hat{p}_i(t)$ is computed in the following way:$\hat{p}_i(t) = 1 - (1 - \hat{p}_i(t-1))R_i(t-1)$, where $R_i(t)$ is the approximated probability that none of the neighbors of node i at time t propagates information, which are expressed as follows: $R_i(t) = \prod_{j \in \text{neighbors}(i,t)}(1 - \hat{p}_j(t)\lambda)$, where neighbors$(i,t)$ is the set of neighbors of node i at time t.
3. $\hat{\sigma}(\mathscr{S})$ is calculated as the expected number of I nodes at time $T+1$ in terms of approximated probability $\hat{p}_i(T+1)$, i.e. $\hat{\sigma}(\mathscr{S}) = \sum_{i=1}^{N} \hat{p}_i(T+1)$.

An example of exact value and its approximate value of σ are shown in Figure 2. A label of an edge in Figure 2 shows the time that the edge appears. Suppose the seed nodes at time 1 is $\mathscr{S} = \{A\}$, and we are going to compute the probability $p_B(4)$ that node B is in status I at $T = 4$. In exact computation, p_B is affected by edge $(1,A,B)$ only, so the final probability is $p_B(4) = \lambda$. It seems that p_B is also affected by edge $(3,C,B)$, but this is not true. If node C is in status I at time $t = 3$, node B is already in status I, so p_B will not be affected with the edge from C to B. In this way, we have to judge whether each edge actually affect the probability in status I in order to perform exact computation. However, this procedure is computationally expensive.

We propose a method for approximating this computation shown above. In this method, all edges that are connected to a node are assumed to affect the probability that the node is in status I. Based on this method, the above probability $p_B(4)$ in Figure 2 is computed as $p_B(4) = \lambda + (1-\lambda)\lambda^3$, which is $(1-\lambda)\lambda^3$ more than true probability. Our approximation method overestimates the probability of a node to be in status I on networks having cycles.

As for computational complexity of our method, computational time for updating $\hat{p}_i$ needs time that is proportional to the number of edges m in a network. So the computational time for the update is $\mathscr{O}(m)$. In order to select k nodes that should be added to $\mathscr{S}$, approximate computation of σ is repeated N times, where N is the number of nodes in the network. Therefore, the total computational time will be $\mathscr{O}(Nmk)$.

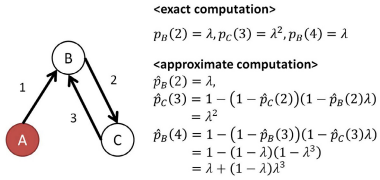


Fig. 2. Approximate computation of $p_B(4)$

Table 1. Statistics of dynamic networks

	nodes	edges	time period	modularity	density
Hospital	75	2,424	5,792	0.367	0.410
Infectious	200	943	469	0.883	0.036
TI model	500	308,000	3,000	0.892	0.006

4 Experiments

4.1 Experimental Settings

We have performed experiments using some dynamic networks and compare the performances of proposed method and some other methods. Three dynamic networks that we used for our experiments are shown in Table 1. Hospital network [VBC$^+$13] shows dynamic proximities of patients and workers in a French hospital. Infectious network [ISB$^+$11] also shows dynamic proximities at a science gallery in Ireland. TI model network is a synthetic network generated by Triad-enhanced Interaction model which is proposed by Jo et al. [JPK11].

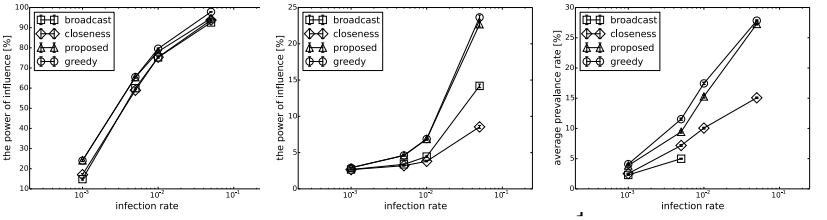
As baseline methods, the following three methods are attempted: (1) a method of selecting nodes of top-k closeness centrality values (closeness method), (2) a method of selecting nodes of top-k broadcast centrality values (broadcast method) and (3) the greedy method based on Monte-Carlo simulation proposed by Berger-wolf et al. (greedy method).

In this experiment, we fix the number of seed nodes $k = 5$ and set infection rate $\lambda = 0.001, 0.005, 0.01, 0.05$ to observe behaviors of methods for values of λ. As for the parameters for broadcast centrality a, for our proposed method λ and for greedy method λ, the same value as infection rate λ is used.

Based on the seed nodes that are selected with our proposed method and the baseline methods, simulations of information propagation based on SI model are performed 1,000 times to calculate the power of influence of seed nodes selected by methods, which is used to evaluate the quality of them.

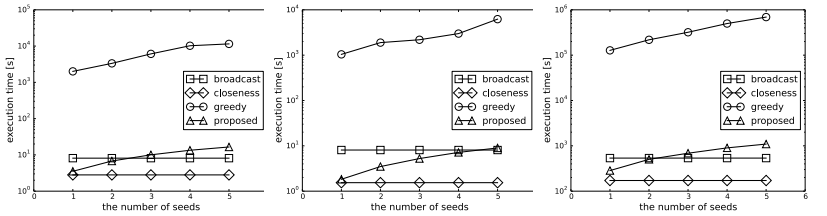
4.2 Results

Results for each network are shown in Figure 3. In this Figure, X axis is infection rate λ, and Y axis is the power of influence of seed nodes selected by each method. For most of values of λ, our proposed method successfully select seed nodes that are more



(a) Hospital Network (b) Infectious Network (c) TI model Network

Fig. 3. The power of influence for values of λ in each network



(a) Hospital network (b) Infectious network (c) TI model network

Fig. 4. Computational time for each network

influential than those selected by two centrality-based methods. But in Hospital network (Figure 3(a)), the power of influence of them are almost equal when $\lambda = 0.05$, and all methods can propagate the information to more than 90% of nodes in the network. On the other hand, in Infectious network and TI model network, the advantage of proposed method becomes larger as the value of λ increases. In TI model network, broadcast centrality cannot be calculated because of the irregularity of matrix $(I - aA_t)$. Compared with greedy method, our proposed method can select seed nodes as influential as the one selected by greedy method even though our proposed method is quite faster than it as shown below.

Computational times for all methods are shown in Figure 4. X axis of the Figure is the number of seed nodes, and Y axis is the computational time. Since closeness method and broadcast method need to compute centralities of all nodes in a network, their computational times are the same regardless of the value of k. On the other hand, computational times of our proposed method and greedy method are proportional to the number of seed nodes k. In all of our cases, the closeness method is the fastest and greedy method is the slowest. Our proposed method is the second or third slowest, but its computational time is still practical even though its performance is almost equal to greedy method which is 500 times slower than the proposed method.

In summary, we can claim that our proposed method can select seed nodes that is as influential as the one obtained with greedy method, which is the most accurate method in the comparison and more accurate than two centrality-based methods, in most of our parameter settings. Computational time of proposed method is slower than two centrality-based methods but is still practical and 500 times faster than greedy method.

5 Discussion

Experimental results in the last section show that for some networks and parameter settings, our proposed method does not outperform two centrality-based methods. One of the reasons for this is that such networks are too dense and they have no community structures. There is no clear definition of community structures especially for dynamic networks. For the sake of convenience, we define "the existence of community structures in a dynamic network" as "the existence of partitions of high modularity[New06] for its aggregated static network". Modularity Q is a function that takes a network and its partition as its input, and a value for showing the goodness of the partition as it output, which is defined as follows:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j),$$

where A is an adjacency matrix of a network, k_i is the degree of node i, C_i is a community that node i belongs to, $m = \frac{1}{2} \sum_{i,j} A_{ij}$ is the number of edges, respectively. High modularity values will be obtained for the partitions whose intra-community densities are high and whose inter-community densities are low. As a method for optimizing modularity, Louvain method [BGLL08] is used.

Modularity values and density for each static aggregate network are shown in Table 1. Modularities of Infectious network and TI model network are very high, while that of Hospital network is not. As for the densities of these static aggregated networks, density of Hospital network is quite high compared with those of other two networks.

If a network is dense and exhibits no community structure, each node in the network can propagate information to many others especially when λ is high. Therefore all methods including centrality-based methods can select very influential seed nodes. On the other hand, if a network is sparse and exhibits community structure, information tends to stay within the communities in which the seed nodes are. In this case, selecting seed nodes from the same communities will be ineffective for information propagation because they may have many overlapping adjacent nodes as we pointed it out as one of the problems of centrality-based influence maximization methods.

6 Conclusion

This paper proposes a method for selecting seed nodes in a dynamic network that are the most influential in information propagation. Experimental results show that our proposed method is effective for some networks compared with the strategies based on centralities for dynamic networks. In comparison between proposed method and greedy method, it is shown that proposed method is as effective as greedy method for some networks, and consistently 500 times faster than it. Our proposed method is especially good for the networks exhibiting community structures.

References

BGLL08. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)

BW07. Berger-Wolf, T.Y.: Maximizing the extent of spread in a dynamic network. DIMACS Technical Report 2007-20, 10 pages (2007)

BZW07. Bai, W.-J., Zhou, T., Wang, B.-H.: Immunization of susceptible–infected model on scale-free networks. Physica A: Statistical Mechanics and its Applications 384(2), 656–662 (2007)

CSH+13. Cheng, S., Shen, H., Huang, J., Zhang, G., Cheng, X.: Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, pp. 509–518 (2013)

CWW10. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1029–1038 (2010)

GPHE11. Grindrod, P., Parsons, M.C., Higham, D.J., Estrada, E.: Communicability across evolving networks. Physical Review E 83(4), 046120 (2011)

HS12. Holme, P., Saramäki, J.: Temporal networks. Physics Reports 519(3), 97–125 (2012)

ISB+11. Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., Van den Broeck, W.: What's in a crowd? analysis of face-to-face behavioral networks. Journal of Theoretical Biology 271(1), 166–180 (2011)

JHC12. Jung, K., Heo, W., Chen, W.: Irie: Scalable and robust influence maximization in social networks. In: ICDM, pp. 918–923 (2012)

JPK11. Jo, H.-H., Pan, R.K., Kaski, K.: Emergence of bursts and communities in evolving weighted networks. PloS One 6(8), e22687 (2011)

JSC+11. Jiang, Q., Song, G., Cong, G., Wang, Y., Si, W., Xie, K.: Simulated annealing based influence maximization in social networks. In: AAAI, pp. 127–132 (2011)

Kat53. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (1953)

KKT03. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003)

New06. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)

PS12. Peyrard, N., Sabbadin, R.: Evaluation of the expected size of a sir epidemics on a graph. UBIAT Resarch Report, RR-2012-1 (2012)

VBC+13. Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B.-A., Comte, B., Voirin, N.: Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. PloS One 8(9), 73970 (2013)