Diffusion Time Minimization on Undirected Weighted Networks: Tree Structure Based Approach

Jeonghoon Kim, Sungchang Kim, Geonsik Yu, Junyeop Lee, Sung-Bong Yang
Department of Computer Science
Yonsei University
Seoul, Korea
Email: {jeonghoonkim90, bidioel, geonsik.yu}@gmail.com, kotzyi@yonsei.ac.kr, sbyang@cs.yonsei.ac.kr

***Abstract*—Many approaches have been proposed for selecting optimal diffusion nodes and maximize the influence. The probabilistic diffusion model that we assumed in this literature is the model which is generally accepted when describing the diffusion occurred by direct contacts such as text messaging. Under this assumption, we found that the influence maximization problem also can be interpreted into diffusion time minimization problems with deterministic link costs. In this paper, we manipulate the deterministic cost characteristic to solve the problem. In our proposal, we avoid the community finding process to remove the excess resource consumption and instead, apply tree structure reorganization process which is much cheaper. After the reorganization, we use root switching on each subproblem to improve the expected performance on the given tree. The conclusion section of this paper shows the practicality of our approach by experiments and comparison to other approaches based on well-known centrality approaches.**

I. Introduction

The problem of influence maximization in social networks has received attention of researchers and companies nowadays. This is because the opportunity of viral marketing so called ”word-of-mouth” is realized to be valuable to many economic subjects. The advent of online social network services such as Facebook, Myspace and Twitter is one of the main factors that shifts the social environment of human connection. This change literally gets rid of the time and distance constraints of social interactions. In this new form of society which is connected by multiple 2-way digital media, the power of information diffusion is amplified by the online social networks. Due to this change, Companies who used the traditional marketing methods relatively lose their market performances and change their policies to the ”word-of-mouth” approach.

Since typical social networks among human society can be represented as graphs, researchers who proposed many practical and applicational methods conducted their studies based on graph theory. More specifically, in terms of the viral marketing, detecting the kernel structure of social networks and selecting valuable nodes in a given sparse and large graph are important issues. Therefore, In the field of network analysis, researchers introduced many methods to detect group structures which are substantial customer clusters and developed other methods to select central nodes which is also important in terms of product marketing. The former, we call those kind of approaches as community detection and the letter we call them centrality based approaches. The problem we concern in this literature and methods we mentioned above is influence maximization on networks. In other words, companies can detect the potential customers who can conduct a vigorous viral marketing activity through the social network in scientific way with those methods and ours.

First, we choose probabilistic diffusion model as our diffusion behavior assumption on the given network and utilize the main characteristic of this assumption or the existence of deterministic diffusion time between any two nodes. Under this assumption, Lu et al. showed that the influence maximization problem also can be interpreted into diffusion time minimization problems in their study [8]. Diffusion maximization based on detecting community structure is proved to be effective enough by existing researches, although its computational cost is expensive [3]. On the other hand, centrality concepts are also introduced to determine how much a node is important in the given network [7]. However both approaches have their own limits in terms of efficiency and effectiveness. The former has higher time complexity and the letter is not effective in selecting multiple diffusion nodes effectively.

Therefore, we concentrated our efforts on solving diffusion time minimization problem both efficiently and effectively. We reorganize the given network by omitting some of the links by applying Dijkstra's algorithm and optimize the subproblem or tree structure by switching its root. In detail, the approach that we propose in this literature has two steps: growing k-tree forest and optimizing the subproblem or tree by root switching. The purpose of the former step is to determine the boundaries of partial optimization by building tree structures. By selecting reorganization of graph, time complexity of our Dijkstra algorithm based process with Fibonacci Heap which is also the bottleneck of our approach accomplish the time complexity of $O(m+nlog(n))$. If we choose clustering based methods such as Girvan and Newman's algorithm, the time complexity will be $O(n^3)$[1]. The latter step is designed to optimize the central node of the tree by considering edge costs. Keep comparing and changing the root node's position and possible candidates we can find optimal node for root in the tree. To show the improvement in terms of the diffusion time, we establish a set of experiments and compare the diffusion time of our approach to the diffusion time of other two existing methods - closeness centrality based approach and diffusion centrality based approach. For the fairness of evaluation, we use the network generator proposed by Lancichinetti et al. [5].

The overall structure of the study takes the form of six sections, including this introduction. Section II reviews the existing researches of our research domain or social network analysis. Section III defines the problem we concerned and gives the diffusion model assumption we choose. The approach that we propose through this literature is described in Section IV and the experimental results that support our algorithm is presented in Section V. Section VI concludes the paper.

II. Related Work

A recent trend in network analysis is focused on finding important participants or meaningful group in the network. More specifically, recent interest in detecting valuable objects is based on the concept of information diffusion. The rapidly grown interest in the information diffusion process and its characteristics is closely related to the appearance of various SNS platforms in recent years such as Facebook. Social network companies who generate massive network type dataset noticed that detecting central nodes in any context is useful in many ways such as building their marketing strategy. Such practical importance, especially in the business perspective accelerates the development of many complex and effective network analysis methods related to the diffusion concept. Kempe et al. designed an approximation algorithm is this social context [4]. The research handles the influence maximization problem and they proved that the approach is valid in several different diffusion models with experiments. Lu et al. also studied the information diffusion [8]. Their study proposed the approximate method to optimize initial diffusion nodes the given weighted network with contact frequencies.

One conventional approach solving this problem is using characteristics of each node so called centrality. From the viewpoint of centrality approach, a node with the higher centrality statistic is considered to be a better node for information diffusion. Although these methods are fast and practical enough on small networks with single diffusion node, centralities have a critical drawback when they are used for large networks which require multiple diffusion nodes. Since centrality values are determined independently for each node, they do not include the information for overlapped effect on diffusion processes. In other words, under centrality criteria, multiple selected nodes for diffusion can be located very close and perform like single node.

Another kind of studies about information diffusion on the static network environment is based on community detection. In other words, in order to find a better node set for effective diffusion, current studies are focusing their efforts to find meaningful node groups efficiently. After grouping nodes in a meaning way, existing approaches select the central node from each group or community. However, we found that the community detection is a highly resource consuming process even though it is a pre-process of the diffusion node selection. Newman and Girvan's method [1] on community detection is generally accepted, although it's time complexity is known to be $O(n^3)$ which is not really practical. Newman, therefore, conducted additional studies on the same problem and reduce the time complexity to $O((m + n)n)$, or $O(n^2)$ on sparse graphs [2].

Therefore, in this research, we sidestep the community finding process which is complex to handle the diffusion time minimization problem efficiently, but still consider the overlap effects of diffusion nodes by using the tree concept to solve the given problem effectively.

A. Existing Approach: Closeness Centrality

The closeness centrality of a node in a certain network represents the reciprocal amount of the sum of the shortest distances to every other node in the network. Since our assumption of information diffusion process defines link costs of the given graph as $|(u, v)|$, the closeness centrality of node u is calculated as follows:

$$\frac{1}{\Sigma_{v\in V}|u, v|} \tag{1}$$

To minimize the overlapped effect of diffusion nodes, closeness based approach remove the selected diffusion node which has the largest closeness value and its related links after every iteration. In other words, after every iteration, V replaced by the set $V' = V\backslash\{n\}$ when the node n is selected in this iteration. When the number of initially informed nodes is 1 which is one of the simplest versions of network analysis, this approach works well. However, since closeness centrality based approach does not consider interactions and the distances between initially selected nodes properly. This blind spot of the approach causes inefficiency in selecting multiple diffusion nodes.

B. Existing Approach: Diffusion Centrality

The diffusion centrality which is proposed by Abhijit et al. [7] is a generalized formula of Katz-Bonacich centrality and eigenvector centrality and can be calculated as follows:

$$\vec{C} = \sum_{t=a...T} (\mathbf{PG})^t \cdot \vec{1} \tag{2}$$

where **P** is a diffusion probability matrix and **G** is an adjacency matrix

III. Preliminaries, Problem Statement and Existing Approaches

A. Graph Representation

Let $G = (V, E)$ denote a weighted and undirected graph consisting of the vertex set V and the edge set E. For two adjacent nodes $u, v \in V$, w_{uv} represent the frequency of communication between the two social network participants. For each node $u \in V$, $d(u)$ is the degree of node u and N_u is the adjacent node set of u, and $d(u)$ can be written as follows: $d(u) = \Sigma_{v\in N_u} w_{uv}$.

B. Assumptions on Diffusion Process

Since we assume that the information diffusion occurs when two participants of the network directly communicate each other just like on mobile messenger platforms, we take *probabilistic diffusion model* to describe our simulation environment for social network and diffusion process. Under this assumption, an inactive node or uninformed person becomes active with some probability $\lambda(u, v) = \frac{w_{uv}}{d(u)}$. This is because the probability $\lambda(u, v)$ represents person u?s interaction with person v proportional for interactions over person u's every neighbor node. In other words, $\lambda(u, v)$ represents the diffusion ratio of one unit contact between the two participants. From a social and behavioral perspective, people are more likely to send informative messages to their closer neighbors than to whom are not. There also exist other popular diffusion models such as *independent cascade model* and *linear threshold model*. However, those models are not appropriate to describe the information diffusion over person to person contacts which we consider in this literature. Although assumptions like linear threshold model can be more reasonable when we simulates information diffusion through bulletin-style SNS platforms like 'My Wall' of Facebook or blog space of 'Myspace', it is not our current issue.

C. Problem Statement

As is well-known in the field of network analysis, the diffusion minimization problem can be described as follows. Let K denote a set of nodes in the given graph $G = (V, E)$ which is informed and activated externally and $|K| = k$ which is also a given condition. Our goal in this problem is to define the set K ⊂ V in order to minimize the expected diffusion time. The expected diffusion time can be calculated deterministically under our assumption. In other words, no matter how many times you simulate the information diffusion process, expected time outcome will be same for the given network $G = (V, E)$ and the set of nodes K. Before we define our problem mathematically, we introduce the definition of expected diffusion on the given network and other derived concepts from it. To explain how to compute the expected diffusion time from K to overall network, we follow the assumption on the expected diffusion time on unit edge given by the probabilistic diffusion model. For two adjacent nodes u and v, the expected diffusion time from u to v is formulated as follows:

$$t(u, v) = \frac{1}{\lambda(u, v)} \cdot \frac{1}{w_{uv}} = \frac{d(u)}{w_{uv}^2} \tag{3}$$

Since the reciprocal number of $\lambda(u, v)$ represents the expected contact frequency between u and v for information diffusion and the reciprocal number of w_{uv} denotes average inter-contact time between u and v. Therefore the product of these two terms is the expected diffusion time from u to v. Under this diffusion time conditions, we can also define expected diffusion time from a to b when they are not directly connected through the concept of the shortest path between two nodes on networks as follows:

$$|(a, b)| = \min \sum_{i=1}^{k-1} t(n_i, n_{i+1}) \text{ where } n_1 = a, n_k = b \tag{4}$$

on the basis we mentioned above, our problem can be stated mathematically as follows:

$$\textbf{Minimize} \max_{v \in V} |(K, v)| \tag{5}$$

where

$$|(K, v)| = \min_{u \in K} |(u, v)| \tag{6}$$

In other words, our goal is to minimize the expected diffusion time from the nodes in K to the farthest node from them. Since this problem is explained to be NP by Zongqing Lu et al., the aim of this research is to establish an algorithm which can find the set K in terms of effectiveness and efficiency by applying tree data structure.

IV. Proposed Method: Tree Based Approach

A. Motivation and Overall Description

Our method is based on the concept of tree data structure which is typically applied to the dataset in order to improve efficiency in searching process. Similar to diffusion time minimization problems which we are mainly concerned with through this paper, the goal of general tree algorithms is to build a structure that has rather even depths or path lengths from a root node to leaf nodes. In addition, It is well known to network analysts that the condition of short path lengths from initially informed nodes to all other nodes is generally helpful in diffusion time minimization. This similarity between two problems implies that if we can rearrange the network structure in proper way and make given network type data a *k*-tree forest, we can interpret roots of the forest as initial diffusion node set which diffuses information more efficiently.

Obviously, Some of link information will be omitted in this link reorganization process and some of the missing edges can be important in the diffusion process. Nevertheless, social networks in the real world are composed of multiple kernels or set of nodes with denser link connections between them and a lot more peripheral nodes with sparse links. We will discuss on this matter in the next section with experiments.

The method that we propose has two steps: growing *k*-tree forest and optimizing the structure by root shifting. The former process is designed to determine the boundaries of optimization by building tree structures and the latter to optimize the central node of the tree by considering edge costs. In following subsection, we will explain the details of our method and give a simple example.

B. k-Tree Expansion

Algorithm 1 Expanding a Forest by Multiple Dijkstra

```
Input: $G, K$
Output: $F$
for each node $r$ in $K$ do
    Queue.enqueue($r$)
    while Queue is not empty do
        $u \leftarrow$ Queue.dequeue()
        activate $u$
        for each $v$ in Neighbor($u$) do
            if $w(v) > w(u) + t(u, v)$ then
                $w(v) \leftarrow w(u) + t(u, v)$
            end if
            if v is not activated then
                Queue.enqueue($v$)
            end if
        end for
    end while
end for
```

Let *G = (V, E)* denote the undirected weighted graph which is connected. In order to pick seed nodes of tree expansion, we establish a degree sorted node table and pick top *k* nodes from it. From these *k* seed nodes which we call *K*, we append branches based on Dijkstra's algorithm for *k* iterations. For each iteration, selected branches are modified by shortest path criterion. As a result, every non-seed node is attached to the closest tree from it. In other words, we produce an optimally partitioned forest. In addition, we also apply Fibonacci heap trick on iteration of Dijkstra's algorithm to reduce the time complexity of our approach on sparse matrix. In the following subsection, we offered an additional process to choose optimize our selection of diffusion node set from this partitioned and simplified structure.

C. Root Shifting

Algorithm 2 Tuning Diffusion Nodes by Root Shifting

```
Input: $F$
Output: $F'$
1:  for for each tree $T$ in $F$ do
2:      flag $\leftarrow$ True
3:      repeat
4:          $r \leftarrow T$.getRoot()
5:          $D \leftarrow \{v|v$ is a direct child of $u\}$
6:          $n_{(1)} \leftarrow v \in D$ with Max $|(u, v, f(v))|$
7:          $n_{(2)} \leftarrow v \in D \backslash \{n_{(1)}\}$ with Max $|(u, v, f(v))|$
8:          if $|(r, n_{(1)}, f(n_{(1)}))| > |(r, n_{(2)}, f(n_{(2)})| + |(n_{(1)}, r)|$
            then
9:              $T$.setRoot($n_{(1)}$)
10:         else
11:             flag $\leftarrow$ False
12:         end if
13:     until flag $=$ True
14: end for
```

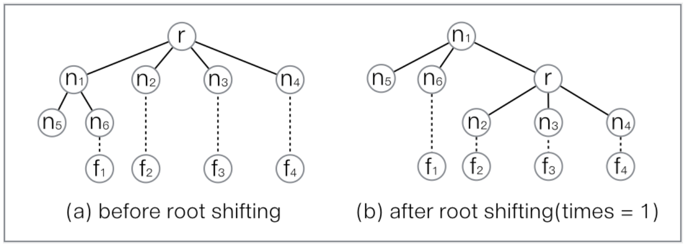
By conducting *k*-trees expansion step, we got approximately divided and simplified problem for initial optimization problem which is NP. In this subsection, we will solve this subproblems by greedily shifting root nodes based on the cost values calculated beforehand. Let *T* be a given tree. We assume without any loss of generality that the node *r* is the current root of *T* and $A = \{n_1, n_2, n_3, \ldots, n_p\}$ is a set of direct child nodes of the root. Let $f(n_i)$ denote the farthest child node of n_i in A. Then we can calculate $|(r, f(n_i)|$ for each farthest node of root's direct child nodes and rewrite the value:

$$|(r, f(n_i))| = t(r, n_i) + |(n_i, f(n_i))| \tag{7}$$

In other words, $|(r, f(n_i))|$ represents the expected diffusion time from root to $f(n_i)$ which is farthest child node of n_i or one of the root's direct child node. For every node in the root's direct child node set A, we can calculate the $|(r, f(n_i))|$ values and find the node $n_{(1)}$ with the largest value and the $n_{(2)}$ with the second largest. Since our goal in this process is to minimize the maximum expected diffusion from the root node, we compare the current maximum time and maximum after root change and determine whether to change the root node or not. The diffusion time from root node to the farthest node from the root will decrease when we shift the root of T from r to $n_{(1)}$, if the tree T satisfies the following shifting condition which is implemented in line 8 of Algorithm 2:

$$|(r, f(n_{(1)}))| > |(r, f(n_{(2)})| + t(n_{(1)}, r) \tag{8}$$

In order to get an optimal root node of our tree, our approach keeps changing its root while condition equation (3) is satisfied.



(a) before root shifting

(b) after root shifting(times = 1)

Fig. 1. Illustration of the roof tuning process, (a) is the given tree and (b) is the state of the tree after one iteration of shifting.

We illustrate one iteration of the root shifting process using an example in Fig. 1. Figure 1a shows a tree T and in the figure, current root node r has four children or $A = \{n_1, n_2, n_3, n_4\}$. Suppose that n_2 be $n_{(1)}$ and equation (2) condition is satisfied. Then we get n_2 as our new root while keeping the given tree T's linkage structure unchanged. Figure 1b shows the result of one iteration of the root shifting process.

V. Performance

TABLE I. Experimental Settings

Parameter	Value	Parameter	Value
n	500, 1000	ξ_1	2
μ_t	0.1, 0.3	ξ_2	1
μ_w	0.1	α	15
β	2	α_{max}	20, 50

This section presents the results of the simulations under different network features and compare two other methods and ours on various fictitious networks. We compare the tree based approach(Tree Based), the closeness centrality based approach(Closeness Based) and diffusion centrality approach, in terms of expected diffusion time. Due to combinatorial complexity of our simulation object domain or networks, we define our target network domain first. For the fairness of evaluation process, we use the artificial networks generated by proposed methods of Lancichinetti et al. [5]. Since our research is conducted on undirected weighted graph domain, we utilize weighted network generator proposed as we mentioned.

In the benchmark network generator [5], there are several parameters confining output network's characteristics. The parameters we used are defined as follows: the number of nodes N; average degree k; the maximum degree k_max; mixing parameter for the topology μ_t; mixing parameter for the weights μ_w; exponent for the weight distribution β; minus exponent for the degree sequence t1; minus exponent for the community size distribution t2.

In order to cover broader experimental domain in the reasonable boundary, we combine the existing experimental settings of the research of Z. Lu et al. [8] and the additional settings we designed. Although network settings proposed by Z. Lu represents partial generality which is accepted by many researchers in advance, we realize that those settings results in generating considerably regular networks. In other words, the distribution of node degrees is far from real world networks due to maximum neighbor constraint. Therefore we assume a new bundle of parameter settings to simulate and compare approaches we mentioned above. To be more specific, we

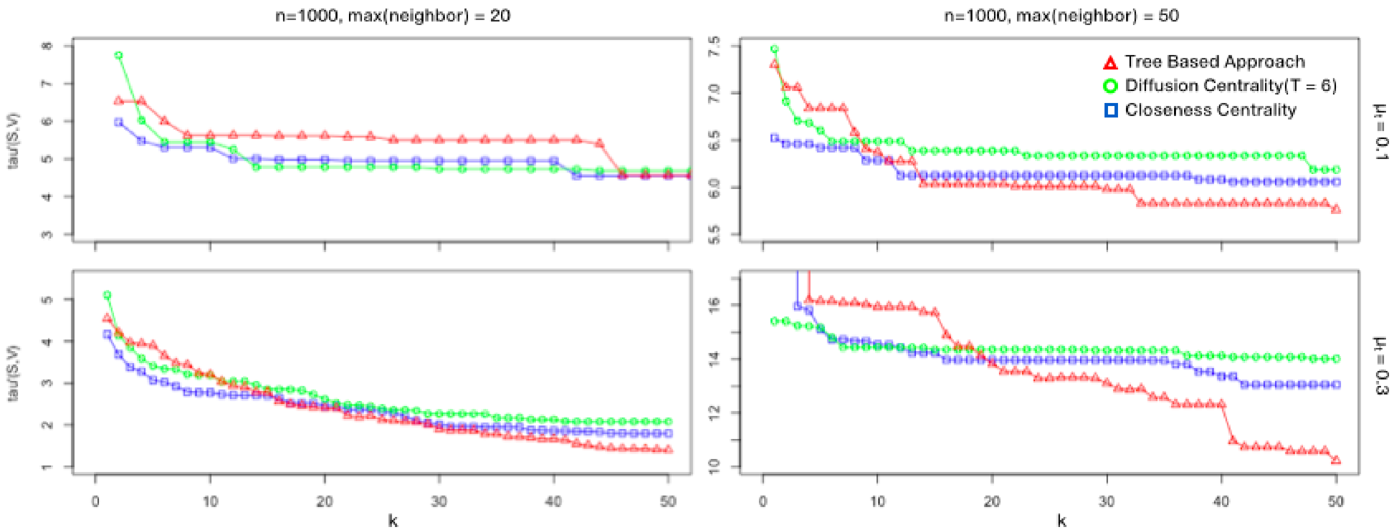


Fig. 2. Expected diffusion time of Closeness Centrality, Diffusion Centrality($T = 6$) and Tree Based Approach with different network settings we designed

choose larger value for maximum neighbor constraint to get a set of networks which have less regularity in their structures. We illustrate all the parameter details about our experiments in Table I.

Since our approach sacrifices some linkage informations from the given network at the first step, It is obvious that the tree based approach we proposed shows better performance on sparse and centralized networks. In the settings that designed by Lu et al. has maximum neighbor setting with the number of 20 and we found that the structure generated by the setting has much higher regularity than real world networks that we found. Therefore we added more settings with less regularity. However on these settings, since the generated networks has the links with outlying values or nearly unreachable, we also removed those nodes to evaluate the general diffusion performance.

Fig. 2 shows the expected diffusion time of the tree based approach and other two comparison methods. We choose k up to 5% of the total number of nodes and conduct experiments. The approach we proposed generally outperform the closeness and diffusion centralities and even more effective when the network structure is irregular and centralized.

In summary, Tree based approach performs better than Closeness Centrality and Diffusion Centrality in terms of expected diffusion time. Specifically, when the mixing parameter for topology is large or the given network is less community structured and the upper bound of node degree is higher or the given network is more centralized, our approach loses less information with reorganization process and perform even better.

VI. Conclusion

In this paper, we proposed tree based approach to solve diffusion time minimization problem on undirected weighted networks. Under the probabilistic diffusion model's assumptions, our approach reduce the influence maximization problem into another problem which is much easier to apply shortest path approach, and we use tree structure to solve it. First, we presented how to reorganize the given network using Dijkstra's algorithm, and then optimize the subproblem by root shifting process. Experiment result shows that our approach is generally more effective in the networks, and even more effective on the centralized and irregular networks. This is because under those conditions, our approach loses less information on reorganization.

A number of additional researches are possible. On the further study, we will focus our effort on finding better seed nodes for tree reorganization process and improve the performance our approach.

References

[1] M. Newman, 'Community detection and graph partitioning', *EPL*, vol. 103, no. 2, p. 28003, 2013.

[2] M. Newman, 'Fast algorithm for detecting community structure in networks', *Physical Review E*, vol. 69, no. 6, 2004.

[3] M. Kimura, K. Saito, R. Nakano and H. Motoda, 'Extracting influential nodes on a social network for information diffusion', *Data Mining and Knowledge Discovery*, vol. 20, no. 1, pp. 70-97, 2009.

[4] D. Kempe, J. Kleinberg and E. Tardos, 'Maximizing the spread of influence through a social network', *Theory of Computing*, vol. 11, no. 1, pp. 105-147, 2015.

[5] A. Lancichinetti and S. Fortunato, 'Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities', *Physical Review E*, vol. 80, no. 1, 2009.

[6] A. Lancichinetti, S. Fortunato and F. Radicchi, 'Benchmark graphs for testing community detection algorithms', *Physical Review E*, vol. 78, no. 4, 2008.

[7] T. Opsahl, F. Agneessens and J. Skvoretz, 'Node centrality in weighted networks: Generalizing degree and shortest paths', *Social Networks*, vol. 32, no. 3, pp. 245-251, 2010.

[8] Z. Lu, Y. Wen and G, Cao, 'Information diffusion in mobile social networks: the speed perspective', *in Proc. of IEEE INFOCOM*, 2014.