Finding Influential Mediators in Social Networks

Cheng-Te Li[1], Shou-De Lin[1], Man-Kwan Shan[2]
Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan [1]
Department of Computer Science, National Chengchi University, Taipei, Taiwan [2]
{d98944005, sdlin}@csie.ntu.edu.tw, mkshan@cs.nccu.edu.tw

ABSTRACT
Given a social network, who are the key players controlling the bottlenecks of influence propagation if some persons would like to activate specific individuals? In this paper, we tackle the problem of selecting a set of *k mediator* nodes as the influential gateways whose existence determines the activation probabilities of targeted nodes from some given seed nodes. We formally define the *k*-Mediators problem. To have an effective and efficient solution, we propose a three-step greedy method by considering the probabilistic influence and the structural connectivity on the pathways from sources to targets. To the best of our knowledge, this is the first work to consider the *k*-Mediators problem in networks. Experiments on the DBLP co-authorship graph show the effectiveness and efficiency of the proposed method.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *Data mining*.

General Terms
Algorithms, Performance, Design.

Keywords
Targeted Marketing, Influential Mediator, Social Networks.

1. INTRODUCTION
Social network plays a significant role as the spread of information and influence for target marketing and immunization setting. The problem, *influence maximization* [4], is to find a subset of influential individuals (as *seeds*) such that they can eventually influence the largest number of people in a social network. Some greedy [6][10] and heuristic [1][2] methods are proposed to effectively and efficiently solve this problem. There are some important variations to tackle different real-world requirements. Leskovec et al. [6] propose to select a set of social *sensors* such that their placements can efficiently detect the propagation of information or virus in a social network. Lappas et al. [5] propose to find a set of *effectors* who can cause an activation pattern as similar as possible to the given active nodes in a social network.

In this paper, we reveal another crucial problem for finding influential *mediators* in a social network. Considering a source node as the seed of influence propagation and a target node as the goal of activation, what are the best mediators to coordinate the spreading process from the source to the target? Potential applications on realistic scenarios include (1) who are the key individuals that hold the bottlenecks of influence propagation to targeted persons given specific initial seeds? (2) In epidemiological setting, given some infected persons and some ones we intend to defend, who are those best few disseminators we should immunize? (3) In the computer network, if some hackers aim at spreading a virus to certain sites, which are the best gateways we should carefully monitor and guard?

WWW 2011, March 28 – April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

Preliminary. First, we adopt the *Independent Cascade* (IC) model [4] for the information propagation. In IC model, in time step t each active node u has a single chance to activate each of its inactive neighbors v with a pre-determined probability $p(u,v)$. If u succeeds, v will become active in step $(t+1)$. Otherwise, u will not activate v again. Second, given the source set of nodes $S \subseteq V$ as the originally active nodes and the target set of nodes $T \subseteq V$, where S and T are mutually exclusive, if a source $s \in S$ can activate a target $t \in T$, we can obtain a *propagation path* $P_{s,t}=<s=v_1,v_2,...,v_m=t>$. And we compute the *activation probability* as $ap(s,t) = \prod_{i=1}^{m-1} p(v_i, v_{i+1})$. If s fails to activate t, then $ap(s,t)=0$. The activation probability between the source set S and target set T is defined as $ap(S,T)=\Sigma_{s \in S, t \in T}\ ap(s,t)$. Third, we further define the *mediation probability* as $mp(S,T,M) = ap(S,T) - ap_M(S,T)$, where $ap_M(S,T)$ is the activation probability from S to T by setting those nodes in M as sinks (i.e., when the information propagates to nodes in M, it can just stops there and make no activations).

Problem Definition. (The *k*-Mediators Problem) Given (1) a social network $G=(V, E, P)$, where V stands for individuals and each undirected edge $(u,v) \in E$ is associated an influence probability $p(u, v) \in [0, 1]$ as weights, (2) a set of source nodes S, (3) a set of target nodes T, and (4) a budget (integer) k, find a set of k nodes (mediators) M with the highest mediation probability $mp(S, T, M)$.

To solve the *k*-Mediators problem, we consider the probabilistic influence and the structural connectivity to develop a greedy heuristic algorithm. Our method consists of three steps. First, we reduce the graph space by retrieving a *reliable induced subgraph* from the social network. Second, we devise a *Mediation-Steiner* algorithm to find the *best propagation tree* connecting sources and targets on leaves. It is greedily combined with the third step to find top-k mediators. The third is to find the k^{th} mediator by a heuristic *proximity-based node selection*. Experimental results demonstrate the effectiveness of the purposed method.

2. THE PROPOSED METHOD
Our method consists of three parts: (1) reliable subgraph extraction, (2) Mediation-Steiner algorithm, (3)proximity-based node selection.

Reliable Subgraph Extraction. Since a real-world social network contains lots of nodes and edges and only a subset of them relates to the user-specified sources and targets, we prune some irrelevant information and retain the effective ones as a reliable subgraph for finding best mediators. The pruning process is controlled by two factors: (a) the probabilistic weights on edges, (b) the connectivity from sources to targets. Those edges with weights below a threshold δ are removed from the original network. The pruning proceeds by increasing the threshold δ and terminates when there exists one source disconnected to any other target. After pruning,

we retrieve the component connecting any source node to any target node and discard other parts. This component is called the reliable subgraph H, which is beneficial to time efficiency.

Mediation-Steiner Algorithm. Based on the reliable subgraph, we find the *Best Propagation Tree* (*BPT*) to capture the most effective pathways of influence propagation from sources to targets. Specifically, we aim at finding a tree that connect the source and target nodes on leaves and has the highest mediation probabilities at all target nodes propagating from some source nodes. We propose to modify the Steiner Tree algorithm [7] to find the best propagation tree. Recall the original Steiner Tree problem is to search out the minimum-cost tree in the input graph containing all required nodes and some intermediate nodes (i.e., Steiner nodes). However, there are three things needed to concern: (a) what we tackle is the influence propagation where the tree cost is measured by product of edge weights instead of sum of edge weights, (b) the required nodes consists of two kinds of nodes (i.e., source and target), and (c) the goal is to route each source to each target by a maximum activation probability. Therefore, we devise a new Steiner Tree algorithm, called Mediation-Steiner, which is shown in line 4–10 in Algorithm 1. Specifically, by adding each source node into *BPT* incrementally, we find the propagation path with the highest activation probability on each target node from those nodes in the current *BPT*. And the corresponding path is added into *BPT*. Note that the function *HighestActivationProb/Path* can be easily derived by modifying the Dijkstra's shortest path algorithm.

Algorithm 1. The Proposed Algorithm.

Input: the social network $G = (V, E, P)$;
the source set S and target set T; a budget (integer) k.
Output: a set of k nodes M.

1: $H = (V_H, E_H, P_H) \leftarrow ReliableSubgraphExtract(G)$.
2: $M = \phi$. // the top mediators
3: **for** i=1 to k **do** // select the top-k mediators in a greedy manner
4: $BPT = \phi$. // the best propagation tree in each run
5: **foreach** $s \in S$ **do**
6: $V_H = V_H \cup \{s\}$.
7: **foreach** $t \in T$ **do**
8: $x^* \leftarrow \text{argmax}_{x \in V_H} HighestActivationProb(x, t)$ in H.
9: **if** $HighestActivationPath(x^*, t) \neq \phi$ **then**
10: $BPT \leftarrow BPT \cup \{HighestActivationPath(x^*, t)\}$.
11: $prox(\{(v|v \in V_{BPT})\}) \leftarrow RandomWalkRestart(BPT, S \cup T)$.
12: $m = \text{argmax}_{v \in V_{BPT} \setminus (S \cup T)} prox(v)$. //proximity-based selection.
13: $M = M \cup \{m\}$.
14: $V_H = V_H \setminus M$.

Proximity-based Node Selection. We consider the structural connectivity to find the bottlenecks of influence propagation as the mediators. The proximity scores with respect to sources and targets are computed by *Random Walk with Restart* [9] in the best propagation tree. We select the top-k mediators by greedily picking the nodes with the highest proximity in *BPT*, as shown in line 3, 11–14 in Algorithm 1.

3. EXPERIMENTAL RESULTS

We conduct the experiments to demonstrate the effectiveness and efficiency of our method. We compile the DBLP bibliography data to a connected co-authorship network, which contains 6,616 nodes and 12,807 edges in some recent premier conferences of data mining (including KDD, ICDM, SDM, PAKDD, and CIKM). The probabilistic weights on edges are determined by the number of co-author between two persons. If #co-author is higher than 20, we set the weight to be 1, otherwise it is set as #co-author/20. One will have higher potential to activate its neighbor if they have more co-works. Since the basic idea of the proposed mediation probability is to find the nodes controlling the bottleneck of influence flows from sources to targets, we compare our method with some alternatives: (1) randomly selecting k nodes in the reliable subgraph, (2) selecting k nodes with the highest CePS-AND score [8], (3) selecting k nodes with the highest betweenness score [3] in the reliable subgraph. We measure the effectiveness by the normalized decay of activation probability $(ap(S,T)–ap_M(S,T))/ap(S,T)$. The results are averaged over 1,000 randomly picked source-target set-pairs, where each source/target set has two nodes. Besides, the number of simulation rounds for independent cascade is 10,000. The result is shown in Figure 1(a). We can observe ours outperforms others, especially when k is small. Besides, for the time efficiency of finding the top-k (k=10) mediators, ours also outperforms CePS. Note that we do not show the runtime of betweenness since it is much longer than ours and CePS.

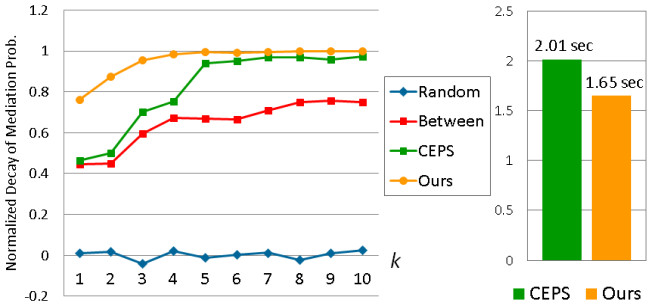

Figure 1(a). Effectiveness comparison our method and alternatives. x-axis the budget k and y-axis is the normalized decay of mediation probability. Higher is better. (b) Comparison of time efficiency.

4. CONCLUSION

We introduce and define the k-Mediators problem which is to find the bottlenecks of influence propagation from given seed nodes to targeted nodes. To solve the problem effectively and efficiently, we consider the pathways of probabilistic propagation and structural affinity to propose a greedy heuristic method, which consisting of pruning irrelevant information, finding the best propagation tree, and selecting the mediators based on the proximity. Evaluations on real-world DBLP co-authorship data show the effectiveness and efficiency of our method.

5. REFERENCES

[1] W. Chen and C. Wang. Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks. In *KDD* 2010.
[2] W. Chen, Y. Wang, and S. Yang. Efficient Influence Maximization in Social Networks. In *KDD* 2009.
[3] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. In Sociometry 1977.
[4] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence through a Social Network. In *KDD* 2003.
[5] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding Effectors in Social Networks. In *KDD* 2010.
[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective Outbreak Detection in Networks. In *KDD* 2007.
[7] H. Takahashi and A. Matsuyama. An Approximate Solution for the Steiner Problem in Graphs. In *Mathematica Japonica* 1980.
[8] H. Tong and C. Faloutsos. Center-Piece Subgraph: Problem Definition and Fast Solutions. In *KDD* 2006.
[9] H. Tong, C. Faloutsos, and J. Y. Pan. Fast Random Walk with Restart and Its Applications. In *ICDM* 2006.
[10] Y. Wang, G. Cong, G. Song, and K. Xie. Community-based Greedy Algorithm for Mining Top-k Influential Nodes in Mobile Social Networks. In *KDD* 2010.