Regional Subgraph Discovery in Social Networks

Cheng-Te Li[1], Man-Kwan Shan[2], Shou-De Lin[1]

[1] Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

[2] Department of Computer Science, National Chengchi University, Taipei, Taiwan

{d98944005, sdlin}@csie.ntu.edu.tw, mkshan@cs.nccu.edu.tw

ABSTRACT

This paper solves a region-based subgraph discovery problem. We are given a social network and some sample nodes which is supposed to belong to a specific region, and the goal is to obtain a subgraph that contains the sampled nodes with other nodes in the same region. Such regional subgraph discovery can benefit region-based applications, including scholar search, friend suggestion, and viral marketing. To deal with this problem, we assume there is a hidden backbone connecting the query nodes directly or indirectly in their region. The idea is that individuals belonging to the same region tend to share similar interests and cultures. By modeling such fact on edge weights, we search the graph to extract the regional backbone with respect to the query nodes. Then we can expand the backbone to derive the regional network. Experiments on a DBLP co-authorship network show the proposed method can effectively discover the regional subgraph with high precision scores.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information Filtering.*

Keywords

Regional subgraph, information backbone, social networks.

1. INTRODUCTION

Social network services (e.g. Facebook, Twitter) have become the major platforms where people can share ideas, participate activities, and communicate with each other. These services allow people to maintain their social relations with others from different countries and regions. That says, for people living in a certain region, their personal social circles could be interwoven with others around the world. Another evidential example for this region-wide social interweaving occurs in the academic collaboration network. Researchers can either work with people in the same country through the colleague-colleague or supervisor-supervisee relations. Though there are many successful proposals on *community detection* in social networks (see the review in [3]), all of them consider only the structure (i.e., tightly intra-connected and loosely inter-connected) to find social groups. The regional information (e.g. country) of people is often neglected. We think the regional prospective can trigger some potential applications: (a) suggesting friends based on regional connections. (b) Identifying scholars in the same region based on certain search criteria (c) In viral marketing, promoting products to customers within the same region.

In this paper, we aim at mining the *regional subgraph* for some user-interested individuals from a *raw* social network. The expected regional network is a subgraph of the original social network and connects all user-specified query nodes that belong to a single region. More importantly, the regional network is expected to include other nodes and links that belong to the same region. Besides, we assume that the regional information (e.g. name, affiliation, and region) is not provided for the discovery task due to privacy issues. In this work we utilize the DBLP co-authorship network for regional subgraph discovery, in which each node is an author associated with some keywords, and each edge is constructed if two authors had ever collaborated. The personal information, including country and affiliation, is used as the ground truth to evaluate our method. We believe such task is a challenging task because (a) no regional knowledge is provided as prior, (b) people in the social network are highly interwoven among different regions, and (c) people in the same region are not necessary directly connected.

Problem Statement. (*Regional Subgraph Discovery*). Given (a) a social network $G=(V,E)$, where each node is associated with a set of keyword labels L, (b) a set of query nodes Q belonging to a certain region r, (c) a filtering threshold δ_f, and (d) a picking threshold δ_p, we aim to mine a subgraph $H=(V_H,E_H)$ as the desired result such that (1) H connects all query nodes, (2) in V_H the percentage of nodes belong to region r is as close to 100% as possible, and (3) in E_H the number of edges whose both end nodes belong to region r is as many as possible.

We mine the regional subgraph by assuming the potential *regional backbone* with respect to the query nodes. The basic idea is there is a hidden backbone for nodes in a region and the query nodes are connected directly or indirectly by the backbone. If we can effectively find the regional backbone connecting the query nodes, it will enhance the discovery of the regional network. By assuming that those belonging to the same region tend to share similar interests and cultures, and thus we model the local interactions between nodes from three aspects of region meanings: frequency, structure, and semantic, on edge weights in the raw social network. Then we develop a *Backbone-Steiner* algorithm to find the regional backbone. In the end, using some heuristics, we expand the regional backbone to obtain the regional subgraph with respect to the query nodes.

Related Works. Existing works about network discovery focus on extracting or constructing the social network by taking the Web as corpus. Y. Matsuo et al. [3] propose a network extraction engine by investigating co-occurrence contexts from Google. J. Tang et al. [5] use probabilistic inference methods to retrieve the academic network from online digital libraries. M. Choudhury et al. [1] explore the effects of different relevance thresholds on inferring social networks. Instead of extracting the entire network, we retrieve the sub-network of a region according to query nodes. To our knowledge, we are the first to tackle such regional subgraph mining in a social network.

2. THE PROPOSED METHOD

Our method consists of three parts: (1) interaction modeling, (2) the Backbone-Steiner algorithm, and (3) heuristic expansion.

Interaction Modeling. Considering the physical meanings of regions, we aim to capture such regional clues from the interactions between individuals and model such clue as edge weights in the raw social network. We investigate three kinds of ideas about regions. The first is *frequency*: people belonging to the same region tend to interact frequently with each other than ones of different regions. In our co-authorship data, we use the number of

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

collaborations between authors as their edge weight. The second is *structure*: those in the same region tend to know each other and form communities. Thus, we compute the number of common friends of two authors as the edge weight. The third one is *semantic*: people belonging to the same region usually experience similar cultures or share similar interests in the network. We use the set of keyword labels L of each author to compute the semantic weight. The weight between author u and v is determined by

$$semanticWeight(u, v) = | L_u \cap L_v | / min\{ | L_u |, | L_v | \}.$$

Backbone-Steiner Algorithm. Given the weighted social graph G=(V,E) and the set of query nodes Q of region r, we are going to find the corresponding regional backbone. Our method, called Backbone-Steiner algorithm, consists of three phases. (1) Query-based expansion: we search the graph from nodes in Q and find the *best expanded tree* T_Q=(V_T,E_T) from the query nodes, where V_T=$Q \cup Q'$ (Q' is the set of expanded nodes). (2) Proximity-based processing: by computing the topological proximities from nodes in V_T, we rule out irrelevant nodes to reduce the search space, and pick the most relevant ones as seed set S. (3) Backbone formation: based on T_Q and the derived seed set S, we greedily compose the regional backbone T_B=(V_B,E_B), which is a subgraph of G. For phase-(a), we employ the *Steiner Tree* algorithm [4] to find the best expanded tree T_Q. 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. For phase-(b), we take advantage of the *Random Walk with Restart* [6] to compute the proximity values. We also utilize the filtering threshold δ_f for filtering and use the picking threshold δ_p to pick the seed set S. For phase-(c), we modify the Steiner Tree method and take nodes in S as the required ones to form the backbone T_B that is extended from the best expanded tree T_Q. The Backbone-Steiner algorithm is shown in the following.

Heuristic Expansion. We compose the regional subgraph by performing the Breadth-First Search (BFS) expansion from the regional backbone T_B. Edge weights are used to determine whether to expand or not. If *weight*(u,v) is larger than τ, where τ is the expansion threshold, the expansion search proceeds. We also use the BFS level (set to 2 in this paper) to control its termination. Note the choice of calculation method for edge weights lead to different quality of discovered regional subgraph. Also note that we set τ to be 0.75 in the experiment after tuning.

Algorithm 1. Backbone-Steiner Algorithm.

```
Input: the social network $G = (V, E, W)$; the query set $Q$;
the filtering threshold $\delta_f$; the picking threshold $\delta_p$.
Output: the backbone $T_B = (V_B, E_B)$.
$T_Q = (V_T, E_T, W_T) \leftarrow SteinerTree(G, Q)$. // query expansion
$proximity(\{(v|v \in G \backslash V_T)\}) \leftarrow RandomWalkRestart(G, V_T)$.
Sort $proximity(\{(v|v \in G \backslash V_T)\})$ in descending order.
$G \leftarrow G \backslash \{v \mid v \in G \backslash V_T$ and $proximity(v)$ below last $\delta_f$ percentage$\}$.
$S \leftarrow \phi$ . // the seed set for backbone formation.
$S \leftarrow S \cup \{v \mid v \in G \backslash V_T$ and $proximity(v)$ in top $\delta_p$ percentage$\}$.
$T_B = (V_B, E_B) \leftarrow T_Q$. // initialize backbone as best expanded tree
while $(S \backslash V_B) \neq \phi$ do // backbone formation by modified Steiner
    $s^* \leftarrow \text{argmax}_{s \in S \backslash V_B} proximity(s)$.
    $v^* \leftarrow \text{argmax}_{v \in V_B} weightedShortestDist(v, s^*)$ in $G$.
    if $weightedShortestPath(v^*, s^*) \neq \phi$ then
        $T_B \leftarrow T_B \cup \{weightedShortestPath(v^*, s^*)\}$.
```

3. EXPERIMENTAL RESULTS

We evaluate the effectiveness of our method by examining if the nodes and edges in the discovered regional subgraph truly belong to the region of the queries. We compile a DBLP co-authorship network, which contains 36,332 nodes and 116,673 edges from 21 primer conferences: {KDD, ICDM, SDM, PAKDD, PKDD, SIGIR, WWW, CIKM, ACL, VLDB, ICDE, SIGMOD, PODS, EDBT, ICML, ECML, AAAI, IJCAI, MM, ICME, MMM}. The keyword labels associated with each author are the textual terms occurring in at least three paper titles that he/she ever participated. In the experiment, we set the region of query nodes is Taiwan. The ground truth (i.e., authors in Taiwan region) is labeled by human. There are totally 1,437 authors and 2,206 co-authorship edges among them. We randomly generate 200 sets of query nodes, where each set contain five Taiwan authors. We compute the average precision and average recall over such query sets: if the discovered regional subgraph contains nodes/edges in the ground truth, the number of hits will increase. The evaluation plan is to present the performance of different stages of our method (i.e., best expanded tree, backbone, and regional network) under diverse interaction models (i.e., frequency, structure, and semantic), where the filtering threshold δ_f is set to be 0.9 and the picking threshold δ_p is set to be 0.0075 (about 170 nodes). The results are shown in Figure 1 and Figure 2. We find the semantic model outperforms the other two, which implies that the knowledge of interests and cultures really help capture the concept of region. We think some noise is introduced in the heuristic expansion stage that hurts the performance.

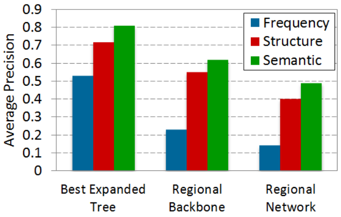

Figure 1. Average precision.

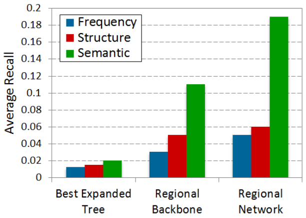

Figure 2. Average Recall.

4. CONCLUSION

This work proposes and solves the problem of regional subgraph discovery in a social network. The central idea of our method is to find the backbone with respect to the query nodes of a certain region. We study the effects of different interaction models on the performance. Experimental results show our backbone-based method with semantic weighting is able to find the regional network with highest precision score. Ongoing work is to investigate what if the query nodes belong to different regions and to identify those individuals shifting from one region to another (e.g. students study abroad or cross-region cooperation) considering the temporal factor.

5. ACKNOWLEDGEMENT

This work was supported by National Science Council, National Taiwan University, and Intel Corporation under Grants: NSC 100-2911-I-002-001, and 10R70501

6. REFERENCES

[1] M.D. Choudhury, W.A. Mason, J.M. Hofman, and D.J. Watts. Inferring Relevant Social Networks from Interpersonal Communication. In *WWW* 2010.

[2] J. Leskovec, K. Lang, and M. Mahoney. Empirical Comparison of Algorithms for Network Community Detection. In *WWW* 2010.

[3] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. POLYPHONET: an Advanced Social Network Extraction System from the Web. In *WWW* 2006.

[4] H. Takahashi and A. Matsuyama. An Approximate Solution for the Steiner Problem in Graphs. *Mathematica Japonica* 1980.

[5] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD* 2008.

[6] H. Tong, C. Faloutsos, and J. Y. Pan. Fast Random Walk with Restart and Its Applications. In *ICDM* 2006.