Personalized Community Detection in Scholarly Network

Zheng Gao
Indiana University Bloomington
1320 E. 10th Street
Bloomington, IN 47405-3907
gao27@indiana.edu

Xiaozhong Liu
Indiana University Bloomington
1320 E. 10th Street
Bloomington, IN 47405-3907
liu237@indiana.edu

ABSTRACT

Most graph clustering methods partition a network into communities based solely on the topology and structure of the network. Due to this, the means via which communities are detected on a network are insensitive to the preferences of a user who is searching the network with a specific, personalized information need. Such partition algorithms may be of diminished value for scholars exploring networks of research if these scholars possess prior preferences on what information they consider relevant. Parts of the graph that align with their research interests are sought out while everything else is irrelevant to their search task. To better address this type of information seeking behavior, we introduce a personalized community detection algorithm that provides higher-resolution partitioning of areas of the network that are more relevant to a provided seed query. This algorithm utilizes the divisive Girvan-Newman approach but incorporates a user's personal preferences as a prior. We show that this personalized algorithm can produce a more fine-tuned partition of a scholarly network when compared to existing prior-insensitive approaches.

Keywords

personalized community detection; graph mining; network analysis

1. INTRODUCTION

A common information-seeking behavior in the academic domain is that of scholars searching for research relevant to their areas of expertise. An ideal user-centered community detection method would provide higher-resolution partitions in areas of the graph most relevant to the user, while the remaining areas of the graph are partitioned in a coarser manner that brings about no loss when considered with respect to the user's information need. The result of such a method would be a graph where fine-grained detail is allocated to areas of interest while the rest of the graph is partitioned at a higher level of abstraction. Although traditional algorithms lead to reasonable network partitions, in their default implementations they operate globally and are insensitive to a user preference for a specific subregion of the graph. Therefore they cannot product this result.

For instance, suppose some experts in information retrieval want to conduct an open-ended search in a network of scholarly research, one not guided by a search keyword but is limited in scope to their domain area. Since they are very familiar with the subdomains and specialties of their field, a graph representation that makes these distinctions is useful. IR experts may be interested in the relationship between their field and other disciplines; medical IR experts may be interested in biology, for example, or image IR experts may be interested in computer vision. Partitioning these related disciplines at the same level of granularity as the home discipline may be distracting at best, and confusing at worst.

To make scholarly network search sensitive to a user's expertise and preferences, we propose an implementation of the Girvan-Newman divisive method to personalize clustering results. Given a user's indication of their search preferences, we employ an edge-based PageRank method to calculate relevance scores on all relationships in the network. Then we combine edge relevance scores and edge betweenness scores together to create an integrated edge-importance metric that is subsequently used to cluster network nodes. Finally we calculate and choose the clustering result with highest modularity.

Section 2 is brief literature review for classic clustering algorithms. Section 3 describes our personalized community detection algorithm via modification of the Girvan-Newman method for community detection. In section 4, using information from a ACM dataset, we generate a homogenous paper network to test our algorithm and compare our result with original divisive method. Our conclusion and future research directions are presented in section 5.

2. LITERATURE REVIEW

The problem of finding community structure in graphs has long been a central research topic in network science. Existing community detection methods can be divided into three categories: partitioning algorithms, spectral algorithms, and dynamic algorithms[3, 4]. Divisive partitioning algorithms start with the entire network and remove edges in some metric-determined order to find community structure, while agglomerative partitioning algorithms find structure by starting with all nodes as isolates and adding edges. Spectral algorithms make use of the spectrum (e.g. eigenvalues) of a

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

similarity matrix derived from the original data to perform dimensionality reduction, thus allowing clustering to be performed in a lower-dimensional space. Dynamic algorithms detect communities based on the probability of the path of a random walk on the network. In 2002, Girvan and Newman first raised edge-betweenness-based divisive method which refers to the number of shortest paths passing through an edge in a graph[7]. Edge betweenness refers to the number of shortest paths passing through an edge in a graph. It serves as a representation of the edge's importance in the network.The method iteratively removes the edge with highest edge betweenness score to naturally divide the graph into several subgraphs.

3. METHODOLOGY

In personalized community detection, our overall goal is to generate clusters with different resolution, dependent on the preferences indicated by the user. This means that clusters containing nodes relevant to the indicated preferences will be divided at a higher level of granularity. In contrast, clusters containing nodes less relevant to the indicated preferences will be larger and more broad in subject scope. If the same personalized algorithm is run with different prior preferences, the resulting partition will be different.

The Girvan-Newman method is the most widely-used divisive method for community detection. The core procedure in Girvan-Newman method is edge betweenness calculation. The edges with the highest edge betweenness scores are iteratively removed, generating isolated subgraphs that are subsequently understood as local communities. In order to modify this method for variable levels of granularity, it is necessary to guarantee that edges connecting nodes more relevant to the users' indicated preferences are more likely to be removed in the earlier iterations of the Girvan-Newman method.

Our method for incorporating user preferences consists of six major steps. The psuedocode is shown in algorithm 1:

Algorithm 1 My algorithm

procedure MyProcedure(Graph<V,E>)
  $P_{node} \leftarrow$ prior node score in Graph
  $SC_{edge} \leftarrow (P_{startnode} + P_{endnode})/2$
  while *iteration* > 0 **do**
    for all *edge* in *Graph* **do**
      $node \leftarrow$ startnode; $T \leftarrow$ out degree of node;
      $E \leftarrow$ incoming edges collection
      $SC_{edge} \leftarrow (SC_{edge} + \alpha \cdot \sum_{e\prime \in E} SC_{e\prime}/T)/(1+\alpha)$
      $iteration \leftarrow iteration - 1$
    end for
  end while
  for all *edge* in *Graph* **do**
    $SP_{edge} \leftarrow$ edge betweenness score for each *edge*
    $f(x) \leftarrow sigmoid(x)$
    $S_{edge} \leftarrow (f(SC_{edge}) + \beta \cdot f(SP_{edge}))/(1+\beta)$
  end for
  while $N(E) > 0$ **do**
    $Graph\prime \leftarrow$ Remove $max(S_{edge})$ from Graph
    if $Modularity(Graph\prime) > Q$ **then**
      $Q \leftarrow Modularity(Graph\prime)$
    end if
    $N(E) \leftarrow N(E) - 1$
  end while
end procedure

It is necessary to have a clear and operational definition of "prior preference" for this particular task. While the Girvan-Newman method operates on edges, our method receives the user's prior preferences for nodes. For step 1, the user is given the ability to select one "seed" node to represent their preferences in the graph. The selected node is then given a prior score of 1, while all other nodes in the network are given a score of 0. To cast a wider net, we can modify this approach by applying a prior score of 1 to the selected seed node and all of its neighbors connected by an edge, that is, all nodes a distance of 1 from the seed. Overall, the prior preference is represented through a mapping function from the user to the nodes in the network.

Step 2 is the calculation of edge-based PageRank which represents edge relevance to user preference[2] . After translating the user's indicated preferences into prior scores P onto nodes, we use a transaction function to transfer the value from nodes to edges. We define the edge prior score $(P_{startnode} + P_{endnode})/2$.

However, because only a few nodes in the network can have prior score 1, most edges will be assigned a score of 0. To smooth the edge scores, we apply PageRank on the edges:

$$SC = \frac{SC + \alpha \cdot \sum_{e\prime \in E} SC_{e\prime}/T}{1+\alpha} \tag{1}$$

Where E refers to all incoming edges to the start node of edge and T refers to out-degree of the start node. After multiple iterations of the PageRank algorithm, the edge relevance scores converge. Edges attached to more relevant nodes will have higher edge scores[1].

Step 3 involves calculating each edge's overall importance score using both edge relevance and edge betweenness. We use sigmoid function to normalize both into (0,1) range and calculate edge importance as:

$$S = \frac{sigmoid(SC) + \beta \cdot sigmoid(SP)}{1+\beta} \tag{2}$$

Step 4 involves removing edges iteratively in the manner of the Girvan-Newman method. After calculating edge importance score, we rank all edges based on edge importance score and at each iteration the highest ranking edge is removed from the graph.

Step 5 is finding the best partition through modularity calculation. After removing the most important edges, we calculate modularity score based on current partition result. Each isolated subgraph will be regarded as a cluster. We use the modularity to evaluate current partition[6]. The formula of modularity for unweighted and undirected graph is

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m}) \cdot \delta(C_i, C_j) \tag{3}$$

Where m is the total number of edges in the graph, and A_{ij} indicates the presence of an undirected edge between node i and node j.k_{ij} represents the degree of the node. The expected number of edgesbetween node i and node j is $\frac{k_i k_j}{2m}$. δ is the Kronecker delta, and has a value of 1 when $C_i = C_j$ and 0 otherwise. Modularity has a range from (-1,1). Traditionally, a higher modularity value is taken to represent a better partition of the graph.

Step 6 is iterating steps 4 and 5 until the partition with the highest modularity score is found.

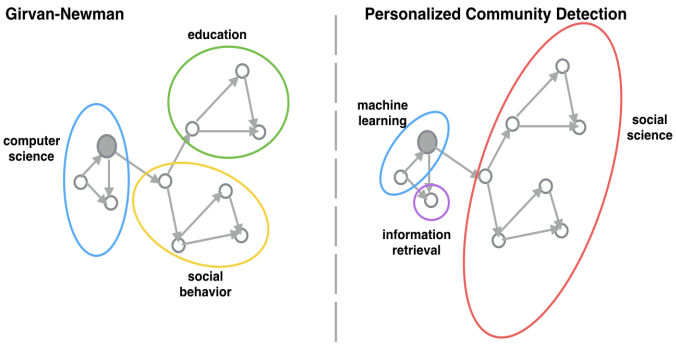

Figure 1: Difference between Girvan-Newman method and personalized community detection method.

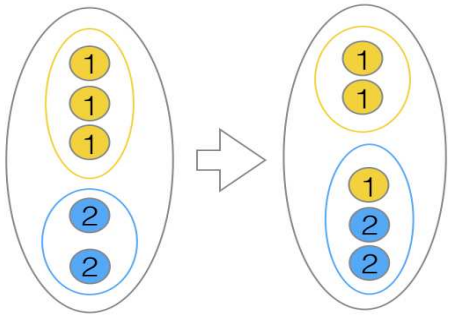

Figure 2: A toy sample for true positive & false negative case calculation

	Paper graph
node type	paper
node number	166,170
edge type	citing
edge number	750,181

Table 1: Dataset Description

The final outcome is a partition of the graph that is at a higher level of granularity in the areas given a higher prior preference score by the user. Figure 1 shows the difference between our method and the original method by Girvan and Newman. The gray node indicates the node selected by the user as a paper of interest. Using the Girvan-Newman method, the granularity of the detected communities is even throughout the whole graph, as seen on the left side of Figure 1. It is insensitive to the user's preference. On the right, we see that the community selected by the user is divided into finer clusters, while those not pertinent are grouped on a coarser level. The user's input causes one particular part of the graph to have a higher partition resolution.

4. EXPERIMENT

4.1 Dataset

The dataset we use is from ACM digital library. In its default format, the network formed by the ACM data is heterogeneous, with nodes representing authors, publications, and venues. Our method requires a homogenous graph, and so we extract an papers-only subgraph from the original data. Table 1 shows the description of the papers graph. We apply our method on only the largest connected component in the graph.

In the end, we decide to use the graph in our approach to test whether our approach can generate better result than original divisive method.

4.2 Experiment Setting

We set alpha in the PageRank equation to a value of 0.5 based on empirical testing result. In order to identify an optimal value for β, we first select the 10 most cited papers of the 15 most influential scholars in the information retrieval domain[1].These 10 papers serve as test seeds whose prior value are set to 1. We then follow the steps above to obtain edge prior scores and use edge-based PageRank to obtain edge relevance scores. The PageRank algorithm is run for 100 iterations. Edge-betweenness scores are calculated using a breadth-first search method. This allows the calculation of the edge importance scores. We calculate the highest modularity using values from the set (0.1, 0.5, 1, 5, 10, 20) for the hyperparameter β and record the resulting partition. We then compare the results across different values of β, and with the partitions produced by the original Girvan-Newman method.

4.3 Result

To create a ground truth to evaluate the efficacy of our method, we manually allocated a set of 15 conferences on the subject of information retrieval into one of the following 5 categories: human-computer interaction, core information retrieval, machine learning, evaluation, and multimedia. In the ACM papers graph, if a paper was published in one of these 15 conferences, it is assigned that conferences corresponding category label. We use this ground truth dataset to determine if the personalized community detection method we have presented performs better in terms of precision, recall, Rand index, and F-value.

This evaluation ground-truth dataset is an assignment of papers to categories. Our method partitions the graph of papers into clusters or groupings that serve as proxies for these categories. Given the ground truth dataset, we consider each pair of nodes assigned to the same cluster as a positive case, and each pair of nodes assigned to different clusters to be a negative case. Each positive pair of nodes that were also in the same cluster by the partitioning method are true positives, while each positive pair of nodes that are in different clusters by the partitioning method are false negatives. A pair of nodes placed in the same cluster by the partitioning method but are not in the same category in the ground-truth dataset are false positives, while a pair of nodes placed in different clusters by the partitioning method and are also in the different clusters in the ground-truth dataset are true negatives. We illustrate our evaluation method using the example in Figure 2.

The left side represents the ground truth, while the right

[1]Microsoft Academic in https://academic.microsoft.com/

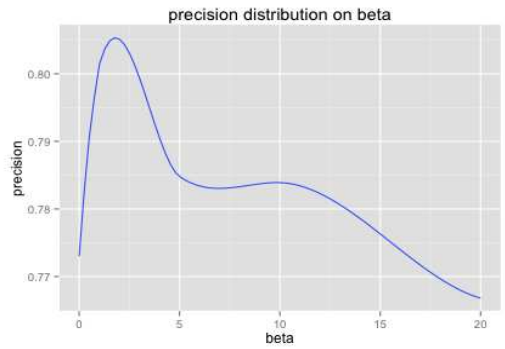

Figure 3: Precision distribution on β

	Paper graph	
	Girvan-Newman	personalized
precision	0.7629	0.8013
recall	0.6320	0.6312
rand index	0.5744	0.5796
F-value	0.6913	0.7003

Table 2: Result Evaluation

side is the hypothetical partition found by a clustering method. In this example, the method has correctly partition all the nodes except one, where a node belonging to cluster 1 was erroneously assigned with the nodes of cluster 2. In the partitioning results, the first (uppermost) partition has two nodes in it. There are 1 pair in this first partition, and it is a correct pairing. There are 3 nodes in the second partition, and therefore there are 3 pairs associated with it. Of those 3 pairs, 1 is correct, but 2 are incorrect. Therefore, in this example case, the partition has 2 true positives, 2 false positives,4 true negatives and 2 false negatives.

Our ground-truth dataset contains 10,548 nodes that represent a paper. For each candidate value of β, we calculated the average precision across the 10 seed papers (Figure 3). The precision peaks at $\beta = 1$ Based on the PageRank formula described above, this value of β indicates that edge relevance is helpful, but if weighted too much will harm performance. Due to this, we chose $\beta = 1$ for the remainder of our evaluation trials. Overall evaluation statistics are presented in Table 2.

Based on these results, we conclude that our personalized community detection algorithm has higher precision, and therefore assigns nodes that are related to each other in the ground-truth dataset to the correct partition more frequently than the default Girvan-Newman method. The recall of our method is slightly lower than the Girvan-Newman, which we believe to be a result of our method's tendency to cluster at a higher resolution. The Rand index is a measure of the similarity between the ground truth and the partitioning result. The result in our algorithm is slightly higher than in the Girvan-Newman method, meaning our method has better fitness when it comes to approximating the ground-truth. The F-value, being the harmonic mean of precision and recall, can be considered a reasonable metric of the efficacy of each algorithm. By F-value, our method slightly outperforms the Girvan-Newman method. The Girvan-Newman method does indeed produce reasonable clusters, but our algorithm, when taking into account prior preference, performs better overall on the papers graph.

5. CONCLUSION

In this paper, we propose a personalized community detection that is an extension of the Girvan-Newman divisive method. We test our approach on a papers network created from data in the ACM digital library. Our results indicate that our enhanced approach is able to incorporate user preference and cluster nodes around user preference with a higher resolution without losing the graph's overall topological structure.

The Girvan-Newman method encounters some problems when applied to some directed graphs with special structure, and so the usual approach is to treat a directed graph as undirected when using this method for community detection. Our method, which is derived from the Girvan-Newman method, suffers from the same problem. Furthermore, the Girvan-Newman method runs with a time complexity of $O(nm)$ on an unweighted network, where n is the number of nodes and m is the number of edges. As such, it has a relatively high computational cost compared to other state-of-the-art partitioning methods like InfoMap. We intend for this personalized community detection algorithm to be able to respond quickly to different users with different preferences, but for each new user preference the method must be run again in its entirety. As such, we are interested in seeing our concept of incorporating our core idea of a user preference prior into other partitioning methods that overcome some of the shortcomings of the Girvan-Newman method. We are considering the use of genetic algorithms, which may allow the graph to be partitioned only once. In such a case, the partition for a new user preference can be calculated from a previous partition through the optimization of some well-defined objective function. We are also interested in seeing how our method performs on more complex heterogeneous graphs[5].

6. REFERENCES

[1] A. Bae, D. Park, Y.-Y. Ahn, and J. Park. The multi-scale network landscape of collaboration. *PloS one*, 11(3):e0151784, 2016.

[2] B. C. Csáji, R. M. Jungers, and V. D. Blondel. Pagerank optimization by edge selection. *Discrete Applied Mathematics*, 169:73–87, 2014.

[3] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[4] E. A. Leicht and M. E. Newman. Community structure in directed networks. *Physical review letters*, 100(11):118703, 2008.

[5] X. Liu, Y. Yu, C. Guo, and Y. Sun. Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 121–130. ACM, 2014.

[6] M. E. Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[7] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.