CS224W Reaction Paper and Proposal

Project Title: Community Detection Using High-Order Structure

Hao Yin [*]

1 Introduction

This project is to perform methodological and empirical research on community detection using local high-order structure. Intuitively, a community in social network is a group of people who interact with each other much more often than with the rest of the world. Community detection has been extensively studied in graph theory and network analysis, and there are tens of models and tools in performing that, such as spectral clustering, Personalized PageRange algorithms, and so on. Most of the previous works aim at finding the subset of nodes that has less incoming and outgoing edges than the edges within the subsets. However, high-order local structure, a.k.a., motif, may contain more information in community level, thus might give better result in community detection. In this project, we are going to examine the existing methods in community detection which analyze edges, and generalize them to fit the motif setting in community detection.

2 Literature Survey

Graph, or network, is a standard way to model complex system and data, where each node represents an entity and each edge represents some interaction or relationship between the corresponding two entities. Under the hypothesis that the world contains various of communities, i.e., group of entities that interact much more often within the group than across the group, community detection has be widely studied in the research of network analysis.

The general methodology in this field of research, especially empirical study, follows the five-part story as described in Leskovec et al. (2008). Besides the part of graph modeling and community hypothesis, most research contains, more or less, each of the following three aspects:

1. Find an objective function on each group (subset) of nodes, which quantifies the "goodness" of a group as a community, i.e., how much more interactions happens within the group as opposed to those across the group;

[*] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. E–mail: yinh@stanford.edu. I would like to thank Austin Benson for his kind help and guidance. This project is also the project I am working on in Prof. Leskovec's group as a research assistant, where I work with Austin Benson.

2. Design an algorithm to exactly or approximately optimizes the objective function, which will give us the group as the best community that is measured by the objective function chosen above;

3. Evaluate the group given by the algorithm. For example, whether the group appear to be a plausible community in the real-world social network, what economic insight can be given by the result, and so on.

In this section, we review and discuss the three most important papers that motivates the project I will be working on.

2.1 Statistical properties of community structure in large social and information networks (Leskovec et al. 2008)

Leskovec et al. focused on the third aspect, i.e., to evaluate the communities and give insights. The major innovation of this paper is the introduction of the *network community profile (NCP) plot*, which measures the quality of the best community for each given size of group. The objective function in this paper is the widely-adopted notion of community goodness called *conductance*, i.e.,

$$\phi(S) = \frac{\texttt{Cut}(S)}{\min\{vol(S), vol(\bar{S})\}}$$

where $\texttt{Cut}(S)$ is the number of edges across the subset of nodes S (i.e., the *cut* of S), $vol(S) = \sum_{v \in S} d_v$ is the sum of degree of all the nodes in S, and $\bar{S} = V - S$ is the complement of S. Formally, the NCP plot is a function on group size, which is the best conductance of all the subset of nodes of the given group size. To compute the NCP plot, for any given group size, we are to solve a constrained optimization problem which is intractable to solve exactly. They used two approximate algorithms, one is the Metis+MQL, the other is Local Spectral Algorithm.

They computed the NCP plot for 70 real-world social and information networks, and the most interesting and important findings are the following:

- The NCP plots for most of the real-world networks are “V”-shaped, that is, the conductance decreases as group size increases at the beginning, reaches its global minimum at the size scale of roughly 100, and then increases afterwards. This suggests that smaller communities is combined into larger, more meaningful communities until it becomes the best community, and then it blends into the rest of the world, making it no longer a meaningful community.

- By examining the global minimum of NCP plot, we find that the best community is usually the largest whisker (i.e., a part of the graph which connects to the rest of graph by a single edge). This means that the graph is of the shape of “jellyfish” or “octopus”, where there is a core with no obvious geometry inside, and some small whiskers that only tenuously connect to the core.

- The observed property of NCP plot can not be reproduced by the commonly used network generating models. However, some sparse random graph model with power-law degree distribution is theoretically proved to have the “V”-shape in NCP plot and whisker-like community. Then they pointed out that the “forest fire” generative model have this desired property.

2.2 Vertex neighborhoods, low conductance cuts, and good seeds for local community methods (Gleich and Seshadhri 2012)

This paper focused on the second aspect, by proposing a new algorithm of only evaluating the neighborhood set of each node and choose the best one. The same as the first paper, this paper also uses conductance as the objective function. This paper proved in theory that if a graph has large clustering coefficient, then there exists a neighborhood set with guaranteed low conductance.

To evaluate this algorithm, the authors implemented it on various real-world network datasets, and compare the conductance with that of other algorithms including Fiedler set, personalized PageRank communities, and whiskers communities. They also compared the NCP Plot of this neighborhood algorithm with personalized PageRank and whiskers communities, and found that this algorithm performs poorly than the other algorithms.

Even though this algorithm does not immediately give a very good community, it can serve as a good seed community used in the personalized PageRank algorithm.

2.3 Higher-order organization of complex networks (Benson et al. 2016)

The biggest innovation of this paper lies in the first aspect, where they used another objective function called motif conductance. Motif is a high-order version (generalization) of edge which relates to more than two nodes. The most simple example of motif is triangle (3-clique) in an undirected graph. Given a motif M, the motif conductance is defined as

$$\phi_M(S) = \frac{\mathtt{Cut}_M(S)}{\min\{vol_M(S), vol_M(\bar{S})\}}$$

where $\mathtt{Cut}_M(S)$ is the number of instances of motif M that have endpoints both in S and in $\bar{S}$, and $vol_M(S)$ is the sum over all the nodes in S of the number of instances each node belongs to. To be clear when using the term *conductance*, in the following, we will refer the widely used conductance as the *edge* conductance.

Like minimizing edge conductance, finding the exact set of nodes that minimizes motif conductance is intractable. To approximately solve the problem, the authors proposed a generalized version of Fiedler Set method (combining the spectral clustering and sweep cut) which is used in approximately minimizes the edge conductance.

The authors implemented this high-order idea and algorithm on various real-world networks, and found new insights into the network organization beyond the analysis upon edges.

3 Discussion and Brainstorming

Among the three aspect of research, the third aspect of evaluating the communities should be central to every research in network analysis since it is why we analyze the social network, and it provides the reason why we should or should not use certain objective function or algorithm.

3.1 Leskovec et al. (2008) and Benson et al. (2016)

Leskovec et al. (2008) provides a novel and interesting way to evaluate the community found by algorithms, and answers a series of questions includes: what is the size of community in an empirical network, what is the shape of the community, the shape of the whole network, and theory corresponding to the empirical observations. The NCP plot can be a useful tool and idea to analyze other algorithms in community detection.

One thing unrealistic in Leskovec et al. (2008) is the existence of whiskers. Whiskers in real-world networks represent the group of people so isolated such that it has only one single connection to the rest of the world, which makes the existence of such community counterintuitive and unrealistic. The reason we observe whiskers from "real-world" data might be that the data (especially the interaction) is not complete, or that some people are not active members in the social network and thus should be excluded from the system. Therefore, in community detection for which we should consider only the real entities, all the whiskers should be removed from the network before we perform any analysis, just like being treated as noise and being ignored. Note that ignoring nodes in network analysis is an option, just like we only consider the largest connected component and perform community detection thereon.

As is compared with Benson et al. (2016), another weakness in Leskovec et al. (2008) is that their use of objective function, i.e., the edge conductance, might not be a good measure in evaluating a group as a community. Edge represents the interaction between two single entities, thus might provide little information in the community level. A motif, such as a triangle, means a strong connection among three people, which is much less likely to happen across different communities. Therefore, the inappropriate use of objective function may result in unrealistic result or incorrect insight in social network structure.

One possible modification of this paper is the perform the similar analysis but substituting edge cut and conductance by motif cut and conductance. The method might be just like what is done in Benson et al. (2016), first transform the original social network into a weighted graph on the same set of nodes, where the weight of each edge represents the number of motif instances the corresponding pair of entities belongs to in the original network, and only consider the largest connected component of this weighted graph. We can then perform the NCP plot analysis on the largest connected component, and uses this method to examine the same datasets to see if there are different finding or insights.

3.2 Gleich and Seshadhri (2012) and Benson et al. (2016)

Gleich and Seshadhri (2012) motivates us to look at the neighborhood set of each node, which might already be a very good community. This idea makes sense, since all the friends might form a good community, or at least a good starting point of community detection where we might find a very good one by modifying the neighborhood set.

As is compared with Benson et al. (2016), one weakness in Gleich and Seshadhri (2012) is that they only consider edges instead of high-order motif structure. One possible extension of their result is to consider how the neighborhood set can give a good motif conductance, thus still being a good community as is measured by the motif conductance.

Another unrealistic result in their paper is that the upper bound on the edge conductance of neighborhood set is too loose when the clustering coefficient κ is low. Note that in real-world networks, the clustering coefficient is not very high (usually 0.2 to 0.6), and the upper bound is only nontrivial (less than 1) when $\kappa > 0.5$. Therefore, their theoretical result does not cast much insight on how the neighborhood set be a community for real-world social network. We will work on it to see if we can find a tighter bound.

4 Project Proposal

Our project is to conduct methodological and empirical research on community detection using local high-order structure, and it will contain the three aspects as is described in Section 2 and Leskovec et al. (2008). To be specific, the project contains the following three components.

4.1 Objective Function

We will use motif conductance as our objective function to quantify the goodness of a group of nodes as a community. The motif conductance is defined in Benson et al. (2016) and has been discussed in the literature survey and discussion sections above. Note that we have not chosen which specific motif we will use in the conductance, and we are going to compare the different motifs in our empirical study on real-world networks to find the best one.

4.2 Algorithm

We are going to compare three algorithms in finding the optimal subset of nodes as a community. Besides working on the theory of each algorithms, we are going to implement them on real-world network datasets to compare their performance, including the quality of the optimal community in terms of motif conductance, and the statistical property of the communities such as the size and shape.

The first one is to first transform the original graph into the weighted graph, as is discussed in Benson et al. (2016), and then apply the Personalized PageRank algorithm to find the approximate optimal community. The Personalized PageRank algorithm in community detection is proposed by Andersen et al. (2006). It can be used to both directed and undirected graph, and weighted graph, and has approximation ration known as the "local" Cheeger Inequality. One limitation of this algorithm is in the transformation of the original graph to the weighted graph, which will only give us an equivalent problem when the motif contains three nodes. With motif on more than three nodes, it will have a worse approximation ratio than the "local" Cheeger Inequality.

The second algorithm is the one described in Benson et al. (2016). After transforming the original graph into weighted graph, we use the spectral clustering method to find the Fiedler Sets and then find the optimal Sweep Set. This algorithm can be applied to both directed and undirected graph, and weighted graph. It has approximation ratio known as the global Cheeger Inequality. Regarding the limitation of this algorithm, besides the one also in the first algorithm that the approximation ration might be worse for motif on more than three edges, another limitation is that we can not compute the NCP plot using this method.

The third algorithm is the one proposed in Gleich and Seshadhri (2012) where we consider the neighborhood set of each node and find the one with lowest motif conductance. Right now we have some theoretical result of this algorithm, saying that if we are interested in k-clique motif in an undirected graph, there exists a neighborhood set with guaranteed motif conductance, and the upper bound on the motif conductance is a function of a high-order clustering coefficient. There are a few limitations of this algorithms. One limitation is that the theory of guaranteed conductance is only valid if we work on undirected graph, interested in k-clique motif, and has high high-order clustering coefficient. In the empirical study, we are going to compute and compare the high-order clustering coefficient on real-world networks to see if it is high enough such that the theory will give us a meaning upper bound. Another limitation is that, the neighborhood method does not performs as well as the Fiedler Set and Personalized PageRank method in the edge conductance setting, and thus not likely to perform well in this motif-conductance setting.

4.3 Empirical Study

We are going to apply and compare different objective functions and algorithm on real-world networks. We are going to use the SNAP Dataset (Leskovec and Krevl 2014), which contains more than 100 network data, of various types and background. Our empirical study will cover the following topics:

(a) Compare the performance of three algorithms in Section 4.2, including the quality of the optimal community in terms of motif conductance, and the statistical property of the communities such as the size. We will also build the NCP plot for the PPR and neighborhood algorithms;

(b) Compare different motifs being used in the motif conductance we are to minimize;

(c) Having found the best algorithm and objective function, we are going to examine the statistical property, especially the NCP plot, of the best community, and try to derive insights from the result;

(d) Computing the high-order clustering coefficients for different networks, and compare them with the regular clustering coefficient.

References

Andersen, Reid, Fan Chung, Kevin Lang. 2006. Local graph partitioning using pagerank vectors. *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, 475–486.

Benson, Austin R, David F Gleich, Jure Leskovec. 2016. Higher-order organization of complex networks. *Science* **353**(6295) 163–166.

Gleich, David F, C Seshadhri. 2012. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 597–605.

Leskovec, Jure, Andrej Krevl. 2014. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

Leskovec, Jure, Kevin J Lang, Anirban Dasgupta, Michael W Mahoney. 2008. Statistical properties of community structure in large social and information networks. *Proceedings of the 17th international conference on World Wide Web*. ACM, 695–704.