CS224W Project Milestone

Project Title: Community Detection Using Local High-Order Structure

Hao Yin *

1 Introduction

Community detection has be widely studied in the research of network analysis, and there are tens of existing tools and techniques to perform this, such as Fielder community (Fiedler 1973), Personalized PageRank algorithm (Andersen et al. 2006), 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 (Benson et al. 2016), thus might give better result in community detection.

Benson et al. (2016) generalized the results on Fielder community to motif level. However, the most commonly used way, as well as the one with best performance, to find community based on edge-level information is the Personalized PageRank algorithm, which can not be directly applied to motif analysis. Thus an interesting question is to generalize the Personalized PageRank algorithm to the motif setting, to compare its performance to the Fielder Community, and edge-level algorithms. In this project, we generalized the Personalized PageRank algorithm from edge level to motif level. This part will be covered in Section 4.

Besides working on algorithm, we also generalized the definition of clustering coefficient to its high-order version, to make the edge-level analysis smoothly generalized to motif level. This part will be covered in Section 3.

2 Basic definitions and notations

Let $G = (V, E)$ be an undirected, unweighted, loop-less graph. We use $n = |V|$ to denote the number of vertices and $m = |E|$ to denote the number of edges. For a vertex v, denote the degree of v by d_v, and we denote the set of its direct neighbors as $N(v)$.

A triple (u, v, w) is called a triangle if $(u, v), (v, w), (u, w) \in E$. Let T be the set of triangles in G, and T_v be the set of triangles with v as an endpoint. Analogously, we use Q to denote the set of 4-cliques in G, and Q_v to denote the set of 4-cliques with v as an endpoint.

*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.

2.1 Clustering coefficient

A wedge is an unordered pair of edges $\{(u, v), (u, w)\}$ that share one common node, and the node u is called the center of the wedge. A wedge is called closed if there is an edge between v and w, i.e., the three nodes u, v, and w forms a triangle. A wedge is called open otherwise. Let W be the set of wedges in G, and W_v be the set of wedges centered at v. Note that $|W_v| = d_v(d_v - 1)/2$, and $|W| = \sum_v |W_v|$.

For a given node u, its local clustering coefficient C_v is defined as the proportion of closed wedges that centered at u, i.e., $C_v = |T_v|/|W_v|$. We also define the global clustering coefficient κ as the proportion of closed wedges in G, i.e., $\kappa = 3|T|/|W|$. Here note that every triangle produces three closed wedges.

2.2 Cut and Conductance

For a given set of nodes, the cut is defined as the number of edges that has one endpoint in S and the other one in $\bar{S}$, where $\bar{S} = V - S$. We denote the cut by $\texttt{Cut}(S)$.

Conductance of a set of nodes S is defined as

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

where $vol(S) = \sum_{v \in S} d_v$ is the sum of degree of all the nodes in S. Conductance is a commonly-used measure on how good this set of nodes is a community.

Benson et al. (2016) generalized the definition of cut and conductance to motif level. 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 cut of a set S, denoted by $\texttt{Cut}_M(S)$, is the number of instances of motif M that has points in both S and $\bar{S}$. Motif conductance is defined as

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

where $vol_M(S)$ 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.

3 Generalization of clustering coefficient

In this section, we consider only undirected graph. We first generalized the definition of clustering coefficient to triangle level in Section 3.1, then shows that this definition will generalized a result on edge conductance by Gleich and Seshadhri (2012) to triangle conductance, and then shows the relationship between the classical clustering coefficient and 3-clustering coefficient, as well as empirical findings, in Section 3.3.

3.1 Definition of 3-clustering coefficient

A 3-wedge $((u, v, w), (u, x))$ is a pair of a triangle and an edge (not an edge of the triangle), and we call u the center of the 3-wedge. A 3-wedge $((u, v, w), (u, x))$ is called closed if the four vertices u, v, w, x form a 4-clique, and is called open otherwise. Let W^3 be the set of 3-wedges in G, and W_v^3 be the set of 3-wedges centered at v. Note that $|W_v^3| = |T_v| \cdot (d_v - 2)$, and $|W^3| = \sum_v |W_v^3|$.

Let κ_3 be the global 3-clustering coefficient, which is the probability that a uniform random 3-wedge is closed, formally,

$$\kappa_3 = \mathbb{P}_{w \sim W^3}[w \text{ is closed}] = \frac{\text{number of closed 3-wedges}}{|W^3|} = \frac{12 \cdot |Q|}{|W^3|} = \frac{12t_3}{t_1 + 4t_2 + 12t_3}.$$

For a vertex v, denote by C_v^3 the local 3-clustering coefficient of v, i.e.

$$C_v^3 = \mathbb{P}_{w \sim W_v^3}[w \text{ is closed}] = \frac{\text{number of closed 3-wedges in } W_v^3}{|W_v^3|} = \frac{3 \cdot |Q_v|}{|W_v^3|},$$

3.2 Theoretical result for neighborhood community

In this subsection, we generalize the result in Gleich and Seshadhri (2012) using our definition of 3-clustering coefficient. Intuitively, it shows that large 3-clustering coefficient implies the existence of neighborhood cuts with low 3-conductance.

Theorem 1 *For any graph G of global 3-clustering coefficient κ_3, then for any $a > 1$, there exists a neighborhood cut with 3-conductance at most $\frac{1-\kappa_3}{1-\kappa_3+12 \cdot \frac{a\kappa_3 - 1}{a(a-1)}}$.*

A detailed proof of this theorem will be given in final report.

3.3 Comparison of C_v and C_v^3

Both C_v and C_v^3 measures the local clusteringness of a node, thus there should be a strong correlation between them. We proved the following results regarding the distribution of C_v^3 as opposed to C_v.

Theorem 2 *For any node in the graph, we have the following results about the local 3-clustering coefficient:*

- $0 \leq C_v^3 \leq \sqrt{C_v}$;
- *In Erdos-Renyi model, we have $C_v^3 = (C_v)^2$.*

Again, a detailed proof of this theorem will be given in final report.

I computed the distribution of local clustering coefficients for all the undirected network in SNAP datasets, and it shows that $C_v^3 \approx C_v$ in most of the network. Note that this result shows that the distribution of 3-clustering coefficient is significantly higher than the baseline Erdos-Renyi model, and small its upper bound. This might imply some local structure of neighborhood, which I will examine further.

4 Combine PPR and triangle conductance

The most commonly used way, as well as the one with best performance, to find community on edge level is using Personalized PageRank algorithm by Andersen et al. (2006). This algorithm is originally designed on undirected unweighted graph to find a subset of nodes with low edge conductance. It can be easily modified to implement on a weighted graph, just like adding parallel edges which the original algorithm permits.

Therefore, to use Personalized PageRank algorithm to find subset with low triangle conductance, we can first apply the technique in Benson et al. (2016) to transform the original graph to a weighted graph, where the weight on the each edge is the number of triangles this edge participates in. Then the problem of minimizing triangle conductance of the original graph is simply minimizing the weighted edge conductance on this transformed weighted graph.

To compare the set given by the Personalized PageRank algorithm using triangle conductance as well as using edge conductance, I compute the F1-measure of the output communities with the ground-truth community, and it shows that triangle conductance gives much higher detection rate. Also, one commonly-observed drawback of Personalized PageRank algorithm on edge-conductance is that it tends to find much larger set than the real community. The experiment shows that, by examining triangle conductance, the Personalized PageRank algorithm tends to output much smaller set than using edge conductance.

I implemented my code on two networks with ground-truth community on SNAP, i.e., `com-DBLP` and `com-Amazon`. For each of the network, I examined 100 communities of size no less than 10. For each of these communities, I take each node as the seed in the Personalized PageRank algorithm, then take the output community of best F1 measure across all seeds as the detection rate of this community. The result can be summarized as the following:

`com-DBLP`	Avg F1 measure	Avg output size	Avg true size
edge conductance	0.04943	519.68000	16.23000
triangle conductance	0.19142	81.29000	16.23000
`com-Amazon`	Avg F1 measure	Avg output size	Avg true size
edge conductance	0.07434	809.15000	34.09000
triangle conductance	0.32071	141.25000	34.09000

As we can see from table, by using triangle conductance, we will have acquire much larger F1 measure and thus high detection accuracy of community recovery. The size of the output community is now comparable to the true size, which indicates that combining triangle conductance and Personalized PageRank algorithm is a very promising way in community detection.

5 Future works

Based on the promising performance of combining Personalized PageRank algorithm and triangle conductance so far, most of the future works will center around how to illustrate the power of this

method. The followings are a list a experiments I am going to work on:

1. Apply the algorithm to more networks with ground-truth communities on the SNAP website, such as `com-LiveJournal` and `com-Orkut`.

2. Apply the algorithm to the network `ego-Facebook`. This network contain egonet information where each node belongs to a few communities. We are going to look at what type of communities will the PPR+triangle conductance method detect across all the communities each node belongs to, and compare this with the PPR+edge to see if there is a difference.

3. Apply the algorithm to Stochastic Block Model. Stochastic Block model is a network generating model with two communities, and the experiment is to see if we can recover the underlying community in the graph generating process. I will apply the algorithm to SBM with different parameter set and find the set of parameter values that the PPR+triangle conductance will lead to a successful detection. I will compare this set to the set by the PPR+edge conductance method, as an evidence of which method is more powerful.

Besides these experiments, I will also implement an on-the-fly version of combining PPR with triangle conductance. Currently the method needs to first list all the triangles in the network to create the transformed graph. This might be too expensive in large networks. A feasible way to get around listing all the triangles at the beginning is to list the triangles only when we have reached the corresponding node. We will compare the computing time of the on-the-fly version and the current version to check the scalarbility of this method.

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.

Fiedler, Miroslav. 1973. Algebraic connectivity of graphs. *Czechoslovak mathematical journal* **23**(2) 298–305.

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.