# Paper Summary

David Miller
CIS 5930: Social Network Mining

February 5, 2018

A Graph Cluster is a graph $G' \subseteq G = \{V, E\}$ that it is partitioned into sets $C = \{C_1, \ldots, C_n\}$ such that $C_i \cap C_j = \emptyset$ for any $i \neq j = 1, \ldots, n$. The graph $G$ is the original graph of data input. If the clustering algorithm breaks edges between vertices in different clusters then $G' \subsetneq G$ since $e_{i,j} \notin G'$ for $v_i \in C_m, v_j \notin C_m$. The main gaol of clustering algorithms are to identify clusters of densely linked nodes given only the graph itself. However, in many cases we only care about a cluster that pertains to a set of nodes $S = \{v_1, \ldots, v_k\}$. This is where local graph clustering comes into play. Local graph clustering is a specific case that takes an additional input in the form of a seed set of vertices. The idea is to identify a single cluster nearby the seed set without ever exploring the entire graph, which makes the local clustering methods much faster than their global counterparts [1].

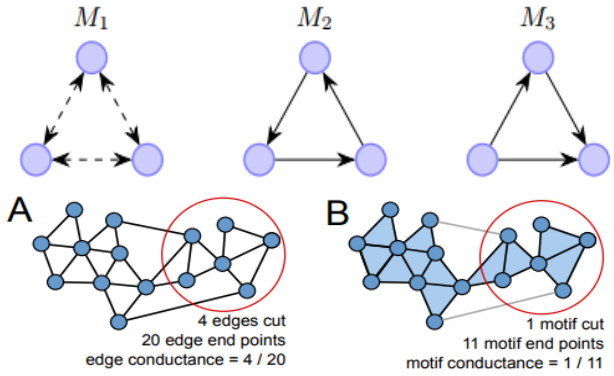


Figure 1

The method in the paper cares not only about undirected graphs, but also directed graphs. Therefore the method searches for motifs as a measure for clustering nodes together. A motif is a characteristic or property that the graph may have. Examples in figure 1 (upper) include a triangle in any direction, a cycle , and a feed forward loop. The lower half of figure one shows a "classical" clustering cut with conductance measure 4/20 (left) versus a triangle motif cut with conductance measure 1/11 (right). We want the conductance measure to essentially be zero since its a heuristic metric on how much we are separating a node to its respective cluster. The cluster found in figure 1 is due to seeding the algorithm with at least one vertex from that cluster. The triangle motif clearly uncovers the cluster better than the "classical" edge density approach. Due to the mathematical nature of the paper, it would be best now to just talk about the advantages of the method talked about in the paper and look at some results. In terms of advantages, the method runs in $\mathcal{O}(\frac{1}{\epsilon(1-\alpha)})$, where $1-\alpha$ is the probability of the random walker "teleporting" back to a seed node $u$ and $\epsilon$ is some prescribed tolerance. Essentially, since the algorithm performs local clustering rather than global, the time required to uncover some cluster $C^*$ with a given set of seeds $S^* \subseteq C^*$ requires less time than efficient global clustering algorithms. For the experiments, a planted partition model and LFR model was used with respective $F1$ scores. The planted partition model generates an undirected unweighted graph with $kn_1$ nodes. Nodes are partitioned into $k$ built-in communities, each of size $n_1$. Between any pair of nodes from the same community, an edge exists with probability $p$ and between any pair of nodes from different communities, an edge exists with probability $q$. Each edge exists independently of all other edges [1]. The LFR model also generates random

graphs with planted communities, but the model is designed to capture several properties of real-world networks with community structure such as skew in the degree and community size distributions and overlap in community membership for nodes [2]. The $F_1$ scores is a measure of the method's accuracy. The results in figure 2 compare the method in the paper using triangle motifs and some set of seeds determined algorithmically versus a "classical" clustering algorithm using edge density. the $x - axis$ is the mixing parameter $\mu$, which specifies the fraction of neighbors of a node that cross cluster boundaries. Not much improvement is made when clusters are easily identifiable ($\mu << 1$) or when clusters are highly entangled with each other ($\mu >> 0$). However, when clusters are not so intertwined or easily identifiable then the method in the paper severely outperforms classical approaches. These types of results make it suitable for consumer based clustering and allow for future work to head in this direction. Examples of this include Netflix using local clustering with seeds as user's movies and shows watched, Amazon using local clustering with using purchased items as seeds, and Spotify using local clustering with using songs listened to as seeds. As long as there are recommendation systems local clustering will have an application.

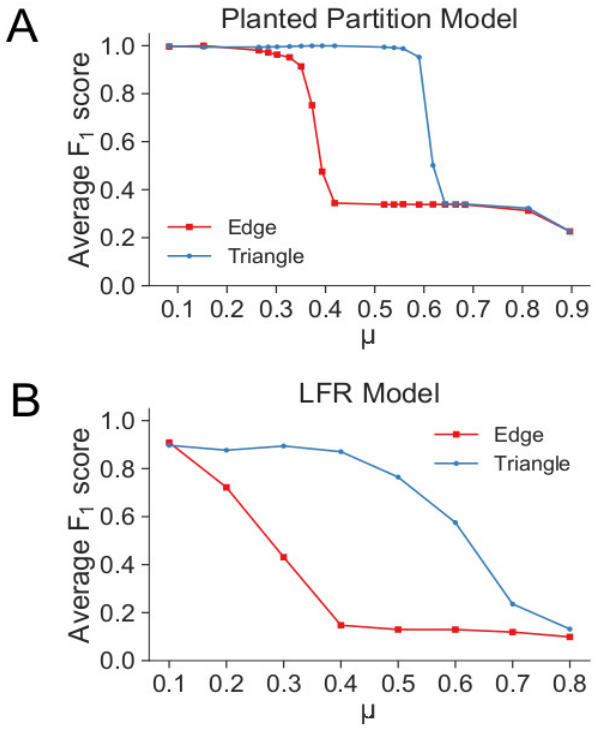


Figure 2

Three strengths I found with the paper are

1. The use of seeds allows for one to uncover a particular cluster of interest rather than clustering globally and picking them out.

2. Due to local clustering rather than global, it runs very fast and allows for scalability.

3. The use of motifs allows the user to detect clusters based on a measurement of interest.

Three weaknesses I found with the paper are

1. The results only compare the method to one other method which may imply there are "classical" global algorithms that may not have better time complexity but may have better $F1$ scores.

2. The search for good seeds is a global algorithm rather than local.

3. There may exist poor choices of motifs that may produce poor results.

Questions for the reader

1. Could the algorithm be used iteratively to cluster globally? IF so will it be faster than currently existing global clustering algorithms?

2. Why does the algorithm perform only as good as the edge clustering algorithm when $\mu << 1$ or $\mu >> 0$?

# References


[1] Hao Yin, Austin R. Benson, Jure Leskovec, David F. Gleich, *Local Higher-Order Graph Clustering*, KDD17, August 13-17, 2017, Halifax, NS, Canada.

[2] A. Lancichinetti and S. Fortunato. *Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities*. Physical Review E, 2009.