Local Higher-Order Graph Clustering

Hao Yin
Stanford University
yinh@stanford.edu

Austin R. Benson
Stanford University
arbenson@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

David F. Gleich
Purdue University
dgleich@purdue.edu

ABSTRACT

Local graph clustering methods aim to find a cluster of nodes by exploring a small region of the graph. These methods are attractive because they enable targeted clustering around a given seed node and are faster than traditional global graph clustering methods because their runtime does not depend on the size of the input graph. However, current local graph partitioning methods are not designed to account for the higher-order structures crucial to the network, nor can they effectively handle directed networks. Here we introduce a new class of local graph clustering methods that address these issues by incorporating higher-order network information captured by small subgraphs, also called network motifs. We develop the Motif-based Approximate Personalized PageRank (MAPPR) algorithm that finds clusters containing a seed node with minimal *motif conductance*, a generalization of the conductance metric for network motifs. We generalize existing theory to prove the fast running time (independent of the size of the graph) and obtain theoretical guarantees on the cluster quality (in terms of motif conductance). We also develop a theory of node neighborhoods for finding sets that have small motif conductance, and apply these results to the case of finding good seed nodes to use as input to the MAPPR algorithm. Experimental validation on community detection tasks in both synthetic and real-world networks, shows that our new framework MAPPR outperforms the current edge-based personalized PageRank methodology.

1 INTRODUCTION

The goal of graph clustering—also called community detection or graph partitioning—is to identify clusters of densely linked nodes given only the graph itself [15]. The vast majority of algorithms optimize a function that captures the edge density of the cluster (a set of nodes), for instance, conductance or modularity. Most methods for clustering are *global* and seek to cluster all nodes of the network. *Local graph clustering*—also known as seeded or targeted graph clustering—is a specific case of this problem 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. Because of its speed and scalability, this approach is frequently used in applications including ranking and community detection on the Web [13, 16], social networks [22], and bioinformatics [24]. Furthermore, the seed-based targeting is also critical to many applications. For example, in the analysis of protein-protein interaction networks, local clustering aids in determining additional members of a protein complex [45].

The theory and algorithms for local approaches are most well developed when using conductance as the cluster quality measure [2, 50]. Conductance, however, is only defined for simple undirected networks. Using principled local clustering methods for networks involving signed edges, multiple edge types, and directed interactions has remained an open challenge. Moreover, current cluster quality measures simply count individual edges and do not consider how these edges connect to form small network substructures, called network motifs. Such *higher-order connectivity structures* are crucial to the organization of complex networks [5, 35, 48], and it remains an open question how network motifs can be incorporated into local clustering frameworks. Designing new algorithms for *local higher-order graph clustering* that incorporate higher-order connectivity patterns has the potential to lead to improved clustering and knowledge discovery in networks.

There are two main advantages to local higher-order clustering. First, it provides new types of heretofore unexplored local information based on higher-order structures. Second, it provides new avenues for higher-order structures to guide seeded graph clustering. In our recent work, we established a framework that generalizes global conductance-based clustering algorithms to cluster networks based on higher-order structures [5]. However, there are multiple issues that arise when this framework is applied to local graph clustering methodologies that we address here.

Present work: Local higher-order clustering. In this paper we develop local algorithms for finding clusters of nodes based on higher-order network structures (also called *network motifs*, Figure 1). Our local methods search for a cluster (a set of nodes) S with minimal *motif conductance*, a cluster quality score designed to incorporate the higher-order structure and handle edge directions [5]. More precisely, given a graph G and a motif M, the algorithm aims to find a set of nodes S that has good motif conductance (for motif M) such that S contains a given set of seed nodes. Cluster S has good (low) motif conductance for some motif M if the nodes in S participate in many instances of M and there are few instances of M that cross the set boundary defined by S. Figure 2 illustrates the concept of motif conductance, where the idea is that we do not count the number of edges that are cut, but the number of times

KDD'17, August 13–17, 2017, Halifax, NS, Canada.

DOI: 10.1145/3097983.3098069

a given network motif M gets cut. This way edges that do not participate in a given motif (say, a triangle) do not contribute to the conductance. Motif conductance has the benefit that it allows us to focus the clustering on particular network substructures that are important for networks of a given domain. For example, triangles are important higher-order structures of social networks [19] and thus focusing the clustering on such substructures can lead to improved results.

Our main approach is to generalize Approximate Personalized PageRank (APPR) [2] to finding sets of provably small motif conductance (Theorem 4.3). The APPR method is a graph diffusion that "spreads" mass from a seed set to identify the cluster. It has an extremely fast running time, which is roughly proportional to the size of the output cluster. Our generalization, the motif-based APPR method, or MAPPR, uses a pre-processing step that transforms the original network into a weighted undirected graph where the weights depend on the motif of interest. This procedure finds all instances of the motif, but does not store the enumeration, which helps to scale to larger networks (for example, if the motif is a clique such as a triangle, no additional memory is needed by our method). We show that running APPR on this weighted network maintains the provably fast running time and has theoretical guarantees on cluster output quality in terms of motif conductance. An additional benefit of our MAPPR method is that it naturally handles directed graphs on which graph clustering has been a longstanding challenge. The original APPR method can only be used for undirected graphs, and existing local approaches for APPR on directed graphs are challenging to interpret [3].

We use MAPPR on a number of community detection tasks and show improvements over the corresponding edge-based methods. We show that using the triangle motif improves the detection of ground truth communities in synthetic networks. In addition, we identify important directed triangle motifs for recovering community structure in directed graphs.

We also show how to identify good seeds for finding local higher-order clusters when the motif is a clique. To do this, we develop a theory around the relationship between 1-hop neighborhoods, motif conductance, and a recently developed higher-order generalization of the network clustering coefficient. Essentially, we show that if the network has a large ℓth-order clustering coefficient C_ℓ, then there exists some node whose 1-hop neighborhood has small ℓ-clique conductance. We use a notion of local optimality in node neighborhood conductances to identify many good seed nodes for MAPPR and find that the resulting clusters capture global trends in the clustering structure of the graph.

In summary, our paper develops simple and flexible methods for local higher-order graph clustering with theoretical guarantees. By going beyond the old edge-based community detection objective functions, our work opens a new door to higher-order clustering and community detection problems that apply to a broad set of network clustering problems.

2 PRELIMINARIES

Before deriving our algorithms, we first go over the basic notation and cluster quality scores that we use throughout the paper. Our datasets will be simple, unweighted, possibly directed graphs $G = (V, E)$ with adjacency matrix A. We denote $n = |V|$ as the number of nodes and $m = |E|$ as the number of edges. Our algorithms will sometimes use a weighted graph $G_w = (V, E, W)$.

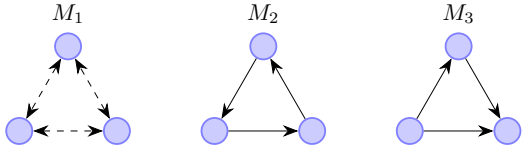

Figure 1: Examples of three directed motifs: a triangle in any direction (M_1), a cycle (M_2), and a feed-forward loop (M_3).

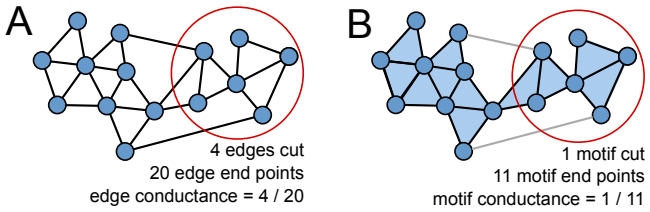

Figure 2: Illustration of (edge) conductance (A) and motif conductance (B) for the same set of nodes in the same graph, where the motif M is the triangle. Our methods finds clusters of nodes based on the motif conductance, where the user can decide which motif M to use for the clustering.

Cut, volume, and conductance. The *cut* of a set of nodes $S \subset V$, denoted by $\text{cut}(S)$, is the number of edges with one end point in S and the other end point in the complement set $\bar{S} = V \backslash S$. The *volume* of a set of nodes S, denoted by $\text{vol}(S)$, is the number of edge end points in S, i.e. $\text{vol}(S) = \sum_{u \in S} d_u$, where d_u is the degree of node u. The *conductance* of a set of nodes $S \subset V$ is

$$\phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(\bar{S}))}.$$

Figure 2A illustrates the concept. When $\text{vol}(S) \leq \text{vol}(\bar{S})$, the conductance measures the ratio of the number of edges leaving S to the number of edges in S. Note that conductance is symmetric—$\phi(S) = \phi(\bar{S})$ since $\text{cut}(S) = \text{cut}(\bar{S})$. Conductance generalizes to weighted networks where $\text{cut}(S)$ is the sum of weights cut and $\text{vol}(S)$ is the sum of weighted degrees. Small conductance indicates a good cluster, and we will use this metric (and the motif conductance defined next) for evaluating cluster quality.

Conductance is recognized as one of the most important graph clustering criterions [39] and is empirically effective at capturing ground-truth communities compared to other popular measures used in community detection [47]. Although minimizing conductance is NP-hard [46], there are approximation algorithms with theoretical guarantees for finding clusters with small conductance [2, 9]. A known issue of using conductance as a global clustering criterion is cluster imbalance, i.e., the detected clusters tend to be of uneven sizes [29, 31]. In local clustering, we seek small clusters containing a seed node, so the imbalance works in our favor.

Motif cut, motif volume, and motif conductance. Benson et al. recently generalized the cut, volume, and conductance measures to account for network motifs [5]. For this paper, we define a network motif M to be any small connected graph (such as a triangle), and an *instance* of M in a graph G is some induced subgraph H of G

that is isomorphic to M. Given a motif M, the *motif cut* of a set of nodes S, denoted by $\text{cut}_M(S)$, is the number of instances of M that have at least one end point (i.e., node) in S and at least one end point in $\bar{S}$ (Figure 2B). The *motif volume* of a set of nodes S, denoted by $\text{vol}_M(S)$ is the number of motif instance end points in S, i.e., the number of times a node is in S, counted over each node in every instance of M. The *motif conductance* for a given motif M is then

$$\phi_M(S) = \frac{\text{cut}_M(S)}{\min(\text{vol}_M(S), \text{vol}_M(\bar{S}))}.$$

In the case that M is an edge, these definitions are simply the original cut, volume, and conductance measures described above. These definitions also accommodate mixtures of motifs (e.g., triangles *and* edges) by counting over the union of instances of each motif type.

Comparing edge and motif conductance. We often compare values of motif conductance to edge conductance (see Figure 2). Although these two objective functions measure different (but related) quantities, they both represent a probability. Edge conductance is equivalently the probability that traversing a random edge adjacent to a randomly selected node from the cluster leads outside the cluster (provided that the volume of the cluster is less than half of the total graph volume). Motif conductance is the probability that a traversing to a random end point of a randomly chosen motif adjacent to a randomly selected node from the cluster leaves the cluster (provided that the motif volume of the cluster is less than half of the total graph volume). Thus, a motif conductance much smaller than an edge conductance is evidence that the higher-order structured exposed by the motif characterizes the cluster structure more clearly.

3 RELATED WORK

We now summarize some additional related work. There are a few methods for global partitioning based on motifs [4, 5, 27, 38, 43]. This paper instead focuses on local clustering methods that examine local regions of the network. There are several methods for finding clusters containing a seed node beyond the personalized PageRank method considered here, including other graph diffusions [8, 26], local spectral methods [32, 33], local modularity maximization [10], flow-based algorithms [36], and minimum degree maximization [11, 40]. All of these local methods optimize edge-based criteria to find clusters, whereas we are focused on finding clusters based on motifs. We generalize the personalized PageRank method because of the algorithm's simplicity, scalability, and theoretical underpinnings. Most related to our approach are a couple of local clustering methods based on triangle frequencies through finding k-trusses containing a seed node [20] or by greedily growing a cluster from a seed node to avoid cutting triangles [38]. In the latter method, the notion of cutting a triangle is a special case of the motif cut discussed above.

Higher-order structures (under the names motifs, graphlets, or subgraphs) are crucial in many domains including neuroscience [41], biology [37], and social networks [44]. Many of our experiments use triangle motifs, which have long been studied for their frequency in social networks [19]. The algorithmic problem of counting or estimating the frequency of various higher-order patterns has also drawn a large amount of attention [1, 6, 12, 23].

4 MOTIF CONDUCTANCE MINIMIZATION

We now develop our local higher-order clustering methodology. We begin by generalizing the Approximate Personalized PageRank (APPR) algorithm of Andersen et al. [2] to quickly find a cluster containing a given seed node with minimal motif conductance. Our algorithm has theoretical guarantees on cluster quality and running time. We then show how the motif conductance of 1-hop neighborhoods in a network can be used to identify many good seed nodes.

4.1 Motif-based Approximate Personalized PageRank (MAPPR)

We now adapt the classical Approximate Personalized PageRank (APPR) method to account for motifs. The essential idea of our approach is to transform the input graph, which is unweighted and possibly directed, into a weighted undirected graph where the weights depend on the motif [5]. We then prove that the fast Approximate Personalized PageRank method on this weighted graph will efficiently find a set with small motif conductance that contains a given seed node. We also explain how previous theoretical results are applicable to this approach, which gives us formal guarantees on running time and cluster output quality in terms of motif conductance.

Background on APPR. The personalized PageRank (PPR) vector represents the stationary distribution of a particular random walk. At each step of the random walk, with a parameter $\alpha \in (0, 1)$, the random walker "teleports" back to a specified seed node u with probability $1 - \alpha$; and with probability α, the walker performs a random walk step. The key idea is that the stationary distribution of this process for a seed node u (the PPR vector p_u) will have large values for nodes "close" to u. We can write the stationary distribution as the solution to the following system of equations $(I - \alpha P)p_u = (1 - \alpha)e_u$, where I is the identity matrix, P is the column-stochastic transition matrix representing the random walk over the graph, and e_u is the indicator vector for node u. Formally, $P = AD^{-1}$, where A is the adjacency matrix, $D = \text{diag}(Ae)$ is the diagonal degree matrix, and e is the vector of all ones.

Andersen et al. developed a fast algorithm for approximating p_u by a vector $\tilde{p}_u$ where $0 \leq D^{-1}p_u - D^{-1}\tilde{p}_u \leq \varepsilon$ component-wise [2]. To obtain a cluster with small conductance from this approximation, a *sweep* procedure is used: (i) sort the nodes by descending value in the vector $D^{-1}\tilde{p}_u$, (ii) compute the conductance of each prefix in this sorted list, (iii) output the prefix set with smallest conductance. Overall, this algorithm is fast (it runs in time proportional to the size of the output cluster) and is guaranteed to have small conductance provided that node u is in a set with small conductance. We will be more specific with the guarantees in the following section, when we derive the analogous theory for the motif-based approach.

Adapting APPR for motifs. We now propose the motif-based APPR (MAPPR) algorithm that finds a cluster with small motif conductance by finding an approximate PPR vector on a weighted graph based on the motif. Given a motif M, MAPPR has three steps: (i) construct a weighted graph W, where W_{ij} is the number of instances of M containing nodes i and j, (ii) compute the approximate PPR vector for this weighted graph, (iii) use the sweep procedure to output the set with minimal conductance. Algorithm 1 formally

Algorithm 1: Motif-PageRank-Nibble method for finding localized clusters with small motif conductance.

Input: Unweighted graph $G = (V, E)$, motif M, seed node u, teleportation parameter α, tolerance ε
Output: Motif-based cluster (set $S \subset V$)

1 $W_{ij} \leftarrow$ #(instances of M containing nodes i and j)
2 $\tilde{p} \leftarrow$ Approximate-Weighted-PPR$(W, u, \alpha, \varepsilon)$ (Algorithm 2)
3 $D_W \leftarrow \mathrm{diag}(We)$
4 $\sigma_i \leftarrow i$th smallest entry of $D_W^{-1}\tilde{p}$
5 **return** $S \leftarrow \arg\min_\ell \phi_M(S_\ell)$, where $S_\ell = \{\sigma_1, \ldots, \sigma_\ell\}$

Algorithm 2: Approximate-Weighted-PPR

Input: Undirected edge-weighted graph $G_W = (V_W, E_W, W)$, seed node u, teleportation parameter α, tolerance ε
Output: an ε-approximate weighted PPR vector $\tilde{p}$

1 $\tilde{p}(v) \leftarrow 0$ for all vertices v
2 $r(u) \leftarrow 1$ and $r(v) \leftarrow 0$ for all vertices v except u
3 $d_w(v) \leftarrow \sum_{e \in E_w : v \in e} W(v)$
4 **while** $r(v)/d_w(v) \geq \varepsilon$ *for some node* $v \in V_w$ **do**
5 /* push operation */
6 $\rho \leftarrow r(v) - \frac{\varepsilon}{2} d_w(v)$; $\tilde{p}(v) \leftarrow \tilde{p}(v) + (1-\alpha)\rho$; $r(v) \leftarrow \frac{\varepsilon}{2} d_w(v)$
7 **for** *each* $x : (v, x) \in E_w$ **do** $r(x) \leftarrow r(x) + \frac{W(v,x)}{d_w(v)} \cdot \alpha\rho$
8 **return** $\tilde{p}$

describes this method. Note that step (i) needs to be done only once, whereas steps (ii) and (iii) would be repeated for multiple subsequent runs.

This method is motivated by the following result of Benson et al. [5], which says that for motifs on three nodes, standard conductance in the weighted graph is equal to motif conductance in the original graph.

Theorem 4.1 ([5], Theorem 5). *Denote the edge conductance in a graph H by $\phi^{(H)}(S)$. Let M be a motif on three nodes, and let G_W be the weighted graph where W_{ij} is the number of instances of M in graph G containing nodes i and j. Then $\phi_M^{(G)}(S) = \phi^{(G_W)}(S)$.*

(When the motif has more than 3 nodes, the weighted graph serves as a principled heuristic for motif conductance [5].) We interpret this result for our purposes: if we can find a set with low *edge* conductance in the weighted graph using APPR, then this set will have small *motif* conductance.

The APPR method is designed for *unweighted* graphs, whereas we want to use the method for weighted graphs. Mathematically, this corresponds to replacing the column stochastic matrix P in the linear system with the column stochastic matrix $P_W = WD_W^{-1}$, where $D_W = \mathrm{diag}(We)$ is the diagonal weighted degree matrix. For the purposes of implementation, this modification is simple. We just need to change the algorithm's push method to push residual weights to neighbors proportional to edge weights (instead of evenly). We state the procedure in Algorithm 2.

For the purposes of theoretical analysis with motifs, it is important that our edge weights are integers so that we can interpret an edge with weight k as k parallel edges. Since all the analysis of APPR permits parallel edges in the graph, we can combine previous results for theoretical guarantees on Algorithm 1. The following result says that our algorithm runs in time proportional to the size of the output set.

Theorem 4.2. *Algorithm 1, after line 1, runs in $O(\frac{1}{\varepsilon(1-\alpha)})$ time, and the number of nodes with non-zero values in the output approximated PPR vector is at most $\frac{1}{\varepsilon(1-\alpha)}$.*

Proof. This follows from Andersen et al. [2, Lemma 2], where we translate G_w into a unweighted graph with parallel edges. □

Although APPR with weighted edges has used before [3, 17], there was never a runtime bound. This result is the first (albeit straightforward) theoretical bound on the runtime of APPR with weighted edges when they arise from integers.

Our next result is a theoretical guarantee on the quality of the output of Algorithm 1 in terms of motif conductance. The proof of the result follows from combining Theorem 4.1 and the analysis of Zhu et al. [50] (an improvement over the analysis of Andersen et al.). The result says that if there is some set T with small motif conductance, then there are several nodes in T for which Algorithm 1 outputs a set with small motif conductance. For notation, let η be the inverse mixing time of the random walk on the subgraph induced by T.

Theorem 4.3. *Let $T \subset V$ be some unknown targeted community we are trying to retrieve from an unweighted graph using motif M. Then there exists $T^g \subseteq T$ with $\mathrm{vol}_M(T^g) \geq \mathrm{vol}_M(T)/2$, such that for any seed $u \in T^g$, Algorithm 1 with $1 - \alpha = \Theta(\eta)$ and $\varepsilon \in [\frac{1}{10\mathrm{vol}_M(T)}, \frac{1}{5\mathrm{vol}_M(T)}]$ outputs a set S with*

$$\phi_M(S) \leq \tilde{O}\left(\min\left\{\sqrt{\phi_M(T)}, \phi_M(T)/\sqrt{\eta}\right\}\right).$$

The final piece we need to consider is the complexity of forming the weighted graph in line 1 of Algorithm 1. For ℓ-clique motifs, we use the method of Chiba and Nishizeki to compute the adjacency matrix W in $O(\ell a^{\ell-2} m)$ time, where a is the arboricity of the graph [7] and m is the number of edges in the graph. This is sufficient for the motifs considered in this paper, and there are also efficient algorithms for counting other types of motifs [34]. Note that this computation can be reused for many subsequent evaluations of the clustering algorithm for different seeds.

Towards purely local methods. Our method precomputes the number of motif instances containing each pair of nodes. Computing W is a (possibly large) upfront cost, but subsequently finding local clusters for any given seed node is fast. The graph weighting procedure could also be done locally by having the push procedure compute W_{vx} "on the fly" for all nodes x adjacent to node v. This suffices for Algorithm 2, but Algorithm 1 needs to know the total volume of the weighted graph to compute the motif conductance scores. To address this, one might use recent techniques for quickly estimating the total ℓ-clique volume on large graphs [21]. We leave the runtime analysis of this purely local method for future work.

Practical considerations. The formal theory underlying the methods (Theorem 4.3) requires multiple apriori unknowable parameters including the inverse mixing time η of the target community and the volume of the output. As practical guidance, we suggest using $\alpha = 0.98$, computing the PPR vector for $\varepsilon = 10^{-2}/\bar{d}_M$, $10^{-3}/\bar{d}_M$, $10^{-4}/\bar{d}_M$, where $\bar{d}_M = \frac{1}{n}\mathrm{vol}_M(G)$ is the average motif-degree of all

nodes, and outputting the set of best motif conductance. The reason for scaling by $\bar{d}_M$ is as follows. Theorem 4.2 bounds the running time by volume as if accessing an edge (i, j) with weight W_{ij} takes $\Theta(W_{ij})$ time (i.e., as if the edges are parallel). However, we merely need to access the value W_{ij}, which takes $O(1)$ time. Scaling ε by $\bar{d}_M$ accounts for the average effect of parallel edges present due to the weights of the motifs and permits the algorithm to do more computation with roughly the same running time guarantees.

Rather than using the global minimum in the sweep procedure in the last step of the Nibble method, we apply the common heuristic of finding the first local minimum [47]. The first local minimum is the smallest set where the PageRank vector suggests a border between the seed and the rest of the graph. It also better models the small size scale of most ground truth communities that we encounter in our experiments.

4.2 Finding many good seed nodes

So far we have proposed MAPPR to find a *single* cluster containing a *given* seed with minimal motif conductance. Now we consider the problem of how to quickly find *many* clusters with small motif conductance. Our approach will examine small network neighborhoods of nodes to identify good seeds for the targeted cluster expansion of MAPPR. We justify the use of these neighborhoods from a technical result that establishes a relationship between 1-hop neighborhoods and clusters of small motif conductance.

Informally, our key theorem is: Real world graphs with large clustering coefficients have a 1-hop neighborhood with small motif conductance for clique motifs. We can find this set since we can compute the motif conductance of all 1-hop neighborhoods efficiently. The result holds for undirected graphs only, so we will be concerned only with undirected graphs in this section. We establish our result in theory in this section and demonstrate the result in practice in Section 5.

The formal theory rests on the idea of higher-order clustering coefficients, which we developed in recent work [49] and briefly review below. As an extreme case of our theory, consider a graph with clustering coefficient 1. Then that graph will be a union of cliques, and any 1-hop neighborhood in that graph is a cluster with motif conductance of zero. The theory developed in this section relaxes this extreme setting and relates large clustering coefficients to finding node neighborhoods with small motif conductance. Our experiments in Section 5.3 show that these node neighborhoods are even better than our theory would predict.

Background on higher-order clustering coefficients. First, we introduce the definition of higher-order clustering coefficients proposed by Yin et. al [49]. The classical clustering coefficient is the fraction of wedges (length-2 paths) in the graph that are closed (i.e., induce a 3-clique, or a triangle). We can alternatively interpret each wedge as a (2-clique, adjacent edge) pair, where the "adjacent edge" shares a single node with the 2-clique (edge). The clustering coefficient is formally $C = 6|K_3|/|W|$, where K_3 is the set of 3-cliques, W is the set of (2-clique, adjacent edge) wedges, and the constant 6 comes from the fact that each 3-clique closes 6 wedges.

The generalization to higher-order clustering coefficients follows by simply looking at the fraction of (ℓ-clique, adjacent edge) pairs, or *ℓ-wedges*, that are "closed", i.e., induce an $(\ell+1)$-clique. Formally, the *ℓth-order clustering coefficient* is

$$C_\ell = (\ell^2 + \ell)|K_{\ell+1}|/|W_\ell|,$$

where $K_{\ell+1}$ is the set of $(\ell+1)$-cliques, W_ℓ is the set of ℓ-wedges, and the $(\ell^2 + \ell)$ comes from the fact that each $(\ell + 1)$-clique closes that many wedges. We can also measure local clustering with respect to a node u. Formally, the *local ℓth-order clustering coefficient* of node u is

$$C_\ell(u) = \ell|K_{\ell+1}(u)|/|W_\ell(u)|,$$

where $K_{\ell+1}(u)$ is the set of $(\ell + 1)$-cliques containing node u and $W_\ell(u)$ is the set of ℓ-wedges centered at u, i.e., the set of (ℓ-clique, adjacent edge) pairs whose intersection is node u.

Theory. Next, we state our main result of this section, which says that if the network exhibits higher-order clustering, i.e., C_ℓ is large, then there is a 1-hop neighborhood with small ℓ-clique conductance. For notation, let $N(u)$ denote the nodes in the 1-hop neighborhood of node u, i.e., $N(u) = \{v \in V \mid (u, v) \in E\} \cup \{u\}$.

Theorem 4.4. *Let graph $G = (V, E)$ have ℓth-order clustering coefficient C_ℓ. Suppose that $\mathrm{vol}_{K_\ell}(N(u)) \le \mathrm{vol}_{K_\ell}(V)/2$ for each node u. Then there exists a node $u \in V$ such that*

$$\phi_{K_\ell}(N(u)) \le \frac{1 - C_\ell}{1 - C_\ell + [C_\ell/(1 + \sqrt{1 - C_\ell})]^2} \quad (1)$$

$$\le \min\{2(1 - C_\ell), 1\}. \quad (2)$$

The bound in Theorem 4.4 is monotonically decreasing and approaches 0 as C_ℓ approaches 1. This result is a generalization of a similar statement for edge conductance [18], but prior results contain only the case of $\ell = 2$ and only use the weaker bound (2) of Theorem 4.4.

We now prove this result via several technical results relating higher-order clustering coefficients and motif conductance, where the motif is a clique. We note that many of the results are generalizations of previous theory developed by Gleich and Seshadhri [18].

The following lemma relates the higher-order clustering coefficient with neighborhood cuts, which we use to prove Theorem 4.4.

Lemma 4.5. $\sum_{v \in V} \mathrm{cut}_{K_\ell}(N(v)) \le (1 - C_\ell) \cdot |W_\ell|$.

Proof. If an ℓ-clique $(u_1, \ldots u_\ell)$ gets cut from $N(v)$, then v must directly connect with one of $u_1, \ldots, u_\ell$, say u_1. Note that the clique $(u_1, \ldots u_\ell)$ and adjacent edge (u_1, v) form an open $(\ell + 1)$-wedge since v can not connect to all of $u_1, \ldots, u_\ell$. Therefore, we have an injective map from any cut clique on the left-hand side of the inequality to an open ℓ-wedge. □

Next, we define a probability distribution on the nodes, $p_\ell(u) = |W_\ell(u)|/|W_\ell|$, which connects the global and local ℓth-order clustering coefficient in the following lemma.

Lemma 4.6. $\sum_{u \in V} p_\ell(u) C_\ell(u) = C_\ell$.

Proof.

$$\sum_{u \in V} p_\ell(u) C_\ell(u) = \sum_{u \in V} \frac{|W_\ell(u)|}{|W_\ell|} \cdot \frac{\ell \cdot |K_{\ell+1}(u)|}{|W_\ell(u)|}$$

$$= \frac{\ell}{|W_\ell|} \cdot \sum_{u \in V} |K_{\ell+1}(u)| = \frac{\ell}{|W_\ell|} \cdot (\ell + 1)|K_{\ell+1}| = C_\ell. \quad \square$$

The following lemma creates a random variable whose expectation is bounded by $1 - C_\ell$, which we use in the proof of Theorem 4.4.

Lemma 4.7. $\sum_{u \in V} p_\ell(u) \frac{\mathrm{cut}_{K_\ell}(N(u))}{|W_\ell(u)|} \le 1 - C_\ell$.

Proof. Using Theorem 4.5, $\sum_{u\in V} p_\ell(u)\frac{\text{cut}_{K_\ell}(N(u))}{|W_\ell(u)|}$
$= \frac{1}{|W_\ell|}\sum_{u\in V}\text{cut}_{K_\ell}(N(u)) \le \frac{1}{|W_\ell|}(1-C_\ell)\cdot|W_\ell| = 1-C_\ell.$ □

We are finally ready to prove our main result, and we will prove the existence using the probabilistic method. Suppose we choose a node u according to the probability distribution $p_\ell(u)$. Let $X = \text{cut}_{K_\ell}(N(u))/|W_l(u)|$ be a random variable. According to Lemma 4.7, $\mathbb{E}[X] \le 1-C_\ell$. Then for any constant $a > 1$, by Markov's inequality, we have $\mathbb{P}[X > a(1-C_\ell)] \le 1/a$. Let $b = (aC_\ell - 1)/(a-1)$, and $p = \mathbb{P}[C_\ell(u) < b]$. Now according to Lemma 4.6, we have that

$C_\ell = \sum_{C_\ell(u)<b} p_\ell(u)C_\ell(u) + \sum_{C_\ell(u)\ge b} p_\ell(u)C_\ell(u) < bp + (1-p).$

Thus $p < (1-C_\ell)/(1-b) = 1 - 1/a$. By the union bound, the probability that $\text{cut}_{K_\ell}(N(u))/|W_\ell(u)| > a(1-C_\ell)$ or $C_\ell(u) < b$ is less than 1. Therefore, there exists some node u such that $\text{cut}_{K_\ell}(N(u)) \le a(1-C_\ell)\cdot|W_\ell(u)|$ and $C_\ell(u) \ge b$. Now we show that, for this u, we have

$$\phi_\ell(N(u)) \le [1-C_\ell] \,/\, [1 - C_\ell + (aC_\ell - 1)/(a-1)]\,.$$

We first find a lower bound on $\text{vol}_{K_\ell}(N(u))$. First, each ℓ-clique cut would contribute at least one into $\text{vol}_{K_\ell}(N(u))$. Second, each $(\ell+1)$-clique in $K_{\ell+1}(u)$ uniquely corresponds to an ℓ-clique in $N(u)$ which is induced by the ℓ nodes in the $(\ell+1)$-clique other than u, thus will contribute ℓ into $\text{vol}_{K_\ell}(N(u))$. Note that each $(\ell+1)$-clique in $K_{\ell+1}(u)$ closes ℓ different ℓ-wedges in $W_\ell(u)$, and there are $C_\ell(u)|W_\ell(u)|$ closed ℓ-wedges. Therefore, by combining all the observations here, we have

$$\begin{aligned}\text{vol}_{K_\ell}(N(u)) &\ge \text{cut}_{K_\ell}(N(u)) + \ell\cdot C_\ell(u)|W_\ell(u)|/\ell \\ &\ge \text{cut}_{K_\ell}(N(u)) + b|W_\ell(u)|.\end{aligned}$$

Now combining that $\text{cut}_{K_\ell}(N(u)) \le a(1-C_\ell)\cdot|W_\ell(u)|$ and our assumption that $\text{vol}_{K_\ell}(N(u)) \le \text{vol}_{K_\ell}(\overline{N(u)})$,

$$\begin{aligned}\phi_{K_\ell}(N(u)) &= \frac{\text{cut}_{K_\ell}(N(u))}{\text{vol}_{K_\ell}(N(u))} \le \frac{\text{cut}_{K_\ell}(N(u))}{\text{cut}_{K_\ell}(N(u)) + b|W_\ell(u)|} \\ &\le \frac{a(1-C_\ell)\cdot|W_\ell(u)|}{a(1-C_\ell)|W_\ell(u)| + b|W_\ell(u)|} = \frac{1-C_\ell}{1-C_\ell+\frac{aC_\ell-1}{a(a-1)}}.\end{aligned}$$

Finally, (1) is obtained by setting $a = (1+\sqrt{1-C_\ell})/C_\ell$. □

Local minima as good seeds. Theorem 4.4 says that there must be at least one node whose 1-hop neighborhood has small motif conductance for clique motifs, provided there is higher-order clustering in the network. We use this as motivation to consider nodes whose 1-hop neighborhoods have small motif conductance as candidate seed nodes for MAPPR. Following the terminology of Gleich and Seshadhri [18], we say that a node u is a *locally minimal* if $\phi_M(N(u)) \le \phi_M(N(v))$ for all neighbors v of u. Between between 1% and 15% of nodes are local minima in the datasets we consider in Section 5.3. In that section, we verify that these local minima are in fact better seeds for MAPPR compared to random nodes, and we show that running MAPPR with all of these seeds is sufficient for reconstructing the global structure of the network.

5 EXPERIMENTS

In this section, we first evaluate the performance of our MAPPR algorithm on networks with ground truth communities or clusters, both on synthetic networks in Section 5.1 and on real-world networks in Section 5.2.[1] Our evaluation procedure of both the edge-based APPR and MAPPR is the following. For each ground truth community, we use every node as a seed to obtain a set and then pick the set with the highest F_1 score for recovering the ground truth. Next, we average of the F_1 scores over all detected communities for the detection accuracy of the method. This measurement captures how well the community can be recovered, and has previously been used to compare seeded clustering algorithms [26].

In Section 5.3, we empirically evaluate the theory of Section 4.2 for finding good seed nodes. We first show the existence of 1-hop neighborhood clusters of small motif conductances in real-world networks, and then use this idea to find seeds upon which running MAPPR will output many clusters with small motif conductance.

5.1 Recovering communities in synthetic networks with MAPPR

We first evaluate our MAPPR method for recovering ground truth in two common synthetic random graph models—the planted partition model and the LFR model. In both cases, we find that using triangle motifs increases the range of parameters in which we are able to recover the ground truth communities.

Planted partition model. 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.

In our experiment, we examine the behavior of MAPPR and the edge-based APPR methods by fixing parameters $n_1 = 50$, $k = 10$, $p = 0.5$, and takings different values of q such that the community mixing level $\mu = [(k-1)q]/[p+(k-1)q]$, which measures the fraction of neighbors of a node that cross cluster boundary, varies from 0.1 to 0.9. For each value of μ, we computed the average of the "mean best" F_1 score described above over 20 random instances of the graph. For MAPPR, we used the triangle motif. We are motivated in part by recent theoretical results of Tsourakakis et al. showing that with high probability, the triangle conductance of a cluster in the planted partition model is smaller than the edge conductance [43]. Here we take an empirical approach and study recovery instead of conductance.

Figure 3A illustrates the results. Using triangles with MAPPR significantly outperforms the edge-based APPR method when $\mu \in [0.4, 0.6]$. In this regime, for any given node, the expected number of intra-community edges and inter-community edges is roughly the same. Thus, the edge-based method degrades in performance. However, the number of intra-community triangles remains greater than the number of inter-community triangles, so the triangle-based method is able to recover the planted partition.

LFR model. 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 [28]. For our purposes, the

[1] As part of this paper, real-world datasets and implementations of the MAPPR algorithms are available at http://snap.stanford.edu/mappr.

Table 1: Recovery of ground truth community structure in undirected graphs using edge-based and motif-based APPR (MAPPR), where the motif is the triangle. Bold numbers denote better recovery or smaller conductance with 5+% relative difference. F_1 score, precision, and recall are all averages over the 100 ground truth communities.

				F_1 score		Precision		Recall		Motif conductance	
Network	$\|V\|$	$\|E\|$	# comms. (sizes)	edge	triangle	edge	triangle	edge	triangle	edge	triangle
com-DBLP	317K	1.05M	100 (10–36)	0.264	0.269	0.342	**0.366**	0.310	**0.329**	0.393	0.384
com-Amazon	335K	926K	100 (10–178)	**0.620**	0.556	0.634	0.660	**0.704**	0.567	0.163	**0.065**
com-Youtube	1.13M	2.99M	100 (10–200)	0.140	**0.165**	0.233	**0.390**	0.147	**0.188**	**0.536**	0.739
com-LiveJournal	4.00M	34.7M	100 (10–10)	0.255	**0.274**	0.216	**0.280**	0.606	**0.672**	0.498	**0.409**
com-Orkut	3.07M	117M	100 (10–200)	0.063	**0.078**	0.072	**0.117**	**0.212**	0.166	0.702	**0.510**
com-Friendster	65.6M	1.81B	100 (10–191)	0.095	**0.114**	0.103	**0.158**	0.204	**0.234**	0.747	**0.622**

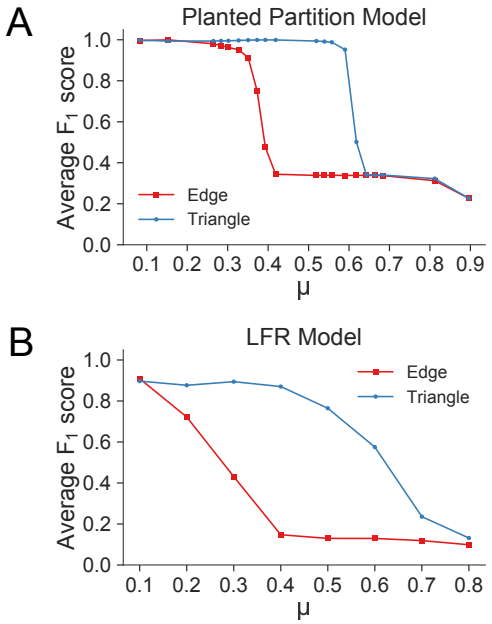

Figure 3: Average F_1 score on detected clusters in the planted partition model (A) and LFR model (B) as a function of the mixing parameter μ that specifies the fraction of neighbors of a node that cross cluster boundaries. We use the classical edge-based APPR and our triangle-based MAPPR to recover ground truth clusters. There is a large parameter regime where the triangle-based approach significantly outperforms the edge-based approach in both models.

most important model parameter is the mixing parameter μ, which is the fraction of a node's edges that connect to a node in another community. We fix the other parameters as follows: $n = 1000$ is the number of nodes, where 500 nodes belong to 1 community and 500 belong to 2; the number of communities is randomly chosen between 43 and 50; the average degree is 20; and the community sizes range from 20 to 50.

We again use the edge-based APPR method and MAPPR with the triangle motif. Figure 3B shows the results. The performance of the edge-based method decays as we increase the mixing parameter μ from 0.1 to 0.4, while the triangle-based method maintains an F_1 score of approximately 0.9 in this regime. For mixing parameters as large as 0.6, the F_1 score for MAPPR is still three times larger than that of the edge-based method, and throughout nearly the entire parameter space, using triangles improves performance.

To summarize, incorporating triangles into personalized PageRank dramatically improves the recovery of ground truth community structure in synthetic models. In the next section, we run experiments on both undirected and directed real-world networks.

5.2 Recovering communities in real-world networks with MAPPR

We now compare the edge- and motif-based APPR methods on real-world networks with ground truth communities. Although these graphs have as many as 1.8 billion edges, the APPR method takes at most a few seconds per seed once the graph is in memory.

Undirected graphs. We analyzed several well-known networks with ground truth community structure constructed from Web data: com-Amazon, com-DBLP, com-Youtube, com-LiveJournal, com-Orkut, and com-Friendster [47]. For each network, we examined 100 communities whose sizes ranged between 10 and 200 nodes. Summary statistics of the datasets and our experiment are in Table 1. In 5 out of 6 networks, MAPPR achieves a higher F_1 score than edge-based APPR. In 3 of the 5 networks, the F_1 score provides a relative improvement of over 5%. In all 6 networks, the average precision of the recovered clusters is larger, and in 4 of these networks, the change is greater than 5%. We suspect this arises from triangles encouraging more tight-knit clusters. For example, dangling nodes connected by one edge to a cluster are ignored by the triangle-based method, whereas such a node would increase the edge-based conductance of the set. In 4 of the 6 networks, recall in the triangle-based method provides relative improvements of at least 5%.

Directed graphs. A major advantage of MAPPR is that it easily handles directed graphs; we simply need to specify the directed motifs. Here, we use the three different directed triangle motifs in Figure 1 (M_1, the undirected triangle; M_2, the cycle; and M_3, the feed-forward loop). We form our motif weighted matrix W with respect to general subgraphs (i.e., *not* induced subgraphs). Thus, a triangle with all six directed edges contains 1 instance of motif M_1, 2 instances of motif M_2, and 2 instances of motif M_3.

We analyze two directed networks. The first is an e-mail network between members of a European research institution (email-Eu), where department membership of researchers are the ground truth

Table 2: Recovery of ground truth communities in directed graphs using edge-based and motif-based APPR for the three triangular motifs in Figure 1. Bold numbers denote (i) cases where a motif-based method's score is a 5+% relative improvement over the edge-based method and (ii) cases where the edge-based method out-performs all 3 motif-based methods by 5+%.

				F_1 score				Precision				Recall			
Network	$\|V\|$	$\|E\|$	# comms. (sizes)	edge	M_1	M_2	M_3	edge	M_1	M_2	M_3	edge	M_1	M_2	M_3
EMAIL-EU	1.00K	25.6K	28 (10–109)	0.398	**0.496**	**0.443**	**0.483**	0.502	**0.580**	**0.605**	**0.660**	**0.754**	0.685	0.577	0.594
WIKI-CATS	1.79M	28.5M	100 (21–192)	0.239	0.246	0.234	0.231	0.334	**0.368**	**0.391**	**0.360**	**0.377**	0.327	0.226	0.328

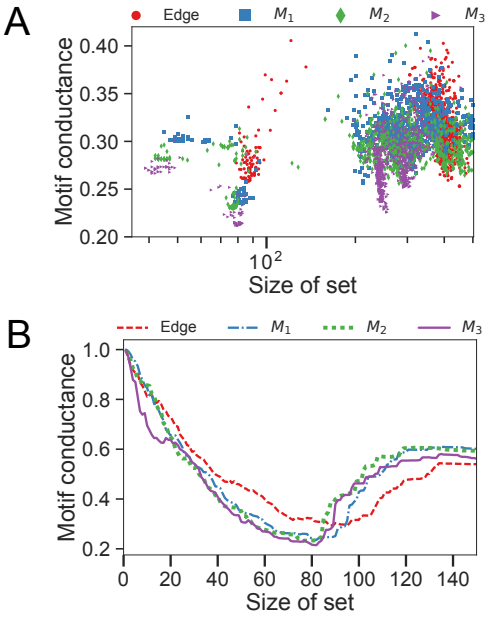

Figure 4: A: Distribution of set size and conductance from using each node in EMAIL-EU as a seed for edge-based APPR and motif-based APPR (MAPPR) with the three motifs in Figure 1. Smaller edge-based cluster concentrate in sizes of 70–100, a regime where also containing many motif-based clusters with smaller conductance. B: Sweep profile for a single seed in EMAIL-EU for edge and the same three motifs. The shape of the curves is similar, but the minima for the 3 motif-based curves occur for smaller set sizes and have smaller motif conductances compared to the curve for edges.

communities. The second network is the English Wikipedia hyperlink network (WIKI-CATS), where the article categories are the ground truth communities (we only consider 100 categories for our analysis). The datasets and recovery results are summarized in Table 2. For both networks, using motif M_1 provides an improvement in F_1 score over the edge-based method. The improvement is drastic in EMAIL-EU (25% relative improvement). In fact, all three motifs lead to substantial improvements in this network. We also see that in both networks, the motifs provide additional precision but sacrifice recall. These tighter clusters are expected for the same reasons as for the undirected networks.

We investigate the results for EMAIL-EU in more detail, as the use of motifs dramatically improves the recovery of ground truth clusters with respect to F_1 score. First, we used every node in the network as a seed for the APPR methods with edges and the three motifs (Figure 4A). The clusters bifurcate into small (< 100 nodes) and large (> 200 nodes) sizes. For the small clusters, the edge-based ones concentrate in sizes of 70–100. In this range, there are several clusters with much smaller motif-based conductance for all three motifs. This provides evidence that the 3 motifs are better models for the community structure in the network. We also see that of the large clusters, the edge-based ones tend to be the largest. Since these sets are larger than the sizes of the communities in the network, this observation provides evidence for why precision is better when using triangle motifs with MAPPR.

Next, we examined the sweep profile for a single seed node in the EMAIL-EU network (Figure 4B). The sweep profile highlights key differences between the output of the motif-based and edge-based algorithms. Although the general shape of the sweep profile is the same across the 3 motifs and edges, the minimum of the curves occurs for a smaller set and at a smaller conductance value for the motifs. A plausible explanation is that the edge-based and motif-based APPR methods are capturing roughly the same set, but the constraint of triangle participation excludes some nodes. The smaller motif conductance values indicate that these motifs are better models for the cluster structure in the network.

5.3 Finding many good seed nodes

We now empirically analyze the theory of Section 4.2. The goal of our experiments here is (i) to demonstrate that there are 1-hop neighborhood clusters of small motif conductance as a test of how well Theorem 4.4 holds in practice, and (ii) to use this idea to quickly find many clusters with minimal motif conductance by running targeted cluster expansion around a subset of the 1-hop neighborhood clusters. Regarding (i), we find that real-world networks exhibit much better results than predicted by the theory and the 1-hop neighborhood with minimal motif conductance is competitive with spectral graph theory approaches. Regarding (ii), we show that locally minimal nodes are better seeds than random nodes. We use this insight to find the global structure of clique conductance clusters more quickly than exhaustive enumeration.

We evaluate 1-hop neighborhood cluster quality in terms of motif conductance for 2-clique (edge), 3-clique (triangle), and 4-clique motifs using four undirected networks where we can exhaustively sample targeted clusters easily: CA-CONDMAT, a co-authorship network constructed from papers posted to the condensed matter category on arXiv [30]; FB-HARVARD1, a snapshot of the friendships network between Harvard students on Facebook in September 2005 [42]; EMAIL-ENRON, an e-mail communication network of the employees of Enron Corporation and their contacts [25]; and WEB-GOOGLE, a Web graph released by Google for a programming contest [31]. Summary statistics for the networks are in Table 4.

Table 3: Comparison of Fiedler clusters and the best 1-hop neighborhood clusters in terms of motif conductance. A star (⋆) denotes when 1-hop neighborhood clusters are the same across different clique sizes, a dagger (†) denotes when Fiedler clusters are the same, and a bullet (•) denotes when the neighborhood and Fiedler clusters are the same.

	M = 2-cliques (edges)				M = 3-cliques (triangles)				M = 4-cliques			
	Neighborhood		Fiedler		Neighborhood		Fiedler		Neighborhood		Fiedler	
	$\|S\|$	$\phi_M(S)$	$\|S\|$	$\phi_M(S)$	$\|S\|$	$\phi_M(S)$	$\|S\|$	$\phi_M(S)$	$\|S\|$	$\phi_M(S)$	$\|S\|$	$\phi_M(S)$
CA-CONDMAT	$23^{\bullet}$	$1.1\cdot10^{-2}$	$23^{\bullet}$	$1.1\cdot10^{-2}$	$17^{\star}$	$2.0\cdot10^{-3}$	$18^{\dagger}$	$1.5\cdot10^{-3}$	$17^{\star}$	$1.4\cdot10^{-4}$	$18^{\dagger}$	$1.1\cdot10^{-4}$
EMAIL-ENRON	$21^{\star}$	$2.4\cdot10^{-2}$	$25^{\dagger}$	$1.7\cdot10^{-2}$	$21^{\star}$	$1.3\cdot10^{-2}$	$25^{\dagger}$	$6.7\cdot10^{-3}$	8	$3.6\cdot10^{-3}$	$25^{\dagger}$	$1.6\cdot10^{-3}$
FB-HARVARD1	4	$2.0\cdot10^{-1}$	1,470	$1.1\cdot10^{-1}$	$5^{\star}$	$1.4\cdot10^{-1}$	1,429	$4.9\cdot10^{-2}$	$5^{\star}$	$7.7\cdot10^{-2}$	1,299	$2.2\cdot10^{-2}$
WEB-GOOGLE	203	$1.5\cdot10^{-3}$	5,335	$9.2\cdot10^{-5}$	$62^{\star}$	$1.8\cdot10^{-4}$	10,803	$5.9\cdot10^{-5}$	$62^{\star}$	$1.1\cdot10^{-5}$	8,001	$2.2\cdot10^{-6}$

Table 4: Summary statistics of datasets. The C_ℓ are the higher-order clustering coefficients [49].

Dataset	$\|V\|$	$\|E\|$	C_2	C_3	C_4
CA-CONDMAT	15,147	75,623	0.30	0.25	0.26
EMAIL-ENRON	18,561	156,139	0.11	0.06	0.05
FB-HARVARD1	13,319	793,410	0.14	0.07	0.07
WEB-GOOGLE	393,582	2,905,337	0.07	0.06	0.07

1-hop neighborhoods have small motif conductance. Plugging the higher-order clustering coefficients from Table 4 into the bound from Theorem 4.4 yields weak, albeit non-trivial bounds on the smallest neighborhood conductance (all bounds are ≥ 0.9 for the networks we consider). However, the spirit of the theorem rather than the bound itself motivates our experiments: with large higher-order clustering, there should be a neighborhood with small clique conductance. We indeed find this to be true in our results.

Table 3 compares the neighborhood with smallest motif conductance for the 2-clique, 3-clique, and 4-clique motifs with the Fiedler cluster obtained by a sweep procedure on the second eigenvector of the normalized Laplacian matrix [14]. Here, the Fiedler cluster represents a method that uses the global structure of the network to compare against the local neighborhood clusters. In all cases, the best neighborhood cluster has motif conductance far below the upper bound of Theorem 4.4. For all clique orders, the best neighborhood cluster always has conductance within a factor of 3.5 of the Fiedler cluster in CA-CONDMAT, EMAIL-ENRON, and FB-HARVARD1. With WEB-GOOGLE, the conductances are much smaller but the best neighborhood still has conductance within an order of magnitude of the Fiedler set. We conclude that the best neighborhood cluster in terms of conductance, which comes from purely local constructs, is competitive with the Fiedler vector that takes into account the global graph structure. This motivates our next set of experiments that uses nodes that induce small neighborhood conductance as seeds for the APPR method.

Local minima are good seeds. So far, we have used our theory to find a single node whose 1-hop neighborhood has small motif conductance for clique motifs. We examine this further by using nodes whose neighborhoods induce good clusters as seeds for MAPPR. Recall that we defined a node u to be locally minimal if $\phi_M(N(u)) \leq \phi_M(N(v))$ for all neighbors v of u. To test whether local minima are good seeds for APPR, we first exhaustively computed MAPPR clusters using every node in each of our networks as a seed. Next, we used a one-sided Mann Whitney U test to test the null hypothesis that the local minima yield APPR clusters with

Table 5: The p-values from Mann-Whitney U tests of the null hypothesis that the motif conductances of sets from MAPPR seeded with local minima are *not less than* the motif conductances of sets from MAPPR seeded with non-local minima. In all but CA-CONDMAT with the standard edge motif, we reject the null at a significance level < 0.003.

motif	CA-CONDMAT	EMAIL-ENRON	FB-HARVARD1	WEB-GOOGLE
edge	0.87	$< 1\cdot10^{-16}$	$8.17\cdot10^{-05}$	$< 1\cdot10^{-16}$
triangle	$2.07\cdot10^{-03}$	$< 1\cdot10^{-16}$	$4.32\cdot10^{-04}$	$< 1\cdot10^{-16}$
4-clique	$7.20\cdot10^{-10}$	$< 1\cdot10^{-16}$	$1.55\cdot10^{-05}$	$< 1\cdot10^{-16}$

motif conductances that are *not less than* motif conductances from using non-local minima as seeds (Table 5). The p-values from these tests say that we can safely reject the null hypothesis at significance level < 0.003 for all cliques and networks considered except for 2-cliques in CA-CONDMAT. In other words, local minima are better seeds than non-local minima.

Finally, we use these local minimum seeds to construct network community profile (NCP) plots for different motifs. NCP plots are defined as the optimal conductance over all sets of a fixed size k as a function of k [31]. The shapes of the curves reveal the cluster structure of the networks. In practice, these plots are generated by exhaustively using every node in the network as a seed for the APPR method [31]. Here, we compare this approach to two simpler ones: (i) using the neighborhood sizes and conductances and (ii) using only local minima as seeds for APPR. In the latter case, between 1% and 15% of nodes are local minima, depending on the network, so this serves as an economical alternative to the typical exhaustive approach.

Figure 5 shows the NCP plots for CA-CONDMAT and FB-HARVARD1 with the triangle and 4-clique motifs. Seeding with local minima is sufficient for capturing the major trends of the NCP plot. In general, the curves constructed from neighborhood information capture the first downward spike in the plot, but do not capture larger sets with small conductance. Finally, the triangle and 4-clique NCP plots are quite similar for both networks. Thus, we suspect that local minima for lower-order cliques could also be used as good seeds when looking for sets based on higher-order cliques.

6 DISCUSSION

Our work enables fast local clustering of graphs in terms of rich, higher-order structures with theoretical guarantees on cluster quality. Our method is also an effective technique for finding clusters in directed graphs, a common data type with relatively few analytic

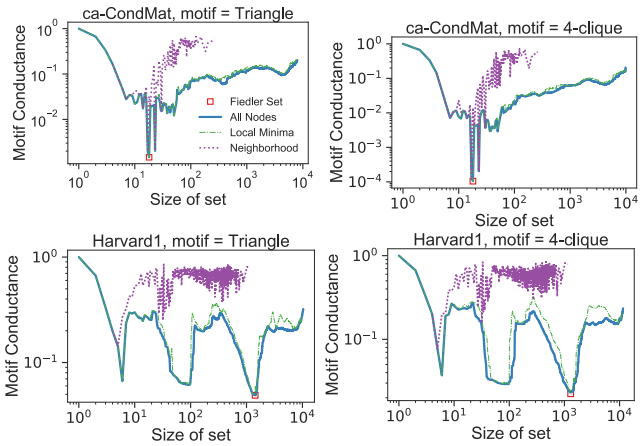

Figure 5: NCP plots for two networks with two different clique sizes. Curves are constructed from MAPPR with all nodes as seeds (blue), MAPPR with just local minima as seeds (green), and all 1-hop neighborhoods (purple). Using local minima as seeds captures the trends of exhaustive PPR using only a fraction of the seeds.

tools, and we found that using directed triangle motifs provided substantial improvements in recovery of communities in a directed e-mail network. We also found triangles critical for recovery in common synthetic models. Lastly, we computed local motif-based clusters for clique motifs through 1-hop neighborhoods and found the centers of 1-hop neighborhoods with small motif conductance to be good seeds. Neighborhoods also revealed correlations between cliques of different orders—in several cases, the same neighborhood has the smallest motif conductance for different clique motifs. Exploring this structure is an interesting avenue for future research.

ACKNOWLEDGMENTS

This research has been supported in part by NSF CCF-1149756, NSF IIS-1422918, NSF IIS-1546488, NSF Center for Science of Information STC CCF-093937, NSF IIS-1149837, Sloan Research Fellowship, DARPA SIMPLEX, Stanford Data Science Initiative, Chan Zuckerberg Biohub, Tencent, and Huawei.

REFERENCES

[1] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield. Efficient graphlet counting for large networks. In *ICDM*, 2015.
[2] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *FOCS*, 2006.
[3] R. Andersen, F. Chung, and K. Lang. Local partitioning for directed graphs using PageRank. *Internet Mathematics*, 2008.
[4] A. R. Benson, D. F. Gleich, and J. Leskovec. Tensor spectral clustering for partitioning higher-order network structures. In *SDM*, 2015.
[5] A. R. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 2016.
[6] M. Bressan, F. Chierichetti, R. Kumar, S. Leucci, and A. Panconesi. Counting graphlets: Space vs time. In *WSDM*, 2017.
[7] N. Chiba and T. Nishizeki. Arboricity and subgraph listing algorithms. *SIAM Journal on Computing*, 1985.
[8] F. Chung and O. Simpson. Solving linear systems with boundary conditions using heat kernel pagerank. In *WAW*, 2013.
[9] F. R. Chung. Four proofs for the cheeger inequality and graph partition algorithms. In *ICCM*, 2007.
[10] A. Clauset. Finding local community structure in networks. *Phys. Rev. E*, 72(2), 2005.
[11] W. Cui, Y. Xiao, H. Wang, and W. Wang. Local search of communities in large graphs. In *SIGMOD*, 2014.
[12] L. De Stefani, A. Epasto, M. Riondato, and E. Upfal. Trièst: Counting local and global triangles in fully-dynamic streams with fixed memory size. In *KDD*, 2016.
[13] A. Epasto, J. Feldman, S. Lattanzi, S. Leonardi, and V. Mirrokni. Reduce and aggregate: similarity ranking in multi-categorical bipartite graphs. In *WWW*, 2014.
[14] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical J.*, 1973.
[15] S. Fortunato. Community detection in graphs. *Physics Reports*, 2010.
[16] U. Gargi, W. Lu, V. Mirrokni, and S. Yoon. Large-scale community detection on youtube for topic discovery and exploration. In *ICWSM*, 2011.
[17] D. F. Gleich and M. W. Mahoney. Using local spectral methods to robustify graph-based learning algorithms. In *KDD*, 2015.
[18] D. F. Gleich and C. Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *KDD*, 2012.
[19] M. S. Granovetter. The strength of weak ties. *American J. of Sociology*, 1973.
[20] X. Huang, H. Cheng, L. Qin, W. Tian, and J. X. Yu. Querying k-truss community in large and dynamic graphs. In *SIGMOD*, 2014.
[21] S. Jain and C. Seshadhri. A fast and provable method for estimating clique counts using turán's theorem. In *WWW*, 2017.
[22] L. G. S. Jeub, P. Balachandran, M. A. Porter, P. J. Mucha, and M. W. Mahoney. Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. *Physics Review E*, 91, 2015.
[23] M. Jha, C. Seshadhri, and A. Pinar. Path sampling: A fast and provable method for estimating 4-vertex subgraph counts. In *WWW*, 2015.
[24] B.-B. Jiang, J.-G. Wang, Y. Wang, and J. Xiao. Gene prioritization for type 2 diabetes in tissue-specific protein interaction networks. *Systems Biology*, 2009.
[25] B. Klimt and Y. Yang. Introducing the Enron corpus. In *CEAS*, 2004.
[26] K. Kloster and D. F. Gleich. Heat kernel based community detection. In *KDD*, 2014.
[27] C. Klymko, D. F. Gleich, and T. G. Kolda. Using triangles to improve community detection in directed networks. In *ASE BigData Conference*, 2014.
[28] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 2009.
[29] K. J. Lang. Fixing two weaknesses of the spectral method. In *NIPS*, 2005.
[30] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 2007.
[31] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 2009.
[32] Y. Li, K. He, D. Bindel, and J. E. Hopcroft. Uncovering the small community structure in large networks: A local spectral approach. In *WWW*, 2015.
[33] M. W. Mahoney, L. Orecchia, and N. K. Vishnoi. A local spectral method for graphs: With applications to improving graph partitions and exploring data graphs locally. *JMLR*, 2012.
[34] D. Marcus and Y. Shavitt. Efficient counting of network motifs. In *ICDCSW*, 2010.
[35] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 2002.
[36] L. Orecchia and Z. A. Zhu. Flow-based algorithms for local graph clustering. In *SODA*, 2014.
[37] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 2007.
[38] K. Rohe and T. Qin. The blessing of transitivity in sparse and stochastic networks. *arXiv:1307.2302*, 2013.
[39] S. E. Schaeffer. Graph clustering. *Computer Science Rev.*, 2007.
[40] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *KDD*, 2010.
[41] O. Sporns and R. Kötter. Motifs in brain networks. *PLOS Biology*, 2004.
[42] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 2012.
[43] C. Tsourakakis, J. Pachocki, and M. Mitzenmacher. Scalable motif-aware graph clustering. In *WWW*, 2017.
[44] J. Ugander, L. Backstrom, and J. Kleinberg. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *WWW*, 2013.
[45] K. Voevodski, S.-H. Teng, and Y. Xia. Spectral affinity in protein networks. *BMC Systems Biology*, 2009.
[46] D. Wagner and F. Wagner. Between min cut and graph bisection. In *MFCS*, 1993.
[47] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 2015.
[48] Ö. N. Yaveroğlu, N. Malod-Dognin, D. Davis, Z. Levnajic, V. Janjic, R. Karapandza, A. Stojmirovic, and N. Pržulj. Revealing the hidden language of complex networks. *Scientific Reports*, 2014.
[49] H. Yin, A. R. Benson, and J. Leskovec. Higher-order clustering in networks. *arXiv:1704.03913*, 2017.
[50] Z. A. Zhu, S. Lattanzi, and V. S. Mirrokni. A local algorithm for finding well-connected clusters. In *ICML*, 2013.