Scalable Motif-aware Graph Clustering

Charalampos E.
Tsourakakis
BU & Harvard University
ctsourak@bu.edu

Jakub Pachocki
Harvard University
meret@seas.harvard.edu

Michael Mitzenmacher
Harvard University
michaelm@eecs.harvard.edu

ABSTRACT

We develop new methods based on graph motifs for graph clustering, allowing more efficient detection of communities within networks. We focus on triangles within graphs, but our techniques extend to other clique motifs as well. Our intuition, which has been suggested but not formalized similarly in previous works, is that triangles are a better signature of community than edges. We therefore generalize the notion of conductance for a graph to *triangle conductance*, where the edges are weighted according to the number of triangles containing the edge. This methodology allows us to develop variations of several existing clustering techniques, including spectral clustering, that minimize triangles split by the cluster instead of edges cut by the cluster. We provide theoretical results in a planted partition model to demonstrate the potential for triangle conductance in clustering problems. We then show experimentally the effectiveness of our methods to multiple applications in machine learning and graph mining.

Keywords

Graph Clustering; Large-scale Graph Mining; Expanders; Community detection

1. INTRODUCTION

Our work is motivated by the following question: *how can we effectively leverage higher-level graph structures, or motifs, for better clustering and community detection in graph structures*? Network motifs are basic interaction patterns that recur throughout networks, much more often than in random networks. We focus here on triangle subgraphs, which have often been suggested as being stronger signals of community structure than edges alone [42]. The use of motifs has been leveraged already in the context of dense subgraph discovery [17], see [27, 37]. For example, social networks tend to be abundant in triangles, since typically friends of friends tend to become friends themselves [41]. Triangles are also important motifs in brain networks [34].

In other networks, such as gene regulation networks, feed-forward loops and bi-fans are known to be significant patterns of interconnection [25], but our techniques extend to other such motifs as well. Despite the intuition that triangles or other structures may be important for clustering and related graph problems [9, 21, 32], there appears to be a gap in terms of useful formalizations of this idea. Our main contribution is a natural and simple formal framework based on generalizing conductance and related notions such as graph expansion, based on reweighting edges according to the number of triangles that contain the edge.

Remark. Recently, Benson, Gleich, and Leskovec published an article in *Science* [10] that proposes the same reweighting framework as ours. Our work [36] and the *Science* paper [10] appeared independently at the same time and share the algorithmic contribution of performing efficiently motif-based clustering on the input graph without constructing a hypergraph whose hyperedges correspond to motifs. In this paper, we have decided to focus on important contributions of our work that do not appear in [10]: a random walk interpretation of the graph reweighting scheme, that provides a principled approach to define the notion of conductance for other motifs; the framework of motif-based graph expanders that provides the theoretical foundations for motif-based graph clustering; our results on the planted partition model; the introduction of a natural heuristic that outperforms a wide variety of popular graph community detection methods, both in terms of output quality and run times; and an experimental evaluation on real-world networks with ground-truth communities.

Contributions. Specifically, our contributions are summarized as follows:

- We formalize intuitions and heuristics in prior work by studying *triangle conductance*, a variation of graph conductance based on triangles. Our definitions generalize to other motifs, but here we focus on triangles. In contrast to prior work [9, 10], we relate the notion of triangle conductance to appropriate random walks on the graph and to a generalization of graph expansion based on triangles instead of edges. When at node u we choose a triangle that u participates in uniformly at random and then choose an endpoint of that triangle, other than u, uniformly at random. We differentiate our new concepts by for example showing that an expander graph [5] is not necessarily a triangle expander and vice versa.

- We provide approximation algorithms for a generalization of the well-studied sparsest cut problem [39], where

the goal now is to minimize the number of triangles cut by a partition. We present this part of our work briefly as it coincides with the algorithmic contribution of the *Science* paper [10].

- We study our reweighting algorithm in the planted partition model, where we provide tight theoretical guarantees on its ability to recover the true graph partition with high probability[1].

- We propose a highly effective heuristic method for detecting communities. Specifically, using publicly available datasets where ground-truth is available, we verify the effectiveness of our framework, and show it takes orders of magnitude less time and obtains similar performance to the best performing competitor Markov clustering (MCL) [14].

Before beginning, we show that our scheme reweighting edges by triangle counts provides significant insights on the community structure of real-world networks. Surprisingly, in many real-world networks we find this simple step immediately disconnects the graph into numerous non-trivial connected components, that we refer as triangle components. Figure 1 shows the distribution of triangle components for the AMAZON, DBLP, and YOUTUBE networks (see Table 1 for a detailed description). Our findings are consistent across all of them: there exists one giant triangle component and then a large number of triangle components with up to few hundreds of nodes. (Trivially all degree one nodes in the original graph become isolated components.)

These findings agree with the "jellyfish" or "octopus" model [35], according to which most networks have a giant "core" with a large number of relatively small "whiskers" dangling around. Furthermore, our findings agree with the findings of [23] that claim that communities have size up to roughly 100 nodes. Our findings show additionally to [23] that no triangles are split between whiskers and the rest of the graph. We generalize this idea for our clustering results and experiments.

Roadmap. Section 2 briefly presents related work. Section 3 presents our algorithmic contributions, and Section 4 studies the performance of our proposed methods and various competitors on graphs with community ground-truth available. Section 5 sets the theoretical foundations for motif-based community detection.

Notation. We use the following notation throughout the paper. Let $G(V, E, w)$ be an undirected graph with non-negative weights; we also use $G(V, E)$ for unweighted graphs. The weighted degree $deg(u)$ of a node $u \in V$ is equal to $deg(i) = \sum_{j \in V} w(i, j)$. For a set of nodes $S \subseteq V$ we define $w(S : \bar{S}) = \sum_{i \in S, j \in \bar{S}} w(i, j)$ as the total weight of the edges leaving S. Also let $vol(S) = \sum_{i \in S} deg(i)$ be the volume of S. For the case of unweighted graphs, we denote $w(S : \bar{S})$ as $e(S : \bar{S})$ for clarity, and we define $t(u), t(u, v)$ as the total number of triangles that contain node u and edge $(u, v) \in E(G)$ respectively. Notice for unweighted graphs graphs, $vol(S) = 2e(S) + e(S : \bar{S})$ where $e(S)$ is the number of edges induced by S.

---

[1]An event A_n holds with high probability (whp) if $\lim_{n \to +\infty} \mathbf{Pr}[A_n] = 1$.

2. RELATED WORK

Communities. Intuitively, a community is a set of nodes with more and/or better intra- than inter-connections. There are different approaches to defining the notion of a community that lead to different mathematical formalizations. For instance, the notion of modularity captures the difference between the connectivity structure of a set of nodes compared the expected structure if edges in the graph were distributed at random [28]. Conductance is one of the most popular measures used in community detection [15, 23, 33, 43]. It quantifies the intuition that the total weight of edges leaving the community should be relatively small compared to the internal weight. It is worth outlining that this intuition is not always true [3]. Specifically, there exist networks with communities whose outgoing number of edges *is not* small compared to the number of internal edges. The notions of k-clique communities [13], i.e., the union of all cliques of size k that can be reached through adjacent k-cliques that share $k - 1$ nodes, and (α, β)-communities [26] have been proposed to tackle communities whose outgoing number of edges *is not* small compared to the number of internal edges.

We formally define graph conductance. For any set $S \subseteq V$ we define its expansion, also known as conductance, by

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

The edge expansion of the graph, also known as graph conductance, is defined as $\phi(G) = \min_S \phi(S)$.

Given a connected graph G, finding cuts with minimum conductance is **NP**-hard. A lot of work has focused on developing approximation algorithms [7, 22, 6]. As noted in numerous works, cf.[23], spectral clustering is considered to be the most practical approach.

Spectral clustering. Cheeger's inequality establishes a bound on edge expansion via the spectrum of the normalized Laplacian matrix representation of the graph. Specifically, let A be the adjacency matrix of G, and D a diagonal matrix containing the weighted degrees in its diagonal. The combinatorial Laplacian is defined as $L = D - A$. The normalized Laplacian is $\mathcal{L} = D^{-1/2} L D^{-1/2}$. It is well-known that the multiplicity of the zero eigenvalue of $\mathcal{L}$ equals the number of connected components of G. Let us assume without any loss of generality that G is connected, hence only one eigenvalue of $\mathcal{L}$ equals 0. The following theorem forms the basis of spectral graph theory.

THEOREM 1 (DISCRETE CHEEGER'S INEQUALITY [6]). *Given a weighted undirected graph $G(V, E, w)$ and its normalized Laplacian matrix $\mathcal{L}$, let the eigenvalues of $\mathcal{L}$ be $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \leq 2$. Then $\frac{\lambda_2}{2} \leq \phi(G) \leq \sqrt{2\lambda_2}$.*

Cheeger's inequality is the basis of spectral clustering [29, 40]. While there exist various versions of spectral clustering, its basic form consists of the following three steps: (i) Compute the eigenvector x of λ_2, and sort its entries so that $x_1 \leq x_2 \leq \ldots \leq x_n$. (ii) Consider subsets $S_i = \{x_1, \ldots, x_i\}$. (iii) Output $S = \arg \min \phi(S_i)$. The output S has conductance $\phi(S) \leq \sqrt{2\lambda_2}$. Cheeger's inequality has recently been generalized to hypergraphs by Louis [24].

Expander graphs. Intuitively an expander is a graph that contains no set S with low conductance. Expander graphs
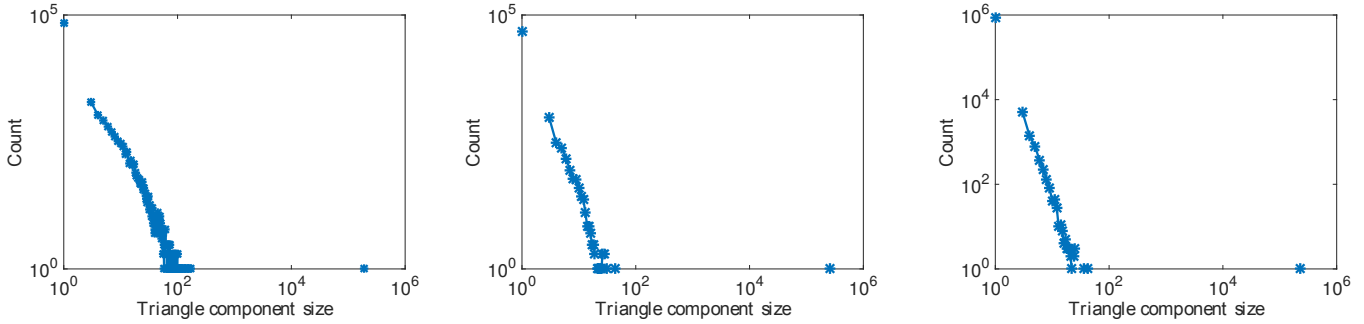
Figure 1: Number of connected components versus size after reweighing each edge with triangle counts for (a) Amazon, (b) DBLP, and (c) Youtube. The original graphs consist of a single connected component.

with constant degree play an important role in a wide variety of applications, including coding theory and hashing. The interested reader may read the excellent monograph of Hoory, Linial, and Widgerson for more details [19]. The formal definition follows.

DEFINITION 1 (EXPANDER). *A graph $G(V,E,w)$ where $w : E \to \mathbb{R}^+$ is an expander if all subsets $S \subseteq V$ with $|S| \leq 0.5n$ have edge expansion $\phi(S) = \Theta(1)$.*

Triangle biased random walks. Motifs, and specifically triangles, have been used in random walks, e.g., [9, 8]. For example, Backstrom and Kleinberg [8] used weighted triangle closing walks as follows: when a random walk is at node u and considers which neighbor of u it should choose, it remembers the previous node in the walk s. If (s,v) is an edge, then the walk is biased towards v. According to their findings, this is a successful heuristic for detecting better quality clusters compared to standard random walks.

3. ALGORITHMS

3.1 Theoretical Framework

Triangle Conductance. Let $G(V,E)$ be an unweighted, undirected graph, and set $\mathrm{vol}_3(S) = \sum_{v \in S} t(v)$. From now on, we denote $\mathrm{vol}(S)$ as $\mathrm{vol}_2(S)$ in order to distinguish vol_2 and vol_3. Also, for a set $S \subseteq V$, define $t_i(S)$ to be the number of triangles with exactly i vertices in S. By double counting we obtain $\mathrm{vol}_3(S) = 3t_3(S) + 2t_2(S) + t_1(S)$. Consider the following biased random walk that utilizes the intuition that triangles play an important role in community detection. When at node u the random walk chooses a neighbor $v \in N(u)$ with probability proportional to $t(u,v)$. Equivalently, when at node u we choose a triangle that u participates in uniformly at random and then choose an endpoint of that triangle, other than u, uniformly at random. Notice that if the random walk starts at a vertex u that does not participate in any triangles, i.e., $t(u) = 0$, then the random walk stays at u. Let $S \subseteq V$ be any set of vertices, and denote by $\phi_3(S)$ the probability of leaving S in one step of the walk conditioned on being at a vertex u chosen from S proportionally to the number of triangles $t(u)$ it participates in[2] . Then,

---

[2]To see why $\phi_3(S)$ equals to the escape probability notice that $\sum_{u \in S} \frac{t(u)}{\mathrm{vol}_3(S)} \frac{0 \times t_3(u) + 0.5 \times t_2(u) + 1 \times t_1(u)}{t(u)} = \frac{t_2(S) + t_1(S)}{\mathrm{vol}_3(S)}$. Here $t_i(u)$ is the number of triangles with i vertices in S (u included).

$$\phi_3(S) = \frac{2t_2(S) + 2t_1(S)}{6t_3(S) + 2t_2(S) + 2t_2(S) + 2t_1(S)} = \frac{t_2(S) + t_1(S)}{\mathrm{vol}_3(S)}.$$

Clearly $\phi_3(S) \in [0,1]$. We define the graph triangle conductance as

$$\phi_3(G) = \min_{S \subseteq V} \frac{t_2(S) + t_1(S)}{\min(\mathrm{vol}_3(S), \mathrm{vol}_3(\bar{S}))}.$$

Notice that the denominator is set to the minimum of the triangle volumes because of the symmetry $t_2(S) + t_1(S) = t_2(\bar{S}) + t_1(\bar{S})$.

3.2 Triangle Spectral Clustering

We provide an efficient approximation algorithm for the triangle conductance problem. Notice that this is essentially a hypergraph problem where each hyperedge corresponds to a triangle. For a given input graph $G(V,E)$ with a set of triangles $\mathcal{T}_G \subseteq \binom{[n]}{3}$, define the 3-uniform hypergraph $\mathcal{H}(V, E_{\mathcal{H}})$, where each hyperedge $e \in E_{\mathcal{H}}$ corresponds to a triangle $u, v, w \in \mathcal{T}_G$. Consider any cut $(S : \bar{S})$ in G and $\mathcal{H}$. The number of triangles $t(S : \bar{S})$ that go across the cut $(S : \bar{S})$ in G is equal to the number of hyperedges going across $(S : \bar{S})$ in $\mathcal{H}$. However, creating $\mathcal{H}$ and then using state-of-the-art semidefinite programming techniques for spectral clustering in [24] is computationally expensive. Our main theoretical result overlaps with the algorithmic contribution of [10], and is stated here without proof, for completeness reasons. Our result provides an efficient way to perform triangle spectral clustering. The interested reader can read our proof on arxiv [36].

THEOREM 2. *Given an undirected, connected graph $G(V,E)$, let $w : E \to \mathbb{R}^+$ be the weight function that assigns to each edge e weight $w(e)$ equal to the number of triangles $t(e)$ that e is contained. Let $H(V, E, w)$ be the weighted version of G. Let the eigenvalues of $\mathcal{L}_H$ be $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n \leq 2$. Then Cheeger's clustering algorithm on $H(V, E, w)$ outputs a cut $(S : \bar{S})$ such that*

$$\frac{\lambda_2(H)}{2} \leq \phi_3(G) \leq \sqrt{2\lambda_2(H)}. \tag{1}$$

Quadratic form for triangle clustering. We define for each triangle $\Delta(u,v,w)$ a $n \times n$ positive semidefinite matrix $L_{\Delta(u,v,w)}$ that is zero except at the intersection of rows

Algorithm 1 TECTONIC

Require: Undirected, unweighted, connected graph $G(V, E)$

Require: Threshold $\theta > 0$

   Count $t(u, v)$ for each $(u, v) \in E$

   Reweight each edge $(u, v) \in E$ by $w(u, v) \leftarrow \frac{t(u,v)}{deg(u)+deg(v)}$

   Remove all edges (u, v) with weight $w(u, v) < \theta$

   Output the resulting connected components

---

and columns indexed by u, v, w. The non-zero entries are $L_{\Delta(u,v,w)}(i, i) = 2$ for $i \in \{u, v, w\}$, and $L_{\Delta(u,v,w)}(i, j) = -1$ for $i \neq j, i, j \in \{u, v, w\}$. In other words, the 3×3 non-zero sub-matrix of $L_{\Delta(u,v,w)}$ indexed by u, v, w equals

$$L_{\Delta(u,v,w)} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

Let $x \in \{0, 1\}^n$ be the indicator vector of a cut $(S, V \setminus S)$. Specifically, let $x(u) = 1$ if and only if $u \in S$. Consider the positive semidefinite matrix $Q = \sum_{\Delta(u,v,w)} L_{\Delta(u,v,w)}$. Notice that

$$x^T Q x = \sum_{\Delta(u,v,w)} \left((x_u - x_v)^2 + (x_u - x_w)^2 + (x_w - x_v)^2 \right)$$
$$= 2t_2(S) + 2t_1(S).$$

The spectral approach has been evaluated in [36], and extensively in [10], and has been shown to be very effective in revealing successfully communities in a wide variety of applications. In the next section, we propose TECTONIC, a significantly faster method compared to spectral clustering that produces high quality output as we will see in Section 4.

3.3 Proposed Method: TECTONIC

In Section 1 we saw that reweighting each edge $(u, v) \in E(G)$ of the graph with weights equal to the triangle count $t(u, v)$ results in disconnecting the graph into multiple connected components. But do these components correlate at all with communities? As we will see in detail in Section 4, they do correlate but there is room for improvement. The main issue with the simple reweighting scheme is that it does not handle well imbalance, i.e., the existence of communities with different numbers of nodes. Our proposed method TECTONIC (Triangle Connected Component Clustering, see Algorithm 1) deals with imbalance by normalizing the triangle weight $t(u, v)$ by the sum of degrees $deg(u) + deg(v)$. Then, it removes all edges with weight less than a predefined threshold θ. It is worth outlining that TECTONIC is amenable to distributed implementation as it relies simply on triangle counting and thresholding.

Our heuristic normalization scheme is inspired by the following observation. Let $\theta = \frac{1}{2}\left(1 - \frac{\theta'}{deg(u)+deg(v)}\right)$. Then two neighboring nodes u, v in G become disconnected after reweighting if and only if

$$\frac{t(u, v)}{deg(u) + deg(v)} < \theta \Leftrightarrow$$
$$\frac{1}{2}(deg(u) + deg(v) - \theta') > t(u, v) \Leftrightarrow$$
$$deg(u) + deg(v) - 2t(u, v) > \theta' \Leftrightarrow$$
$$|N(u) \cup N(v)| - |N(u) \cap N(v)| > \theta' \Leftrightarrow$$
$$dist^2(A^{(u)}, A^{(v)}) > \theta',$$

where $N(u) = \{v : (u, v) \in E(G)\}$, and $dist(A^{(u)}, A^{(v)})$ is the Euclidean distance between the u-th and v-th row of the adjacency matrix representation of G.

4. EXPERIMENTAL RESULTS

4.1 Experimental setup

Table 1 shows the three networks we use in our experiments together with the number of nodes n and the number of edges m. We use three social and information graphs for which ground-truth about the community structure is available [2]. For all datasets we use the top 5000 ground-truth communities, as provided by SNAP.

As our competitors we use a list of popular graph clustering methods: MCL [14], Infomap [31], the Girvan-Newman (GN) algorithm [18], the Louvain method [11], the Clauset-Newman-Moore (CNM) [12], Cfinder [4], spectral clustering (SC) [29], and triangle spectral clustering (tSC) [10, 36]. For the Girvan-Newman algorithm, we use the implementation available at SNAP, and for spectral clustering (SC,tSC) we use the Python sklearn library. For all other methods we use the original implementations provided by the authors. Methods that had not completed after several hours were stopped. Our code will become available at https://github.com/tsourolampis/tectonic. Our results were obtained by setting $\theta = 0.06$. We discuss the choice of θ in the next Section, as a rule of thumb we suggest this value.

We count triangles exactly using MACE [1, 38]. All experiments run on a laptop with 1.7 GHz Intel Core i7 processor and 8GB of main memory. Triangle counting took 0.56, 1.25 and 6.6 seconds for Amazon, DBLP, and Youtube graphs respectively.

Name	n	m
AMAZON	334 863	925 872
DBLP	317 080	1 049 866
YOUTUBE	1 134 890	2 987 624

Table 1: Datasets used in our experiments.

4.2 Community detection

Table 2 shows our experimental findings. For each method we use we report the average precision and recall over all 5 000 ground-truth communities. We compute the precision and recall of a given partition as follows: for each ground-truth community S, we find the community S' in the partition that has the largest intersection size with S. Then, we compute how well S' matches S by computing precision and recall. The overall precision and recall that we report is
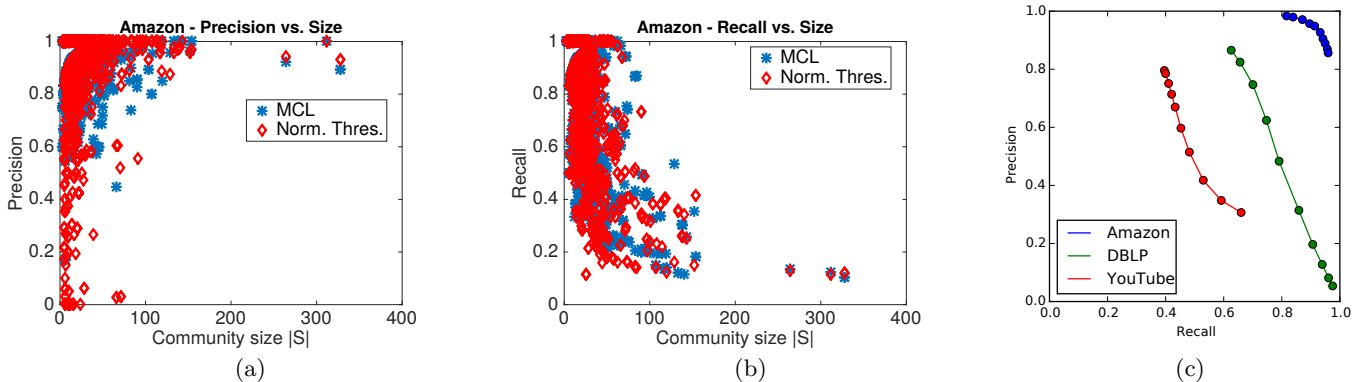
Figure 2: (a) Precision, and (b) Recall vs. ground-truth community size for the Amazon graph using MCL [14], the best competitor, and our method TECTONIC (Norm. Thres.). (c) Precision vs. recall for our method for various threshold values ranging from 0.01 to 0.1 with a step of 0.01.

Method	Amazon			DBLP			YouTube		
	p	r	T	p	r	T	p	r	T
MCL	95.6	90.1	736.54	55.1	81.7	1 166	39.9	60.6	19 187.1
Louvaine	50.0	14.7	9.00	50.20	12.13	10.38	50.13	27.55	55.8
CFinder	-	-	> 5h	-	-	> 5h	-	-	> 5h
GN	-	-	> 5h	-	-	> 5h	-	-	> 5h
CNM	-	-	> 5h	-	-	> 5h	-	-	> 5h
Infomap	50.0	14.8	63.0	50.16	12.13	64.0	50.00	27.6	204
SC	-	-	> 5h	-	-	> 5h	-	-	> 5h
tSC	-	-	> 5h	-	-	> 5h	-	-	> 5h
Thres. 0	85.2	96.0	4.62	4.0	100.0	1.65	22.5	70.8	6.92
Thres. 1	94.1	81.1	4.61	12.0	91.4	1.65	36.1	59.7	6.92
Thres. 2	97.1	67.7	4.62	23.0	81.6	1.65	45.0	53.9	6.92
Thres. 3	98.0	52.4	4.62	35.7	71.4	1.65	49.6	50.3	6.93
TECTONIC	94.9	91.3	4.62	48.3	79.1	1.65	66.7	43.3	6.92

Table 2: Average precision (p), average recall (r) over all ground-truth communities, and total run time (T) in seconds for MCL and our method using various threshold values. The run times for our method include the run time for triangle counting (0.56, 1.25 and 6.6 secs respectively).

averaged over all ground-truth communities. Method Thres. 0 refers to just reweighting edges by triangle counts, as in Figure 1. Methods Thres. 1,2,3 take this idea further, by removing edges whose weight is less or equal than 1,2,3 respectively. Surprisingly, this simple reweighting reveals a lot about the community structure. For example, as soon as we add triangle weights the single connected component of Amazon breaks up into 77 811 components. When we remove all edges whose weight is 1, we obtain 139 456 components. Similarly for threshold values 2, 3 we find 199 693 and 250 572 connected components. Precision and recall show that these components correlate well with the ground-truth communities. The same is true for the other two datasets. Analyzing further the ground-truth communities shows that they typically have low conductance ϕ_2. Therefore, on these datasets low values of ϕ_2 and ϕ_3 are positively correlated. Nonetheless, reweighting by triangle counts may immediately reveal the community structure or lower the conductance further, i.e., $\phi_3(S) < \phi_2(S)$. Even in the latter case, this facilitates the algorithmic discovery of such communities.

In terms of run times, our methods are significantly faster than other methods. At one extreme, CFinder, GN, CNM, SC, tSC do not produce any output after running for at least 5 hours. Actually, GN does not produce any output after running for at least 10 hours. Louvain is the fastest method among competitors but produces significantly lower quality output compared to MCL. Infomap has a similar behavior to Louvain, but is slightly slower. Our method only requires a few seconds, as it only needs to compute the degree sequence, the triangle counts, and the connected components.

TECTONIC provides state of the art performance that can compete with MCL in terms of quality but is significantly faster. For instance, on the YouTube graph it is more than 2 741 times faster than MCL. Figures 2(a),(b) show a detailed view of precision and recall as a function of the community size for MCL and our normalized thresholding method. Figure 2(c) plots precision vs. recall for our method for various threshold values ranging from 0.01 to 0.1 with a step of 0.01 for all three datasets. Our choice for the threshold in Table 2 was the middle choice 0.06. As the threshold increases, precision increases and recall decreases. Finally, it is worth outlining that many points correspond-

ing to communities in Figures 2(a),(b) fall on the top of each other. We provide a detailed view of recall versus precision in the form of heat maps in the full version of our work [36].

5. THEORETICAL FOUNDATIONS

5.1 Preliminaries

We use a powerful probabilistic result from [16] to prove our main results in Section 5.3.

DEFINITION 2 (READ-k FAMILIES). *Let $X_1, \ldots, X_m$ be independent random variables. For $j \in [r]$, let $P_j \subseteq [m]$ and let f_j be a Boolean function of $\{X_i\}_{i \in P_j}$. Assume that $|\{j | i \in P_j\}| \leq k$ for every $i \in [m]$. Then, the random variables $Y_j = f_j(\{X_i\}_{i \in P_j})$ are called a read-k family.*

THEOREM 3 (CONCENTRATION OF READ-k FAMILIES). *Let $Y_1, \ldots, Y_r$ be a family of read-k indicator variables with $\mathbf{Pr}[Y_i = 1] = q$. Also, let $Y = \sum_{i=1}^r Y_i$. Then for any $\epsilon > 0$,*

$$\mathbf{Pr}[Y \geq (1+\epsilon)\mathbb{E}[Y]] \leq e^{-\frac{\epsilon^2 \mathbb{E}[Y]}{2k(1+\epsilon/3)}} \qquad (2)$$

$$\mathbf{Pr}[Y \leq (1-\epsilon)\mathbb{E}[Y]] \leq e^{-\frac{\epsilon^2 \mathbb{E}[Y]}{2k}}. \qquad (3)$$

5.2 Planted partition model

The following example illustrates the benefit of using the triangle biased walk we described in Section 3.1 instead of the standard random walk. Let $G \sim G(nk, k, p, q)$ be a graph sampled from the planted partition model on nk vertices, with k clusters each with exactly n vertices. Specifically, let $\Psi : V \to [k]$ be the partition function and let any pair of distinct vertices $u, v \in V(G)$ connect with probability p if $\Psi(u) = \Psi(v)$ and with probability $q < p$ otherwise. For the sake of simplicity, assume p, q are two distinct constants.

LEMMA 1. *Let $G \sim G(kn, k, p, q)$ be an unweighted graph. Let $H(V, E, w)$ be the auxiliary graph derived from G where the graphs edges (u, v) are weighted as $w(u, v) = t(u, v)$, i.e., according to the number of triangles that contain edge (u, v). Consider random walks X_t and Y_t on the vertices of G and H, respectively, where the random walk on G is the standard random walk on the random walk on H a neighbor proportionally to the weights on the edges. Then with probability $1 - o(1)$ over the choice of G, for all vertices u,*

$$\mathrm{Pr}(\Psi(X_{t+1}) = \Psi(X_t) \mid X_t = u) < \mathrm{Pr}(\Psi(Y_{t+1}) = \Psi(Y_t) \mid Y_t = u).$$

In plain words, Lemma 1 shows that the random walk on H is more likely to stay in the same component of the planted partition than the random walk on G. Leveraging these ideas further, we can show that in the planted partition model, reweighting edges by triangle counts can completely reveal the cluster structure.

Proof of Lemma 1

We provide the intuition of the proof by working with expectations; the full proof uses relies on concentration of all values around their expectations, which follows from concentration of measure.

For the random walk on G, a vertex u has $p(n-1)$ neighbors in expectation in the same partition, and qn neighbors in expectation in each other partition. For simplicity we

use pn as the expectation for the number of neighbors in the same partition as asymptotically the difference does not matter. Thus $\mathrm{Pr}(\Psi(X_{t+1}) = \Psi(X_t) \mid X_t = u) = \frac{p}{p+q(k-1)}$ with our approximations.

For the random walk on H, we first determine the expected vertex weights. If $(u, v) \in E(G)$, and $\Psi(u) \neq \Psi(v)$, then $\mathbb{E}[w(u, v)] = 2(n-2)pq + (k-2)nq^2$. The first term corresponds to triangles where the third vertex is in the same component as u or v, the second term to triangle where the third vertex is in another component. Similarly, if $\Psi(u) = \Psi(v)$, then $\mathbb{E}[w(u, v)] = (n-2)p^2 + (k-1)nq^2$. Again for simplicity we avoid lower order terms and use weights $2npq + (k-2)nq^2$ and $np^2 + (k-1)nq^2$ for the two cases.

For the random walk on H, there are in expectation $(n-1)p$ neighbors in the same partition, and $(k-1)nq$ neighbors in the other partitions. Hence the total expected weight of edges to neighbors in the same partition is (again, approximately) $np(np^2 + (k-1)nq^2)$, against $(k-1)nq(2npq + (k-2)nq^2)$ to other partitions. We thus find that

$$\mathbf{Pr}[\Psi(Y_{t+1}) = \Psi(Y_t) \mid Y_t = u] =$$
$$\frac{p^3 + (k-1)pq^2}{p^3 + 3(k-1)pq^2 + (k-1)(k-2)q^3}$$

The following chain of statements are equivalent:

$$\frac{p^3 + (k-1)pq^2}{p^3 + 3(k-1)pq^2 + (k-1)(k-2)q^3} > \frac{p}{p + q(k-1)} \Leftrightarrow$$
$$(k-1)p^3 q + (k-1)pq^3 > 2(k-1)p^2 q^2 \Leftrightarrow$$
$$2pq < p^2 + q^2.$$

The last statement follows from the arithmetic mean-geometric mean inequality, with strict inequality as $p \neq q$.

The high probability result follows from the fact that all expectations are correct *whp* up to lower order terms due to concentration. Hence with more non-instructive work we find that *whp* for all vertices u:

$$\mathbf{Pr}[\Psi(X_{t+1}) = \Psi(X_t) \mid X_t = u] = \frac{p}{p + q(k-1)} + o(1);$$

$$\mathbf{Pr}[\Psi(Y_{t+1}) = \Psi(Y_t) \mid Y_t = u] =$$
$$\frac{p^3 + (k-1)pq^2}{p^3 + 3(k-1)pq^2 + (k-1)(k-2)q^3} + o(1).$$

The result follows. ∎

We also outline how in the planted partition model reweighting edges by triangle counts can recover the cluster structure. (This is a phenomenon observed on real data as well, see Figure 1 in Section 1.) For example, set $p = \frac{3\log n}{\sqrt{n}}$, $q = \frac{\log n}{\sqrt{n}}$, and let $G \sim G(2n, 2, p, q)$ be a graph sampled according to the planted partition model. The weight of an edge within a cluster w_{in} has expectation $\binom{n-1}{1}p^2 + \binom{n}{1}q^2 \approx 10\log^2 n$, and similarly the expectation of the weight w_{out} of an edge crossing clusters is $\mathbb{E}[w_{out}] = 6\log^2 n$. By Chernoff bounds, we obtain that $\mathbf{Pr}[w_{in} < 8\log^2 n] = o(n^{-2})$ and similarly $\mathbf{Pr}[w_{out} > 8\log^2 n] = o(n^{-2})$. A union bound over all possible $\binom{n}{2}$ edges yields that with high probability all edges within a cluster have weight at least $8\log^2 n$ and all

edges crossing clusters have weight at most $8\log^2 n$. It follows immediately that removing edges with weight less than $8\log^2 n$ recovers the two clusters. A more complete analysis with bounds on the required "gap" between p and q needed to recover clusters will appear in the full version.

5.3 Triangle expanders

We extend the notion of an expander graph to a triangle expander.

DEFINITION 3. *A graph $G(V,E)$ is a triangle expander if all subsets $S \subseteq V$ with $|S| = s \leq 0.5n$ have constant triangle expansion, i.e., $\phi_3(S) = \Theta(1)$.*

We prove that triangle expanders exist.

THEOREM 4. *Let $G \sim G(n,p)$ with p equal to $\frac{\log(n)}{n^{1/3}}$. With high probability, G is a triangle expander.*

Notice that for this range of p, the expected number of edges is $O(n^{\frac{5}{3}}\log n)$. An interesting open problem is to show the existence of sparser triangle expanders. We make the following conjecture.

Conjecture 1: $G \sim G(n,p)$ with p equal to $\frac{\log(n)}{n^{2/3}}$ is a triangle expander *whp*.
Also, an interesting question is whether triangle expansion implies edge expansion. Our result is stated as the following theorem.

THEOREM 5. *There exist edge expanders that are not triangle expanders. Similarly, under conjecture 1, there exist triangle expanders that are not edge expanders.*

Our construction works not only under conjecture 1, but for any triangle expander that has diameter at least 3.

Proof of Theorem 4

Consider any cut $(S : \bar{S})$. We prove concentration results for the number of triangles $t(S : \bar{S})$ cut by $(S : \bar{S})$, and for the triangles induced by S separately. Then, we combine the two concentration results to prove that $\phi_3(G) = \Theta(1)$.

Define an indicator variable $X_{uv} = \mathbf{1}(u \sim v)$ for each pair of distinct vertices $u, v \in V$. Notice $\mathbb{E}[X_{uv}] = p$. Let ϵ be a fixed constant.
Number of triangles $t(S : \bar{S})$ cut by $(S, \bar{S})$. For each value $s = 1, \ldots, 0.5n$, define Q_s to be the event

$$Q_s = \exists S \subseteq V : |S| = s, \left|t(S : \bar{S}) - \mathbb{E}\left[t(S : \bar{S})\right]\right| > \epsilon \mathbb{E}\left[t(S : \bar{S})\right].$$

The random variable $t(S : \bar{S})$ is the sum of two multivariate polynomials,

$$t(S : \bar{S}) = \underbrace{\sum_{u \in S, v \notin S} X_{uv}X_{vw}X_{uw}}_{T_1(S)} + \underbrace{\sum_{u,v \in S, w \notin S} X_{uv}X_{vw}X_{uw}}_{T_2(S)}.$$

The two polynomials are equal to the number of triangles which have exactly one and two vertices in S respectively. By the independence of the random variables $\{X_{uv}\}$ and the linearity of expectation, $\mathbb{E}[T_1(S)] = \binom{s}{1}\binom{n-s}{2}p^3$, and $\mathbb{E}[T_2(S)] = \binom{s}{2}\binom{n-s}{1}p^3$. Therefore,

$$\mathbb{E}\left[t(S : \bar{S})\right] = \frac{\log^3(n)}{n}\left(\binom{s}{1}\binom{n-s}{2} + \binom{s}{2}\binom{n-s}{1}\right).$$

We prove that there exists a constant $c = c(\epsilon)$ such that

$$\mathbf{Pr}\left[\left|t(S : \bar{S}) - \mathbb{E}\left[t(S : \bar{S})\right]\right| > \epsilon\mathbb{E}\left[t(S : \bar{S})\right]\right] \leq e^{-cs\log^3 n}.$$

We apply Theorem 3. Here, $m = \binom{n}{2}$, $r = \binom{s}{1}\binom{n-s}{2} + \binom{s}{2}\binom{n-s}{1}$. We define the family of variables $Y_{uvw} = X_{uv}X_{vw}X_{uw}$ for each triple of vertices u, v, w such that either $u \in S, v, w \notin S$ or $u, v \in S, w \notin S$. This is a read-k family of variable where $k \leq n$. We apply Equation (2)

$$\mathbf{Pr}\left[t(S : \bar{S}) \geq (1+\epsilon)\mathbb{E}\left[t(S : \bar{S})\right]\right] \leq \exp\left\{-\frac{\mathbb{E}\left[t(S : \bar{S})\right]\epsilon^2}{2nk(1+\epsilon/3)}\right\}$$
$$\leq \exp\left\{\frac{0.01\epsilon^2 s\log^3 n}{2(1+\epsilon/3)}\right\}$$
$$= e^{-C(\epsilon)s\log^3(n)}$$

By applying Equation (3)

$$\mathbf{Pr}\left[t(S : \bar{S}) \leq (1-\epsilon)\mathbb{E}\left[t(S : \bar{S})\right]\right] \leq e^{-C'(\epsilon)s\log^3(n)},$$

where $C'(\epsilon) = 0.005\epsilon^2$. By taking two union bounds we get for any constant $\epsilon > 0$,

$$\mathbf{Pr}\left[Q_s\right] \leq \binom{n}{s}e^{-\min(C(\epsilon),C'(\epsilon))s\log^3 n} \leq \left(\frac{en}{s}\right)^s e^{-\min(C(\epsilon),C'(\epsilon))s\log^3 n}$$
$$= o(n^{-1}),$$

and therefore by a union bound,

$$\mathbf{Pr}\left[\cup_{s=1}^{0.5n}Q_s\right] \leq no(n^{-1}) = o(1).$$

Number of triangles $T_3(S)$ induced by S. In order to prove that $G \sim G(n,p)$ is a triangle expander *whp*, it suffices to show that for all sets $S \subseteq V$, $T_3(S) = O(\mathbb{E}[T_1(S) + T_2(S)])$ *whp*. We express $T_3(S)$ as the multivariate polynomial $T_3(S) = \sum_{u,v,w \in S} X_{uv}X_{vw}X_{uw}$. Notice that $\mathbb{E}[T_3(S)] = \binom{s}{3}p^3$.
In the following, we prove that $T_3(S)$ does not exceed twice its expectation *whp*. We consider two cases, depending on the cardinality of the set $S \subseteq V$.
● CASE 1: $s = o(n)$ Consider any fixed set $S \subseteq V$ such that $|S| = s = o(n)$. For any cardinality $s = o(n)$, we can write $s = \frac{n}{\omega(n)}$ where $\omega(n)$ is an appropriately chosen slowly growing function such that $\omega(n) \to +\infty$ as $n \to +\infty$. We obtain

$$\mathbf{Pr}\left[T_3(S) \geq 2\mathbb{E}\left[t(S : \bar{S})\right]\right] \leq e^{-sn\log^2 n}.$$

By taking a union bound over all possible subsets $S \subseteq V, s = o(n)$ we obtain that

$$\mathbf{Pr}\left[\exists S : S \subseteq V, s = o(n), T_3(S) \geq 2\mathbb{E}\left[t(S : \bar{S})\right]\right] \leq$$
$$\sum_{s \leq 0.5n}\binom{n}{s}e^{-sn\log^2 n} = o(1).$$

• Case 2: $s = \Theta(n)$

Fix any set $S \subseteq V$ such that $s = \alpha n$ for some constant $\alpha \leq 0.5$. By applying Equation (2) with $\epsilon = 1$ we obtain

$$\mathbf{Pr}\left[T_3(S) \geq 2\mathbb{E}\left[T_3(S)\right]\right] \leq e^{-\frac{\epsilon^2}{2(1+\epsilon/3)}\frac{p^3\binom{s}{3}}{n}}$$

$$\leq e^{-n \log^2 n}.$$

By taking a union bound over all possible subsets $S \subseteq V, s = \Theta(n)$ we obtain that

$$\mathbf{Pr}\left[\exists S : S \subseteq V, s = \Theta(n), t_3(S) \geq 2\mathbb{E}\left[T_3(S)\right]\right] \leq$$

$$\sum_{s=\Theta(n)}\binom{n}{s}e^{-n\log^2 n} \leq \sum_{s=\Theta(n)}\left(\frac{en}{s}\right)^s e^{-n\log^2 n} = o(1).$$

<u>Triangle conductance ϕ_3.</u> By combining our concentration results for $T_3(S), t(S : \bar{S})$, we obtain that *whp* for any set $S \subseteq V, |S| \leq 0.5n$

$$\phi_3(S) \geq \frac{(1-\epsilon)\mathbb{E}\left[t(S : \bar{S})\right]}{3 \times 2\mathbb{E}\left[t(S : \bar{S})\right] + 2(1+\epsilon)\mathbb{E}\left[t(S : \bar{S})\right]}$$

$$\geq \frac{2(1-\epsilon)}{7 + 4\epsilon} = \Theta(1).$$

Therefore, $G \sim G(n, \frac{\log n}{n^{1/3}})$ is a triangle expander *whp*.

Proof of Theorem 5

Since a bipartite network contains no triangles, and there exist bipartite expander graphs, the first direction is trivial. Nonetheless, we provide a non-trivial construction.

(i) Let $G \sim G(n, \frac{\log n}{n^{1/3}})$. We modify G in such a way that we maintain its edge but not its triangle expansion.

<u>Claim 1: Volume is concentrated.</u> We prove that for any $S \subseteq V$, $\text{vol}_2(S) \in [(1 - \epsilon)\mathbb{E}\left[\text{vol}_2(S)\right], (1 + \epsilon)\mathbb{E}\left[\text{vol}_2(S)\right]]$ *whp*. It suffices to show that for *each* vertex $v \in V(G)$, $deg(v) \in [(1 - \epsilon)\mathbb{E}\left[deg(v)\right], (1 + \epsilon)\mathbb{E}\left[deg(v)\right]]$ *whp*. Notice, $deg(v) \sim Bin(n - 1, p)$. The claim is easily proved by applying Chernoff and taking a union bound over n vertices.

<u>Claim 2: Edges crossing cut are concentrated.</u> We prove that for all sets $S \subseteq V$, the number of edges $e(S, \bar{S})$ that cross the cut $(S, \bar{S})$ are concentrated around the expectation. First, notice that $e(S, \bar{S}) \sim Bin(s(n - s), p)$. We define for each possible size $s = 1, \ldots, 0.5n$ the event

$$Q_s = \exists S \subseteq V : |S| = s, e(S : \bar{S}) \notin [(1-\epsilon), (1+\epsilon)]\mathbb{E}\left[e(S : \bar{S})\right].$$

We apply Chernoff and union bounds.

$$\mathbf{Pr}\left[\cup_{s=1}^{0.5n}Q_s\right] \leq \sum_{s=1}^{0.5n}\binom{n}{s}2e^{-\epsilon^2/3\frac{s(n-s)\log n}{n^{1/3}}}$$

$$\leq 0.5no(n^{-1}) = o(1)$$

<u>Claim 3: Edge conductance is constant whp.</u> By combining claims 1,2 we obtain that for any set $S \subseteq V$ with less than $0.5n$ vertices

$$\phi_2(S) \geq \frac{(1-\epsilon)ps(n-s)}{(1+\epsilon)\left(2\binom{s}{2} + s(n-s)\right)} = \Omega(1).$$

Recall that G is also a triangle expander, namely for all sets $S \subseteq V$ with $s \leq 0.5n$ $\phi_3(S) = \Theta(1)$.

Consider the following modification to G. Pick a subset S with $s = n^{2/3}$ vertices and any $X \subseteq S$ with $n^{2/3-\gamma}$ vertices, where $\gamma = \frac{1}{10}$. We add a clique on X by adding in expectation

$$(1 - \frac{\log n}{n^{1/3}})\binom{|X|}{2} = (1 - o(1))\binom{n^{\frac{2}{3}-\gamma}}{2}$$

extra edges. Let G' be the resulting graph. Now, we prove that G' is an edge but not a triangle expander.

$$\phi_2'(S) \approx \frac{pn^{2/3}(n - n^{2/3})}{pn^{2/3}(n - n^{2/3}) + 2p\binom{n^{2/3}}{2} + \binom{|X|}{2}} \to 1.$$

It is also easy to check that the conductance of X and any subset of it is constant. For instance,

$$\phi_2'(X) \approx \frac{p|X|(n - |X|)}{\binom{|X|}{2} + p|X|(n - |X|)} \to 1.$$

However, the triangle conductance of S becomes

$$\phi_3'(S) \approx \frac{p^3\binom{s}{2}\binom{n-s}{1} + \binom{s}{1}\binom{n-s}{2}}{\binom{|X|}{3} + p^3\binom{s}{3} + \binom{s}{2}\binom{n-s}{1} + \binom{s}{1}\binom{n-s}{2}} =$$

$$\frac{n^{5/3}\log^3 n}{n^{2-3\gamma} + n^{5/3}\log^3 n} = o(1),$$

since $3\gamma < 1/3$.

(ii) We provide a general construction that can be applied to modify any graph that is both an edge and a triangle expander of diameter at least 3 to a graph that is a triangle expander but not an edge expander. Notice that $G \sim G(n, p)$ with $p = \frac{\log(n)}{n^{2/3}}$ has diameter at least 3, and under conjecture 1 is a triangle expander *whp*. Since the diameter is at least 3, there exists a pair of nodes u, v such that $dist(u, v) \geq 3$. We add an edge of arbitrarily large weight between u, v. Since $dist(u, v) \geq 3$, the number of common neighbors $|N(u) \cap N(v)|$ between u and v is 0, so the new edge (u, v) does not change the triangle conductance. However the edge conductance of $\{u, v\}$ becomes arbitrarily close to 0 as we increase the weight of the edge. ∎

Motif-based conductance. The framework we developed for the case of triangles naturally extends to other clique motifs. For instance, if the motif of interest is a clique on four nodes, then we define the K_4-conductance ϕ_4 of a set of nodes $S \subseteq V$ as $\phi_4(S) = \frac{3c_3+4c_4+3c_1}{12c_4+9c_3+6c_2+3c_1}$, where c_i is the number of K_4 with i nodes in S. Defining appropriate random walks for general motifs, and deriving the conductance in a principled way is an interesting question.

6. CONCLUSION

As triangles are a natural indicator of community, we have suggested formalizing the importance of triangles by considering reweighting edges according to the number of triangles the edge participates in. While our framework is simple, we have shown that it is quite powerful, both in the more theoretical planted partition model and on real-world graph

experiments. Another advantage of our approach is that it is amenable to distributed implementations. Furthermore, it strengthens already existing approaches based on conductance and spectral clustering. It also can generalize naturally to other graph motifs.

Our work suggests several natural open directions. First, we might consider variations on the reweighting scheme. For example, for each edge in the graph we might use a weight of the form $1 + \alpha t(e)$ for some parameter α; this way edges would still have some weight even if they were not part of any triangle. More generally, understanding how to set appropriate or approximately optimal edge weights based on motifs for different applications seems quite interesting. Also, it is worth exploring the effect of approximate motif counting algorithms, e.g., [20, 30], on the clustering performance. Second, we believe the notion of triangle conductance has further consequences from a theoretical perspective. It would be of interest to better understand its behavior in random graphs, and applications to graph clustering algorithms. Finally, we have not focused on whether our specific choice of reweighting by triangles might lead to especially efficient algorithms designed for this case.

Acknowledgements

7. REFERENCES

[1] Mace. http://research.nii.ac.jp/~uno/codes.htm.

[2] Stanford network analysis project. http://snap.stanford.edu/data/index.html.

[3] B. Abrahao, S. Soundarajan, J. Hopcroft, and R. Kleinberg. On the separability of structural classes of communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

[4] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.

[5] N. Alon, Z. Galil, and V. D. Milman. Better expanders and superconcentrators. *Journal of Algorithms*, 8(3):337–347, 1987.

[6] N. Alon and V. D. Milman. λ_1, isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory, Series B*, 38(1):73–88, 1985.

[7] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009.

[8] L. Backstrom and J. Kleinberg. Network bucket testing. In *Proceedings of the 20th international conference on World wide web*, pages 615–624. ACM, 2011.

[9] A. R. Benson, D. F. Gleich, and J. Leskovec. Tensor spectral clustering for partitioning higher-order network structures. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC*, 2015.

[10] A. R. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[12] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004. Implementation available at https://www.cs.unm.edu/~aaron/research/fastmodularity.htm.

[13] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical review letters*, 94(16):160202, 2005.

[14] S. v. Dongen. Graph clustering by flow simulation. 2000.

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

[16] D. Gavinsky, S. Lovett, M. Saks, and S. Srinivasan. A tail bound for read-k families of functions. *Random Structures & Algorithms*, 2014.

[17] A. Gionis and C. E. Tsourakakis. Dense subgraph discovery: KDD 2015 tutorial. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[18] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[19] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[20] M. Kolountzakis, G. Miller, R. Peng, C. Tsourakakis. Efficient triangle counting in large graphs via degree-based vertex partitioning. *Internet Mathematics* 8:161–185, 2012.

[21] C. Klymko, D. Gleich, and T. G. Kolda. Using triangles to improve community detection in directed networks. *arXiv preprint arXiv:1404.5874*, 2014.

[22] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM (JACM)*, 46(6):787–832, 1999.

[23] J. Leskovec, K. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

[24] A. Louis. Hypergraph markov operators, eigenvalues and approximation algorithms. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 713–722, New York, NY, USA, 2015. ACM.

[25] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[26] N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. Finding strongly knit clusters in social networks. *Internet Mathematics*, 5(1-2):155–174, 2008.

[27] M. Mitzenmacher, J. Pachocki, R. Peng, C. E. Tsourakakis, and S. C. Xu. Scalable large near-clique detection in large-scale networks via sampling. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 815–824. ACM, 2015.

[28] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[29] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[30] R. Pagh and C. E. Tsourakakis. Colorful triangle counting and a mapreduce implementation. *Information Processing Letters*, 112(7):277–281, 2012.

[31] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[32] V. Satuluri, S. Parthasarathy, and Y. Ruan. Local graph sparsification for scalable clustering. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 721–732. ACM, 2011.

[33] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[34] O. Sporns and R. Kötter. Motifs in brain networks. *PLoS Biol*, 2(11):e369, 2004.

[35] S. L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the internet topology. In *Global Telecommunications Conference, 2001. GLOBECOM'01. IEEE*, volume 3, pages 1667–1671. IEEE, 2001.

[36] C. Tsourakakis, J. Pachocki, and M. Mitzenmacher. Scalable motif-aware graph clustering. *arXiv preprint arXiv:1606.06235*, 2016.

[37] C. Tsourakakis. The k-clique densest subgraph problem. *24th International World Wide Web Conference (WWW)*, 2015.

[38] T. Uno. An efficient algorithm for solving pseudo clique enumeration problem. *Algorithmica*, 56(1), 2010.

[39] V. V. Vazirani. *Approximation algorithms*. Springer Science & Business Media, 2013.

[40] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[41] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[42] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[43] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge & Information Systems*, 42(1):181–213, 2015.