Triangle-aware Spectral Sparsifiers and Community Detection

Konstantinos Sotiropoulos
Boston University
ksotirop@bu.edu

Charalampos E. Tsourakakis
ISI Foundation and Boston University
tsourolampis@gmail.com

ABSTRACT

Triangle-aware graph partitioning has proven to be a successful approach to finding communities in real-world data [8, 40, 51, 54]. But how can we explain its empirical success? Triangle-aware graph partitioning methods rely on the count of triangles an edge is contained in, in contrast to the well-established measure of effective resistance [12] that requires *global* information about the graph.

In this work we advance the understanding of triangle-based graph partitioning in two ways. First, we introduce a novel triangle-aware sparsification scheme. Our scheme *provably* produces a spectral sparsifier with high probability [46, 47] on graphs that exhibit strong triadic closure, a hallmark property of real-world networks. Importantly, our sampling scheme is amenable to distributed computing, since it relies simply on computing node degrees, and edge triangle counts. Finally, we compare our methods to the Spielman-Srivastava sparsification algorithm [46] on a wide variety of real-world graphs, and we verify the applicability of our proposed sparsification scheme on real-world networks.

Secondly, we develop a data-driven approach towards understanding properties of real-world communities with respect to effective resistances, and triangle counts. Our empirical approach is mainly based on the notion of ground-truth communities in datasets made available originally by Yang and Leskovec [53]. We perform a study of triangle-aware measures, and effective resistances on edges within, and across communities, and we discover certain interesting empirical findings. For example, we observe that the Jaccard similarity of an edge used by Satuluri [40], and the closely related Tectonic similarity measure introduced by Tsourakakis et al. [51] provide consistently good signals of whether an edge is contained within a community or not.

CCS CONCEPTS

• **Theory of computation → Sparsification and spanners**; • **Computing methodologies**;

ACM Reference Format:
Konstantinos Sotiropoulos and Charalampos E. Tsourakakis. 2021. Triangle-aware Spectral Sparsifiers and Community Detection. In *Proceedings of Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467260

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467260

1 INTRODUCTION

Community detection is the task of grouping the set of nodes of a graph into clusters of densely connected nodes [42]. Given the diverse nature of real-world networks, the notion of what constitutes a *good cluster* depends on the application context. In recent years, motif-aware[1] community detection has proven to be successful in various applications. Motifs are basic interaction patterns that recur throughout networks, much more often than in random networks [34]. For instance, Satuluri, Parthasarathy, and Ruan designed an elegant method for finding communities [40]. The core idea lies in reweighing each edge $e = (u, v)$ in the graph by the Jaccard similarity coefficient, i.e., a function of the number of triangles $t(u, v)$ the edge is contained in, and the endpoints' degrees $deg(u), deg(v)$. For the precise definition of the edge Jaccard similarity, see Table 2. Benson et. al. in their SCIENCE paper introduce motif-aware graph partitioning [8]. Tsourakakis et. al. introduce the notion of motif-expanders, and give a foundational way of thinking about motif-aware graph partitioning using appropriately defined random walks. These random walks naturally give rise to motif-aware notions of the graph conductance [51]. Furthermore, they provide a similar spectral method to [8], and a heuristic called *Tectonic* similar in spirit to the method of Satuluri et al. [40]. These works show empirically that motif-aware graph partitioning solves effectively various real-world problems ranging from bioinformatics to social networks [8], and outperforms various popular community detection methods on data where ground-truth communities are available [51]. Since then, lots of other works have followed up showing a variety of other applications, e.g., [16, 21, 29, 30, 35].

Despite the empirical success of motif-aware graph partitioning, and certain theoretical advances, a deeper understanding of why it succeeds on real-world network data evades us. It is natural to ask how can it be that localized information about each edge can lead us to reveal the global community structure of the graph? In this work we focus on the *triangle motif* (i.e., a clique on three nodes) that plays an important role in social networks. The latter tend to be abundant in triangles, and this is considered a hallmark property [43] due to the fact that friends of friends tend to become friends themselves. To make a high-level analogy, the power of *convex optimization* is based on the fact that from local information, we can draw a safe conclusion about the global minimum of a convex function [10], a property certainly not true for all functions.

In our setting, the well established measure of the *effective resistance*[2] requires *global* information about the graph, and captures the importance of an edge with respect to connectivity. Specifically,

[1]The authors will be using the term *motif-aware* as coined by Tsourakakis et al. [51] instead of *higher order* as coined by Benson, Gleich and Leskovec [8]. The latter term is also used in a different way within the same context of graph partitioning using spectral methods, see [26].

[2]Formally, the effective conductance $C_{eff}(u, v)$ between two nodes u, v in a weighted undirected graph $G(V, E, w)$, $w : E \to \mathbb{R}^+$ is the value of the current from u to v when the potential difference between u and v is set to 1, and the edge

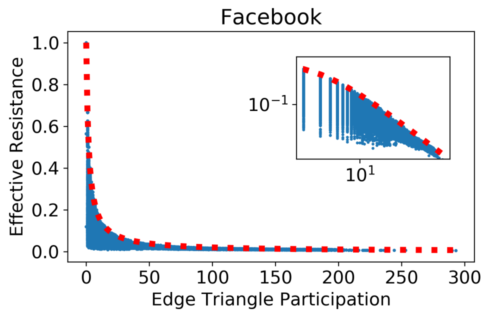


Figure 1: Effective Resistance of an edge, as a function of the number of triangles it participates for the Facebook graph, see Table 3

	Triangles	Eff. Res. $\delta = 0.25$	Eff. Res. $\delta = 0.1$
Amazon	0.7	136	891
DBLP	0.9	115	721
Flickr	12.5	366	2377
YouTube	2.3	349	2412

Table 1: Running time in seconds (secs) for triangle listing and approximate effective resistance computation within multiplicative $(1 \pm \delta)$ error using the fastest available implementation, where $\delta = 0.25, 0.1$ respectively on four real-world networks, see Table 3.

the effective resistance of an edge is equal to the probability that it is included in a random spanning tree of the graph [12]. It is worth remarking that the effective resistance plays a major role in spectral graph theory and its applications [3, 46, 47], and graph clustering [3]. Intuitively, an edge within a community has lower effective resistance than an edge going across two communities.

• We prove (see Theorems 3.3, 3.4 for the precise statements) that sampling each edge inversely to its overlap similarity or its triangle count (see Table 2) produces a spectral sparsifier with high probability. For the case of triangle-dense graphs (see Definition 3.1, and [20] for a closely related random graph model) we prove that it suffices to sample $O(\frac{n \log n}{\epsilon^2})$ edges to obtain a $(1 \pm \epsilon)$-sparsifier (see Definition (1)).

The key intuition behind our sampling is illustrated in Figure 1; this figure plots the effective resistance of each edge versus its triangle count for the Facebook graph (see Table 3). The results are representative of what we observe on real-world networks. For the edges that participate in few triangles, we observe that the effective resistance may range from low to large values. Therefore, our sampling schemes sample edges with small triangle counts with large probability, to make sure that edges that are critical for connectivity are maintained. We sparsify more aggressively edges with large triangle counts since they have low effective resistance.

Table 1 shows the runtimes in seconds for counting the number of triangles each edge is contained in using a straight-forward

Symbol	Definition	Description
$N(u)$	$\{v : (u,v) \in E(G)\}$	Neighborhood
$deg(u)$	$\|N(u)\|$	Node degree
S	$S \subseteq V$	Node subset
$e(S)$	$\|\{(u,v) \in E : u, v \in S\}\|$	# Induced edges
D	$diag(deg)$	Degree matrix
A	(standard)	Adjacency matrix
L	$D - A$	Graph Laplacian
$R_{eff}(u,v)$	$(e_u - e_v)^T L^\dagger (e_u - e_v)$	Effective Res.
$t(u,v)$	$\|\{z : (u,z), (z,v) \in E(G)\}\|$	# Triangles of edge $e = (u,v)$
d_e	$\min(deg(u), deg(v))$	Min-degree of edge $e = (u,v)$
$OC(u,v)$	$\frac{\|N(u) \cap N(v)\|}{\min(deg(u), deg(v))}$	Overlap sim.
$J(u,v)$	$\frac{\|N(u) \cap N(v)\|}{\|N(u) \cup N(v)\|} = \frac{t(u,v)}{deg(u)+deg(v)-t(u,v)}$	Jaccard sim.
$Tect(u,v)$	$\frac{t(u,v)}{deg(u)+deg(v)}$	Tectonic sim. [51]
$RMSE(\hat{\theta})$	$\sqrt{\frac{\sum_{t=1}^{T}(\hat{\theta}_t - \theta_t)^2}{T}}$	RMSE

Table 2: Notation.

triangle listing algorithm, and the running time for computing the effective resistance of each edge using the state-of-the-art Laplacian solver [45]. While the seminal work of Spielman and Srivastava outputs a $(1+\epsilon)$-sparsifier of *any* graph G, having $O(\frac{n \log n}{\epsilon^2})$ edges by sampling each edge proportionally to its effective resistance, we see that our approach scales better to large-scale graphs.

• We complement our foundational contributions with a data-driven approach for better understanding the power of local information in real-world graphs. Our data-driven approach is inspired by the work of Yang and Leskovec who introduced a benchmark of graphs with ground-truth communities. Despite the fact that their notion of ground-truth communities has received a fair amount of criticism by the seminal work of Peel, Larremore, and Clauset [38], this benchmark is the closest available to some acceptable notion of ground-truth communities[3]. Furthermore, even for groups characterized mostly by their meta-data, we find that certain but not all triangle-aware measures can be helpful in uncovering them.

Our findings show that local information can be more than just a good proxy of global connectivity measures such as the effective resistance. We find that in some datasets effective resistance of edges spanning different communities are *less* on average than the effective resistance of edges within communities. This finding challenges the intuition that an edge crossing two communities should always have higher effective resistance [3]. Furthermore, we find that even if triangle counts can also be misleading in a similar way to effective resistance (i.e., edges across communities can participate on more triangles, than edges within), *degree normalized* triangle counts are consistently good signals. The latter empirical finding advances our understanding of Satuluri's [40], and Tsourakakis et al. [51] community detection methods.

• We perform experiments on a wide variety of real-world graphs. Specifically, we compare our sparsification schemes against the Spielman-Srivastava [46] sparsification with respect to the quality

weights are interpreted as conductances. The effective resistance $R_{eff}(u,v)$ is defined as $R_{eff}(u,v) = \frac{1}{C_{eff}(u,v)}$.

[3]These should be called labeled groups rather than ground-truth communities according to the findings of [38].

of the produced sparsifier, and we observe that on real-world graphs our approach produces faster high-quality spectral sparsifiers.

Finally, Table 2 introduces the notation we will use in the rest of the paper.

2 RELATED WORK

Community detection. Given an undirected graph $G(V, E)$, how do we partition its vertex set into k communities $C = \{C_1, \ldots, C_k\}$ [14, 23, 42]? Intuitively, a community is a set of nodes with more and/or better intra- than inter-connections. An intuitive definition of a community is based on internal edge density, namely it is a set of nodes that induces lots of edges. This definition naturally drives the development of dense subgraph discovery methods [17], including the densest subgraph problem and other variations. Another intuitive definition of a community is based on how many edges leave it compared to the number of edges inside it. This intuition captures the fact that a member of a community tends to connect more frequently with members of the same community compared to outside agents. We discuss briefly two important measures that capture this intuition, modularity and conductance. Modularity is a popular quality index for graph clustering. Specifically, for a given partition $C = \{C_1, \ldots, C_k\}$ of the vertex set V into k clusters, the modularity score $Q(C)$ is defined as $Q(C) = \sum_{i=1}^{k} \left(\frac{e(C_i)}{|C_i|} - (\frac{\sum_{v \in C_i} deg(v)}{2m})^2 \right)$ [18], where m is the number of edges in the network. This measure quantifies the difference between the number of edges induced by each set of nodes C_i compared to the expected number of induced edges by C_i in a random graph with the same degree distribution, and sums the differences over all k clusters. Optimizing modularity over all possible partitions of V into k sets is a popular objective [36], and also NP-hard [11].

The notion of conductance plays a major role in graph clustering. The conductance $\phi(S)$ of a set of nodes $S \subseteq V$ is defined as $\phi(S) = \frac{e(S:\bar{S})}{2e(S)+e(S:\bar{S})}$, where $e(S : \bar{S})$ is the number of edges with one endpoint in S and one endpoint in $\bar{S} = V \backslash S$, and $e(S)$ is the number of induced edges by S. The graph conductance ϕ_G is defined as $\phi_G = \min_{S \subset V} \phi(S)$. A lot of work has focused on developing approximation algorithms [5], but as noted in numerous papers, spectral clustering [4] is considered to be the most practical approach at the moment among methods that come with solid theoretical guarantees. A wide variety of heuristics has been also proposed. For instance, Louvain's method is a near-linear time, popular heuristic for modularity optimization [9].

The conductance $\phi(S)$ of any set can be thought of as an escape probability from S of an appropriately defined standard random walk, and ϕ_G is related to the rate of convergence to the stationary distribution of a random walk [28]. When ϕ_G is large, then the graph is well-connected, and vice versa, for a set S to be a good community, $\phi(S)$ has to be small [23]. Various variants of conductance have been proposed, and have found numerous applications in social networks' analysis [19, 27].

Breaking the circular reasoning around community detection. As Abrahao et al. eloquently state [2] "... an attempt to capture community structure by maximizing a given objective function may represent an un-realistic expectation". An important line of work in recent year is *data-driven community detection*, see [2, 27, 53]. These works play a major role in understanding and evaluating measures and algorithms on meaningful communities. They aim to understand what a community *really* is. Thus, they break the "chicken-egg" problem: first define a "community" measure, and then optimize it to find a community. Yang and Leskovec [53] gathered 230 social, collaboration, and information networks with explicit ground-truth, and studied various graph statistics. We use some of these datasets in Section 4. In a subsequent work, Peel, Larremore and Clauset [38] challenge the notion of ground-truth communities. As they point out, by ground-truth, we usually refer to a set of nodes sharing common metadata labels. However, sometimes these metadata are irrelevant to the network structure, that community detection algorithms try to uncover. We experimentally observe that, even in those cases, triangle aware measures can provide a more accurate signal of whether two nodes belong to the same class/community than global measures.

Motif-aware community detection. Motifs are building blocks of complex networks [34]. Which motif is important for a complex network depends on its nature [8]. Here, we focus on the triangle motif that plays an important role in social networks [43].

Tsourakakis et al. defined the following random walk: 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 the random walker chooses a triangle that u participates in uniformly at random and then chooses an endpoint of that triangle, other than u, uniformly at random. This random walk naturally defines the notion of *triangle conductance* [51]. Benson et al. define the triangle conductance as the ratio of the number of triangles (and more generally of other motifs) cut by the partition by the minimum of motif volumes [8]. These two notions of triangle conductance differ by at most a constant, and hence are equivalent from an approximation algorithms' optimization perspective, but the two approaches differ methodologically. Both works propose to reweigh each edge e by its triangle counts t_e, and then perform spectral clustering.

Two motif-aware heuristics based on triangle counts that work well on detecting communities in real-world networks are the Tectonic heuristic [51], and the method of Satuluri et al. that we refer to as the SPR method [40]. These methods compute the Tectonic and Jaccard similarity edge scores, and retain the edges with the highest similarity scores. We describe these two similar methods in pseudocode, see Algorithm 1. The connected components of the sparsified graphs correlate well with ground-truth communities across a variety of datasets, and they speed up graph clustering algorithms without sacrificing quality as shown in [40, 51].

A key intuition that we formalize in this work is the following. Roughly speaking, if we keep the edges that Algorithm 1 removes, then we can obtain (under mild conditions) a spectral sparsifier. Furthermore, our data-driven approach explains the empirical finding of Tsourakakis et al. [51], namely that Tectonic performs better in community detection, compared to the triangle spectral clustering algorithm [8, 51].

Spectral sparsifiers. The notion of a spectral sparsifier was originally introduced by Spielman and Teng in their seminal work [48], and generalized the notion of cut sparsifiers, that were originally introduced by Karger [7, 24]. Given an undirected graph G whose

Algorithm 1 Tectonic [51] and the SPR method [40]

Input: Graph G, fraction $s\%$, threshold θ
Output: Sparsified Graphs $G_{tectonic}, G_{SPR}$
$V_{G_{tectonic}}, E_{G_{tectonic}} \leftarrow V_G, \emptyset$
$V_{G_{SPR}}, E_{G_{SPR}} \leftarrow V_G, \emptyset$
for each edge $e = (u, v)$ **do**
    Calculate Tectonic similarity $Tect_{uv}$ (Tectonic)
    Calculate Jaccard similarity J_{uv} (SPR)
end for
Sort edges such that $J_{e_1} \geq \ldots \geq J_{e_m}$ (SPR), and set $E_{G_{SPR}}$ to be the top $s\%$ edges
$E_{G_{tectonic}} \leftarrow \{(u, v) \in E : Tect_{uv} \geq \theta\}$
Return $G_{tectonic}, G_{SPR}$

Laplacian matrix representation is L_G (see Table 2), and a parameter $\epsilon > 0$, we say that a graph H is a $(1+\epsilon)$-spectral sparsifier of G if it satisfies the following property for all vectors x:

$$(1-\epsilon)x^T L_G x \leq x^T L_H x \leq (1+\epsilon)x^T L_G x \quad (1)$$

Spielman and Srivastava proved that sampling each edge e proportionally to its effective resistance $R_{eff}(e)$ (with replacement), and reweighing it appropriately if it is kept in a sample of $O(\frac{n \log n}{\epsilon^2})$ edges results in a spectral sparsifier with constant probability. Later Marcus, Spielman, and Srivastava proved constructively the existence of spectral sparsifiers with $O(\frac{n}{\epsilon^2})$ edges [32]. These algorithmic advances rely on theorems that have deep theoretical consequences [33]. Spectral sparsifiers find applications in solving Laplacian systems [?], while can be also used to speed-up cut queries in very large dense graphs.

Triangle dense subgraphs. It is a well-known fact that edge sparsity and high triangle density are properties observed across a wide variety of graph datasets [43]. This phenomenon arises naturally due to various biases/mechanisms (e.g. homophily), but also due to recommender systems. For instance Facebook and other online social networks perform triangle-based friend recommendation [49, 50]. Recently Gupta et al. [20] proved interesting structural results for triangle dense graphs. From a graph theoretic perspective, there exist various approaches to modeling triangle dense graphs. For instance, Fox et al. define the notion of a c-closed graph, where for every pair of vertices u, v with at least c common neighbors, u and v are adjacent [15]. Finally, it is worth remarking that counting triangles is an algorithmic primitive with a rich set of exact and approximation algorithms [13, 22, 37].

Theoretical preliminaries. To prove our key result, we apply the following result from concentration of measure for random matrices, see [52, Corollary 3]. The matrix norm is the operator norm.

Theorem 2.1. *Let $X_k \in \mathbb{R}^{d\times d}$ be independent random matrices such that $X_k \succeq 0$ and $\|X_k\| \leq M$ for all $1 \leq k \leq n$. Let $S_n = \sum_{k=1}^n X_k$ and $E = \sum_{k=1}^n \|\mathbb{E}[X_k]\|$. Then for every $\epsilon \in (0, 1)$ we have*

$$\mathbf{Pr}[\|S_n - \mathbb{E}[S_n]\| > \epsilon E] \leq d \cdot e^{-\frac{\epsilon^2 E}{4M}}.$$

3 PROPOSED METHOD

Triangle-dense graphs. In the following we introduce the following definition of a *triangle-dense graph*.

Definition 3.1 (Triangle-dense graph). We call a graph $G(V, E)$ triangle-dense if and only if $t_{uv} = \Theta(\min(deg(u), deg(v)))$ for each $(u, v) \in E(G)$.

Notice that always $t_{uv} \leq \min(deg(u), deg(v))$. Definition 3.1 implies that the number of triangles of an edge is asymptotically the same as the minimum degree of its endpoints. Equivalently, this implies that the ratio $\frac{t_e}{d_e}$ is constant for all edges $e \in E(G)$. In Section 4.3 (see Figure 3) we verify that it is reasonable to assume that many real-world networks are approximately triangle-dense.

Our next lemma upper bounds the effective resistance of an edge e as a function of the number of its triangles t_e.

Lemma 3.2. *The effective resistance $R_{eff}(e)$ of any edge $e \in E(G)$ satisfies the following inequality*

$$R_{eff}(e) \leq \frac{2}{t(e)+2},$$

where $t(e)$ is the number of triangles that contain edge e.

Proof. Consider an edge $e = (u, v)$. and let $S_{uv} = N(u) \cap N(v)$ be their common neighbors. Clearly, $|S_{uv}|$ is equal to $t(u, v)$, the number of triangles that edge e is contained in. Let's define the subgraph H^e to be the subgraph of G with vertex set $V_H = \{u, v\} \cup S_{uv}$ and all the induced edges among them. We will also refer to H^e as the *book of triangles* defined by edge e. Observe that $|V_H| = 2 + t(u, v)$, and $|E_H| = 1 + 2 \cdot t(u, v)$. Since H is a subgraph of G, $L_H \preccurlyeq L_G$ and therefore in the $(t(u, v) + 1)$-dimensional space spanned by L_H, $L_G^\dagger \preccurlyeq L_H^\dagger$ [4]. Recall that the effective resistance $R_{\text{eff}}(e)$ of the edge e is equal to $(e_u - e_v)^T L_G^\dagger (e_u - e_v)$ where e_i is the i-th unit vector, $i = 1, \ldots, n$. Combining the above, we obtain that

$$R_{\text{eff}}^G(e) = (e_u - e_v)^T L_G^\dagger (e_u - e_v) \leq (e_u - e_v)^T L_H^\dagger (e_u - e_v) = R_{\text{eff}}^H(e).$$

We compute the effective resistance of edge e in its book of triangles as follows. We have in parallel $t(u, v) + 1$ resistors. The resistor corresponding to edge e has resistance 1, whereas the rest that correspond to $(u, w), (w, v)$ have resistance 2. The effective conductance between the endpoints u, v is equal to the sum of the conductances in parallel, i.e., $C_{\text{eff}}^H(u, v) = \sum_{z \in S_{uv}} \frac{1}{2} + 1 = \frac{t(u,v)}{2} + 1$. Therefore the effective resistance of edge e in H is equal to

$$R_{\text{eff}}^H(e) = \frac{1}{C_{\text{eff}}^H(u, v)} = \frac{2}{t(u, v) + 2}.$$

□

Our sampling scheme is inspired by this inverse relationship between the effective resistance of an edge and its triangle counts. Specifically, we design two triangle-aware sampling schemes. Consider the following sampling scheme.

[4] where $B \preccurlyeq A$ means that for any vector $x \neq 0$, $x^T(A - B)x \geq 0$

$$p_e = \frac{q_e}{Z} \text{ where } q_e = \frac{d_e}{2+t_e}, Z = \sum_e \frac{d_e}{2+t_e}. \quad (2)$$

Algorithm 2 Triangle-aware Spectral Sparsifiers

Input: Graph G, $\epsilon \in (0,1)$
Output: Sparsifier $\tilde{G}$
for each edge $e = (u,v)$ **do**
  Compute q_{uv} according to Equation (2) ▷ Second Sparsification scheme uses Equation (3)
end for
$Z \leftarrow \sum_{e\in E} q_e$
$p_{uv} \leftarrow \frac{q_{uv}}{Z}$ for each edge $e = (u,v)$
Sample m' edges from G according to the multinomial distribution $\{p_e\}_{e\in E}$ as follows.
for each sampled edge e **do**
  if e is sampled first time **then**
    Add e to $\tilde{G}$ with weight $(m'p_e)^{-1}$
  end if
  if e is re-sampled **then**
    Increase the weight of edge e in $\tilde{G}$ by $(m'p_e)^{-1}$
  end if
end for
Return $\tilde{G}$

Our main theoretical result is stated as the next theorem, and shows that Algorithm 2 can produce under certain mild conditions a spectral sparsifier without computing explicitly effective resistances.

Theorem 3.3. *Let $G(V,E)$ be a connected, triangle-dense graph of minimum degree c, and let $\tilde{G}$ be the output of Algorithm 2 with $m' = \Theta(\frac{Z\log n}{c\epsilon^2})$ samples. Then, $\tilde{G}$ is a $(1+\epsilon)$ spectral sparsifier of G with high probability.*

The main tool we use in our analysis is Theorem 2.1, that can be seen as a Chernoff analog about the concentration of random matrices.

Proof. We associate with each edge $e = (i,j)$ a weighted version of its edge-Laplacian L_e, $R_{ij} = \frac{1}{p_{ij}} L_e = \frac{1}{p_{ij}}(e_i - e_j)(e_i - e_j)^T$. Notice that $p_{ij} > 0$ according to equation 2. Let X be the random matrix that takes value R_{ij} with probability p_{ij}. Its expectation satisfies

$$\mathbb{E}[X] = \sum_e p_e \frac{1}{p_e} L_e = \sum_e L_e = L_G.$$

Consider the average of m' samples as generated by Algorithm 2 $L_{\tilde{G}} = \frac{1}{m'}\sum_{i=1}^{m'} X_i$. Clearly, $\mathbb{E}\left[L_{\tilde{G}}\right] = L_G$. We now show that using m' samples results in strong concentration of $L_{\tilde{G}}$ around its expectation. We follow the key transformation of Spielman Srivastava [46], see also [44, Section 17.4]. Let $L_G^{+/2}$ be the square root of the pseudo-inverse of L_G, and define $\Pi = L_G^{+/2} L_G L_G^{+/2}$. Notice that Π is not the usual identity matrix, but this is only because L_G has a single eigenvalue equal to zero, and its null space is the all-ones vector. Therefore Π is the identity matrix on the range of L_G, and its trace equals $n-1$. Therefore, we can rewrite the expectation of $L_{\tilde{G}}$ as follows $\mathbb{E}\left[L_G^{+/2} L_{\tilde{G}} L_G^{+/2}\right] = \Pi$.

Define for each sample $i = 1,\ldots,m'$ the random variable $Y_i = L_G^{+/2} X_i L_G^{+/2}$. Notice that $\mathbb{E}[Y_i] = \Pi$, and its operator norm is upper bounded by $\frac{2Z}{c}$ since

$$\begin{aligned} \|Y_i\| &\le \max_{(i,j)\in E_G} \frac{1}{p_{ij}} \left(L_G^{+/2}(e_i - e_j)\right)\left(L_G^{+/2}(e_i - e_j)\right)^T \\ &= \max_{(i,j)\in E_G} \frac{Z(t_{ij}+2)}{\min(deg(i), deg(j))} \underbrace{(e_i - e_j)^T L_G^{+}(e_i - e_j)}_{=R_{eff}(i,j) \le \frac{2}{2+t_{ij}}} \\ &\le \frac{2Z}{c}. \end{aligned}$$

We apply Theorem 2.1 by observing that $L_G^{+/2} L_{\tilde{G}} L_G^{+/2} = \frac{1}{m'}\sum_{i=1}^{m'} Y_i$, and we obtain

$$\mathbf{Pr}\left[\left\|\frac{1}{m'}\sum_{i=1}^{m'} Y_i - \Pi\right\| > \epsilon\right] \le n\exp\left(\frac{-\epsilon^2 cm'}{8Z}\right).$$

When $m' \ge \frac{16Z\log n}{c}$ we obtain a $(1+\epsilon)$ spectral sparsifier with probability at least $1 - \frac{1}{n}$. □

Theorem 3.3 states that triangle density, combined with high minimum degree yields a spectral sparsifier with high probability. To see why $\tilde{G}$ has asymptotically less edges than G notice that the upper bound $\frac{2Z}{c}$ we obtain on the operator norm of each random variable Y_i is asymptotically $\Theta(\frac{m}{c})$. When the minimum degree is any (even slowly) growing function of n, i.e., $c = \omega(n)$ where $\omega(n) \longrightarrow +\infty$, then we obtain a sparsifier with significantly fewer edges. While our results are not optimal for generating spectral sparsifiers (see [6, Theorems 3,4] for how the minimum degree condition yields spectral sparsifiers for expanders), they are amenable to implementation in a distributed framework.

Let's consider also an example. The clique graph K_n is a triangle-dense graph, and $c = n-1$. Furthermore, $Z = \sum_{e\in\binom{[n]}{2}} \frac{n-1}{n-2+2} = \frac{(n-1)^2}{2}$. According to the theorem, it suffices to sample $O(n\log n/\epsilon^2)$ edges. While this example is not surprising in the sense that our sampling scheme becomes uniform sampling, our result states that when these two conditions concur, then we can use simple triangle counts to produce spectral sparsifiers.

Our second sampling scheme is inspired by Lemma 3.2. We pretend that the upper bound is tight and instead of sampling proportionally to the effective resistance as in the Spielman-Srivastava scheme, we sample proportionally to the upper bound.

$$p_e = \frac{q_e}{Z} \text{ where } q_e = \frac{2}{2+t_e}, Z = \sum_e \frac{2}{2+t_e}. \quad (3)$$

Using the same proof technique as in Theorem 3.3, we obtain the following theorem.

Theorem 3.4. *Let $G(V,E)$ be a connected, triangle-dense graph, and let $\tilde{G}$ be the output of Algorithm 2 with $m' = \Theta(\frac{n\log n}{\epsilon^2})$ samples. Then, $\tilde{G}$ is a $(1+\epsilon)$ spectral sparsifier of G with high probability.*

Table 3: Datasets used in our experiments

Name	Nodes	Edges	Triangles
Amazon [1]	334,863	925,872	667,129
BlogCatalog [39]	10,312	333,983	5,608,664
ca-HepPh [1]	11,204	117,649	3,357,890
ca-AstroPh [1]	17,903	197,031	1,350,014
ca-CondMat [1]	21,363	91,342	171,051
ca-GrQc [1]	4,158	13,428	47,779
DBLP [1]	317,080	1,049,866	2,224,385
Email-Enron [1]	33,696	180,811	725,311
Facebook [1]	4,039	88,234	1,612,010
Flickr [39]	80,513	5,899,882	271,601,126
PPI [39]	3,852	37,841	91,461
Youtube [1]	1,134,890	2,987,624	3,056,386

We omit the details as the proof follows precisely the proof of Theorem 3.3, but we outline the main difference. The number of samples needed is $O(\frac{Z \log n}{\epsilon^2})$. The normalizing constant Z is $\Theta(n)$ for triangle dense graphs. To see why, notice that there exists some constant c such that for any edge e, $t_e \geq c \min d_e$.Therefore,

$$
\begin{aligned}
Z &= \sum_{e \in E} \frac{2}{2 + t_e} \leq \sum_{e \in E} \frac{2}{2 + cd_e} \\
&\leq \sum_{e=(u,v) \in E} \left(\frac{2}{2 + c \cdot deg(u)} + \frac{2}{2 + c \cdot deg(v)} \right) \\
&= \sum_{u \in V} \frac{2deg(u)}{2 + c \cdot deg(u)} = \Theta(n).
\end{aligned}
$$

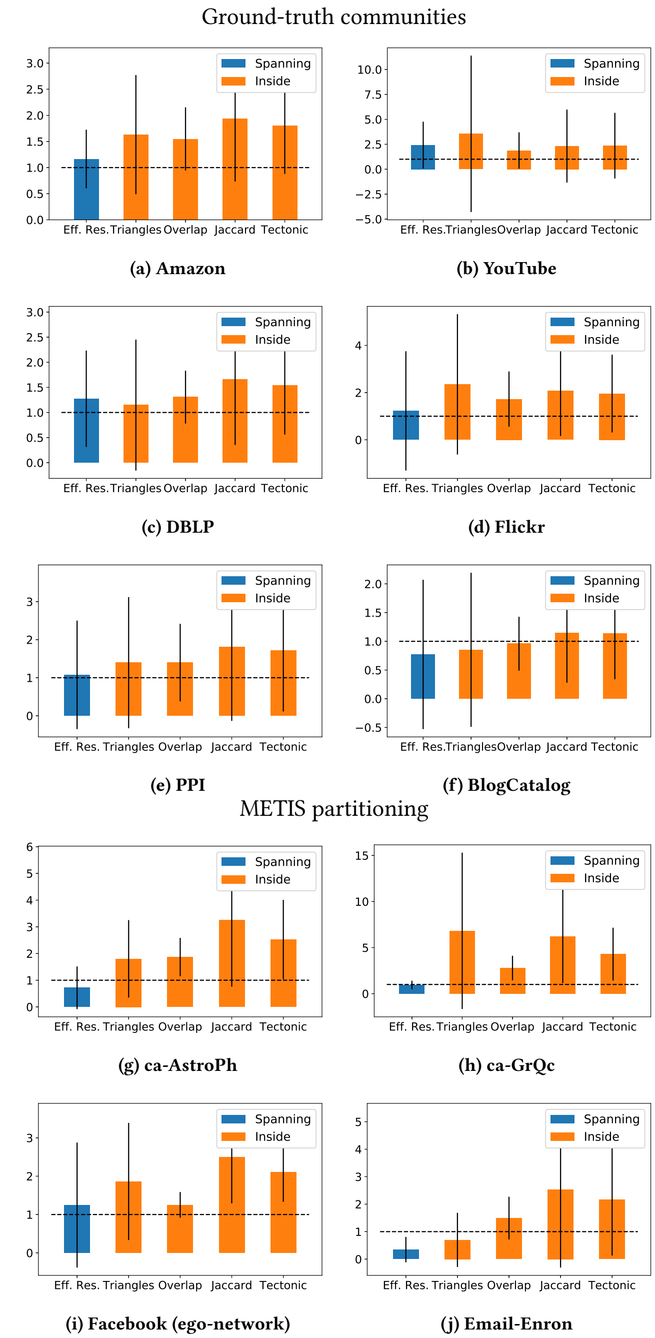
For the last step of the inequality, we use the fact that each node u is incident to $deg(u)$ edges, and each contributes a term $\frac{2}{2+c \cdot deg(u)}$.

4 EXPERIMENTATION

4.1 Experimental setup

In our experiments, we use real-world datasets of undirected networks that are summarized in Table 3. For six of those (Amazon, DBLP, YouTube, PPI, BlogCatalog and Flickr) we also have knowledge of the ground-truth communities. The networks we use are mainly social networks that observe strong triadic closure, which was the motivation of our work. Extension to other types of networks (e.g. bipartites, where other motifs should be explored) is an area for future research. If a network is disconnected, we only keep its largest connected component. All experiments were performed on a personal laptop with a 2.3GHz Dual-Core Intel i5 CPU and 16GB of memory. For smaller networks (less than 15,000 nodes) we calculate exactly the effective resistance of the edges using the formula from Table 2, and for larger networks we get a (1 ± 0.05)-approximation. We choose not to focus on reporting running times, since as Table 1 shows, our sparsification scheme using exact triangle listing is *significantly* faster than the fastest available software for approximate effective resistance computation.

Our code for the experiments is at: https://www.dropbox.com/s/0p0ybkpx19jt3ii/codeKDDTriangleAware.zip?dl=0.



(a) Amazon (b) YouTube (c) DBLP (d) Flickr (e) PPI (f) BlogCatalog (g) ca-AstroPh (h) ca-GrQc (i) Facebook (ego-network) (j) Email-Enron

Figure 2: Mean and Standard Deviation of different measures for edges *spanning* communities and edges lying exclusively *inside* a community. Values are normalized such as the lowest expected value is equal to 1. E.g., in Amazon, Jaccard Similarity for edges connecting nodes in the same community is on average two times higher than those spanning different communities. Notice the inconsistent findings for Effective Resistance.

4.2 Communities, triangles, and effective resistances

Our first task is to investigate the extent to which different measures provide a strong signal of whether an edge connnects two nodes belonging to the same community or it spans different communities. We observe that local edge information can be a better indicator than the widely accepted measure of the effective resistance for deciding whether an edge spans two communities, or is contained within a community in *real data*. Figure 2 shows our empirical findings on graphs where (i) either ground-truth communities are available, or (ii) the partitions are generated by METIS [25], a popular graph partitioning tool. For each edge $e = (u, v)$, we compute the effective resistance $R_{eff}(u, v)$, the number of triangles $t(u, v)$ containing edge (u, v), the Overlap Similarity $OC(u, v)$, the Jaccard similarity $J(u, v)$, and the closely related Tectonic similarity $Tect(u, v)$. We distinguish two categories of edges, i.e., edges within a community, and edges spanning two communities. For the effective resistance we report $\frac{R_{eff}^{spanning}}{Avg(R_{eff}^{inside})}$, while for the rest of the measures the numerator corresponds to the value of the metric for edges connecting nodes within the same community, and the denominator to the average of the measure for edges spanning different communities. This choice was made for consistency reasons; the effective resistance of an edge across two communities is according to common wisdom expected to be larger than the effective resistance of an edge inside a community. In other words, the ratio is expected to be greater than 1. Similarly, we expect that edges within communities will participate in more triangles on average than edges that span two different communities. Therefore, we expect all ratios to be greater than 1. As we see in Figure 2, for networks where the ground-truth community is known, triadic measures provide a stronger signal of whether an edges spans two communities or lies inside one community, than the effective resistance metric. Most importantly, the degree normalization performed by both Jaccard and Tectonic measures seems to help distinguish between the two cases. For example, considering the BlogCatalog dataset, averages for effective resistance and triangle counts are in contrast of what someone would expect – edges spanning communities in BlogCatalog have on average lower effective resistance and participate in more triangles, than those within a community. However, both Jaccard and Tectonic measures, more accurately distinguish between the two cases. This finding is consistent across all datasets, whether the notion of ground-truth community follows the network structure (as in networks (a), (b), (c)), as well when this is not always the case, as in networks (d), (e), and (f). These findings, therefore, explain the success of triangle-aware graph partitioning schemes, as well give insights into the structure that real-world communities follow. For instance, it is common to observe that edges connecting two communities have on average less triangles and involve at least one high degree node, hence $\frac{t(u,v)}{deg(u)+deg(v)}$ is small. This observation also explains why Tectonic and Jaccard similarity work better than the overlap similarity on real data; frequently, an edge crossing communities involves one high degree endpoint and one low degree endpoint.

In order to validate our findings even further, for each measure, we perform Welch's t-test[5] for the two sets - edges with both endpoints inside a community (A) and edges spanning communities (B). For each dataset and measure we calculate the $t-statistic$ of the test that measures the difference of the mean values of two sets normalized by the standard error. The $t-statistic$ for both Tectonic and Jaccard Similarity is consistently larger in magnitude than the one for effective resistance (with the exception of the YouTube dataset, a sparse network with few triangles too), while as pointed earlier it consistently indicates that the average of the two sets follows common wisdom, as opposed to effective resistance (e.g. for BlogCatalog). We note that we also tested similar normalizations (e.g., dividing by the sum of the endpoints) for the effective resistance metric, as we did for triangle counts, however we did not observe any improvement. The same was true for other modifications proposed for the effective resistance measure, as those laid out in Luxburg et al. [31], like subtracting the sum of the reciprocals of the degrees. Moreover, for large networks that we use an approximation of the effective resistance metric, we did not see any significant difference in consecutive runs.

For networks that we do not have knowledge of the ground truth communities, we partition the graphs using METIS, a popular graph partitioning scheme. We partition each graph into 20 almost equally sized sets. We use METIS default parameters, except for the following: We choose the best in terms of the edge-cut out of 100 different partitionings (default is 1), while we set the imbalance factor (the size of largest set, divided by the average) to be at most 1.1 (default is 1.03). As we observe in Figure 2, the edges spanning different partitions are better characterized by triadic measures (as Tectonic and Jaccard similarity), rather than the effective resistance metric. Therefore, this finding is not only true for networks with ground truth communities, but also for the partitions generated by one of the most popular graph partitioning schemes. While this finding is tied to the algorithm used by METIS to create partitions, rather the network structure itself, it is of interest, as METIS is widely used by practitioners to partition a network into different communities.

4.3 Motif-aware Spectral sparsifiers

Are real-world graphs triangle dense? In our results in Theorems 3.3 and 3.4, we relied on the assumption that the network under consideration is triangle dense, meaning that the number of triangles an edge participates is asymptotically the same as its minimum degree (at least a constant fraction of it). Though this assumption may seem idealized, in Figure 3 we see that in reality this is not at all far from true for the majority of edges in a graph. Specifically, in this figure we depict a box-plot for the Overlap Similarity of the edges, for every network used in our experiments. The lower and upper bounds of the box represent the 25 and 75-percentiles, with a line at the median, while whiskers extend from the box to show the range of Overlap similarity. What we observe is indicative of the level that real world networks adhere to the triadic closure principle: For all but three networks, the median value of overlap similarity is really high, most often close or even above 0.5.

[5] Due to space constraints, we will include the detailed crunching numbers, and statistic values in an extended version of our work.

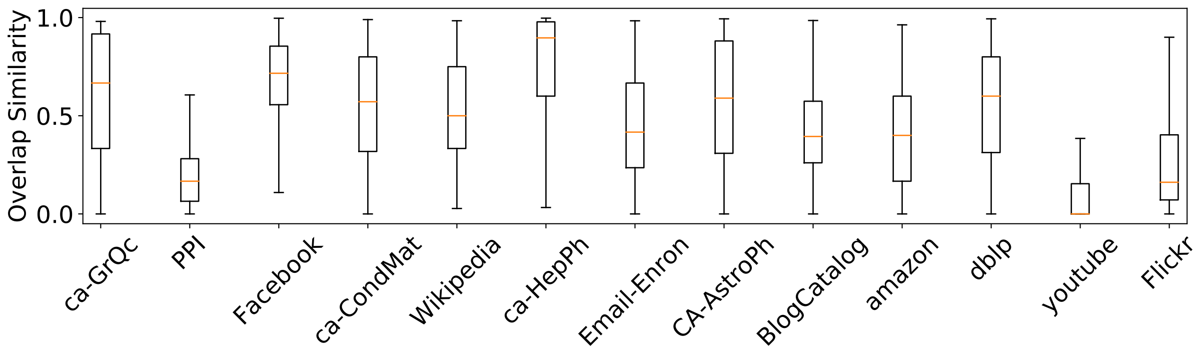


Figure 3: Box Plot for the Overlap Similarity for various networks, sorted from left to right increasingly by the number of edges.

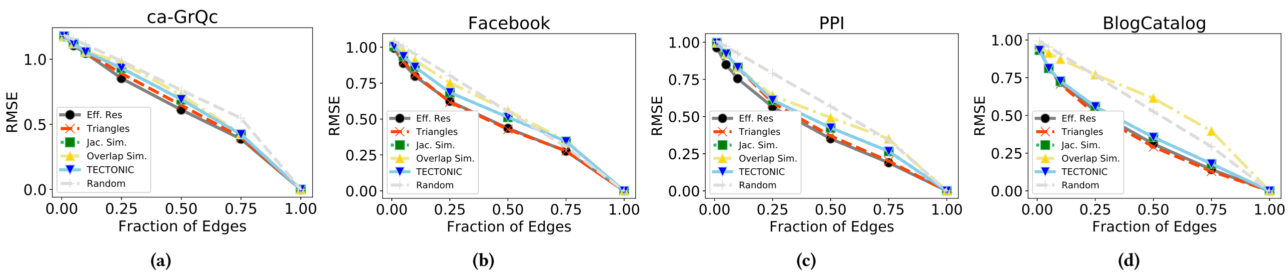


Figure 4: Root Mean Squared Error (RMSE) of the eigenvalues of the projection matrix, when including a fraction of the edges sorted by different measures

This phenomenon is especially profound for social networks. For example, in all co-authorship networks (ca-GrQc, ca-CondMat, ca-HepPh, ca-AstroPh and DBLP) the median value is above 0.5. Also, for some networks, like Facebook and ca-HepPh, even more than 75% of the edges have an Overlap Similarity value of 0.6 and above. This empirical observation strengthens the ability of motif-aware graph sparsification schemes to result into spectral sparsifiers, using a significantly smaller amount of edges than the original graph, as we will demonstrate subsequently.

Spectral Sparsification In Section 3 we explained that if the matrix $\tilde{\Pi} = L_G^{+/2} L_H L_G^{+/2}$ is close to $\Pi = L_G^{+/2} L_G L_G^{+/2}$, where H is a sparsified instance of G, then L_H is close to L_G. Moreover, for a connected undirected network, Π has $n-1$ eigenvalues equal to 1 and one eigenvalue equal to 0. As both SPR and Tectonic methods sort the edges of the graph according to a specific measure, we do the same here with the inclusion of simple triangle participation, all in increasing order, and effective resistances in decreasing order. Afterwards, we start including a fraction of those edges creating a graph H, and measure the Root Mean Squared Error (see Table 2 for the definition) between the eigenvalues of Π and $\tilde{\Pi}$. Figure 4 presents our findings for 4 representative networks (co-authorship, social and biological). Not a surprise, effective resistance is the measure that orders edges more effectively with respect to their significance in retaining the spectral information of a graph. What is surprising, however, is that for the real networks we consider, triadic measures perform comparably well. Especially, just ordering edges by their triangle counts leads to almost equally good results with effective resistance, while SPR and Tectonic that order triangle counts with some regularization (e.g. Tectonic divides the triangles an edge participates by the sum of the degrees of the endpoints of an edge) perform comparably well and always much better than a random ordering baseline, the only exception being the Overlap Similarity measure for BlogCatalog. This observation explains to a great extent the success these techniques have in graph partitioning schemes. For real world-networks, they remove edges that are important in retaining the spectral information of a graph and subsequently its connectivity. As we also see previously, edges having low score in triadic measures, not only are important for the connectivity of a graph, but also are a stronger signal of whether an edge separates communities as those observed in real networks than the effective resistance metric.

5 OPEN PROBLEMS

In this work we demonstrated the power edge triangle counts for spectral sparsification, and also shed light into the performance of two well performing community detection heuristics [41, 51]. An interesting broad question is the following. How well we can predict global graph properties from local edge and node properties on real world networks? Our work is a step towards this research direction.

6 ACKNOWLEDGEMENTS

CT acknowledges support from Intesa Sanpaolo Innovation Center. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] July 2020. Stanford Network Analysis Project. (July 2020). http://snap.stanford.edu/data/index.html.

[2] Bruno Abrahao, Sucheta Soundarajan, John Hopcroft, and Robert Kleinberg. 2014. A separability framework for analyzing community structure. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8, 1 (2014), 1–29.

[3] Vedat Levi Alev, Nima Anari, Lap Chi Lau, and Shayan Oveis Gharan. 2017. Graph Clustering using Effective Resistance. *arXiv preprint arXiv:1711.06530* (2017).

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

[5] Sanjeev Arora, Satish Rao, and Umesh Vazirani. 2009. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)* 56, 2 (2009), 5.

[6] Joshua Batson, Daniel A Spielman, Nikhil Srivastava, and Shang-Hua Teng. 2013. Spectral sparsification of graphs: theory and algorithms. *Commun. ACM* 56, 8 (2013), 87–94.

[7] András A Benczúr and David R Karger. 2000. Augmenting undirected edge connectivity in O (n^2) time. *Journal of Algorithms* 37, 1 (2000), 2–36.

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

[9] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[10] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

[11] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. 2007. On finding graph clusterings with maximum modularity. In *Graph-Theoretic Concepts in Computer Science*. Springer, 121–132.

[12] Andrei Z Broder. 1989. Generating random spanning trees. In *FOCS*, Vol. 89. Citeseer, 442–447.

[13] Norishige Chiba and Takao Nishizeki. 1985. Arboricity and subgraph listing algorithms. *SIAM Journal on computing* 14, 1 (1985), 210–223.

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

[15] Jacob Fox, Tim Roughgarden, C. Seshadhri, Fan Wei, and Nicole Wein. 2020. Finding Cliques in Social Networks: A New Distribution-Free Model. *SIAM J. Comput.* 49, 2 (2020), 448–464. https://doi.org/10.1137/18M1210459

[16] Anuththari Gamage, Eli Chien, Jianhao Peng, and Olgica Milenkovic. 2020. Multi-MotifGAN (MMGAN): Motif-Targeted Graph Generation And Prediction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4182–4186.

[17] Aristides Gionis and Charalampos E. Tsourakakis. 2015. Dense Subgraph Discovery. In *KDD*.

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

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

[20] Rishi Gupta, Tim Roughgarden, and C Seshadhri. 2014. Decompositions of triangle-dense graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*. ACM, 471–482.

[21] Ling Huang, Chang-Dong Wang, and Hong-Yang Chao. 2018. A harmonic motif modularity approach for multi-layer network community detection. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1043–1048.

[22] Madhav Jha, C Seshadhri, and Ali Pinar. 2013. A space efficient streaming algorithm for triangle counting using the birthday paradox. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 589–597.

[23] Ravi Kannan, Santosh Vempala, and Adrian Vetta. 2004. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)* 51, 3 (2004), 497–515.

[24] David R Karger. 1994. Using Randomized Sparsification to Approximate Minimum Cuts.. In *SODA*, Vol. 94. 424–432.

[25] G. Karypis and V. Kumar. 1995. Metis-unstructured graph partitioning and sparse matrix ordering system, version 2.0. (1995).

[26] James R Lee, Shayan Oveis Gharan, and Luca Trevisan. 2014. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)* 61, 6 (2014), 37.

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

[28] David A Levin and Yuval Peres. 2017. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc.

[29] Pan Li, Hoang Dau, Gregory Puleo, and Olgica Milenkovic. 2017. Motif clustering and overlapping clustering for social network analysis. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[30] Pei-Zhen Li, Ling Huang, Chang-Dong Wang, and Jian-Huang Lai. 2019. EdMot: An edge enhancement approach for motif-aware community detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 479–487.

[31] Ulrike V Luxburg, Agnes Radl, and Matthias Hein. 2010. Getting lost in space: Large sample analysis of the resistance distance. In *Advances in Neural Information Processing Systems*. 2622–2630.

[32] Adam Marcus, Daniel A Spielman, and Nikhil Srivastava. 2013. Interlacing families I: Bipartite Ramanujan graphs of all degrees. In *2013 IEEE 54th Annual Symposium on Foundations of computer science*. IEEE, 529–537.

[33] Adam W Marcus, Daniel A Spielman, and Nikhil Srivastava. 2015. Interlacing families II: Mixed characteristic polynomials and the Kadison?Singer problem. *Annals of Mathematics* (2015), 327–350.

[34] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.

[35] Federico Monti, Karl Otness, and Michael M Bronstein. 2018. Motifnet: a motif-based graph convolutional network for directed graphs. In *2018 IEEE Data Science Workshop (DSW)*. IEEE, 225–228.

[36] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.

[37] R. Pagh and Charalampos E. Tsourakakis. 2012. Colorful triangle counting and a mapreduce implementation. *Inform. Process. Lett.* 112, 7 (2012), 277–281.

[38] Leto Peel, Daniel B Larremore, and Aaron Clauset. 2017. The ground truth about metadata and community detection in networks. *Science advances* 3, 5 (2017), e1602548.

[39] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 459–467.

[40] Venu Satuluri, Srinivasan Parthasarathy, and Yiye Ruan. 2011. Local graph sparsification for scalable clustering. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 721–732.

[41] V. Satuluri, S. Parthasarathy, and Y. Ruan. 2011. Local graph sparsification for scalable clustering. In *SIGMOD '11*. 721–732.

[42] Satu Elisa Schaeffer. 2007. Graph clustering. *Computer Science Review* 1, 1 (2007), 27–64.

[43] C Seshadhri, Aneesh Sharma, Andrew Stolman, and Ashish Goel. 2020. The impossibility of low-rank representations for triangle-rich complex networks. *Proceedings of the National Academy of Sciences* 117, 11 (2020), 5631–5637.

[44] Daniel Spielman. [n. d.]. Lecture 17, Sparsification by Effective Resistance Sampling. ([n. d.]). Available at https://www.cs.yale.edu/homes/spielman/561/lect17-15.pdf.

[45] Daniel Spielman. 2021. Laplacians.jl. (2021). https://github.com/danspielman/Laplacians.jl.

[46] Daniel A Spielman and Nikhil Srivastava. 2011. Graph sparsification by effective resistances. *SIAM J. Comput.* 40, 6 (2011), 1913–1926.

[47] Daniel A Spielman and Shang-Hua Teng. 2004. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*. ACM, 81–90.

[48] Daniel A Spielman and Shang-Hua Teng. 2011. Spectral sparsification of graphs. *SIAM J. Comput.* 40, 4 (2011), 981–1025.

[49] Jessica Su, Aneesh Sharma, and Sharad Goel. 2016. The effect of recommendations on network structure. In *Proceedings of the 25th international conference on World Wide Web*. 1157–1167.

[50] Charalampos E Tsourakakis, Petros Drineas, Eirinaios Michelakis, Ioannis Koutis, and Christos Faloutsos. 2011. Spectral counting of triangles via element-wise sparsification and triangle-based link recommendation. *Social Network Analysis and Mining* 1, 2 (2011), 75–81.

[51] Charalampos E. Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. 2017. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1451–1460.

[52] Roman Vershynin. 2009. A note on sums of independent random matrices after Ahlswede-Winter. *Lecture notes* (2009).

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

[54] Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. 2017. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 555–564.