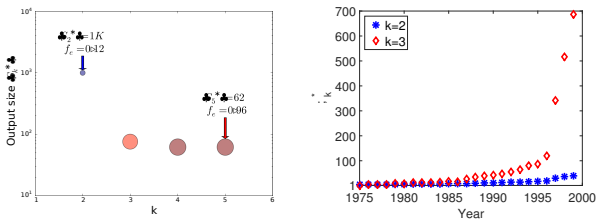
Motif-Driven Graph Analysis

Charalampos E. Tsourakakis
Harvard University
Email: babis@seas.harvard.edu

***Abstract*—In this talk I will present data-driven algorithms for *dense subgraph discovery* [11], [16], and *community detection* [18] respectively. The proposed algorithms leverage graph motifs to attack the large near-clique detection problem, and community detection respectively. In my talk, I will focus on triangles within graphs, but our techniques extend to other motifs as well. The intuition, that has been suggested but not formalized similarly in previous works, is that triangles are a better signature of community than edges. For both problems, we provide theoretical results, we design efficient algorithms, and then show the effectiveness of our methods to multiple applications in machine learning and graph mining.**



(a) (b)

Fig. 1. (a) Finding large near-cliques [16] (b) Anomaly detection in PATENTS CITATION network [11].

I. INTRODUCTION

Our work is motivated by the following high-level question: *can we design data-driven algorithms that exploit input to successfully attack computationally challenging, including NP-hard, problems?* In this talk, we will focus on effectively leveraging higher-level graph structures, known as motifs, for dense subgraph 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. For example, social networks tend to be abundant in triangles, since typically friends of friends tend to become friends themselves [19]. Triangles are also important motifs in brain networks [14]. In other networks, such as gene regulation networks, feed-forward loops and bi-fans are known to be significant patterns of interconnection [10], but our techniques extend to other such motifs as well.

II. DENSE SUBGRAPH DISCOVERY

Numerous high-impact applications rely on dense subgraph discovery [6], including anomaly detection and security, community detection in social networks, and the Web graph, detection of protein complexes in protein interaction networks, and extraction of highly correlated entities from a relevance graph. The latter type of graphs is used to encode pairwise similarities among entities, e.g., timeseries, and is ubiquitous in data science applications. Two major formulations for dense subgraph discovery are the maximum clique problem and the *densest subgraph problem* (*DSP*). The former is a notoriously hard problem, whereas the latter is poly-time solvable and lies at the core of large-scale data mining. The DSP maximizes the average degree $\frac{2e(S)}{|S|}$ over all possible subgraphs $S \subseteq V$, where V is the vertex set of the graph. Here, $e(S)$ is the number of edges induced by the vertex set S. Variants of the densest subgraph problem suitable for directed graphs exist as well [7], [5].

My work introduced the *k-clique densest subgraph problem* [16], a significant advance in routines with rigorous theoretical guarantees for scalable extraction of large near-cliques from networks [17]. The *k-clique densest subgraph problem* maximizes the average number of induced k-cliques over all possible subgraphs. The original intuition behind designing this family of objectives, which contain DSP as the special case $k = 2$, is that triangles are a better signature for community participation compared to edges. Figure 1(a) shows what we observe on a large social network with roughly 19 000 vertices and 200 000 edges and is representative of what we observe on real-world networks. As k grows, the size of the optimal sets S_k^* drops and the edge density $f_e(S_k^*) = e(S_k^*)/\binom{|S_k^*|}{2}$ grows. Notice the sudden change from $k = 2$ to $k = 3$ and that for $k = 5$ we are able to find a large near-clique on 62 nodes. The dots are scaled according to f_e. The same behavior is observed across social networks, autonomous systems, blog and Web graphs [11], [16]. An interesting open question is whether we can use stochastic graph models to explain the behavior observed in Figure 1(a).

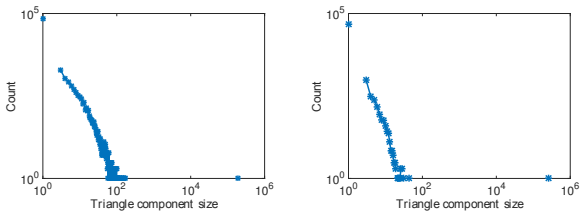


Fig. 2. Number of connected components versus size after reweighing each edge with triangle counts for (a) Amazon, and (b) DBLP. The original graphs consist of a single connected component.

III. Community Detection

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. Despite the intuition that triangles or other structures may be important for clustering and related graph problems [3], [8], there appears to be a gap in terms of useful formalizations of this idea. It is worth noting that independently from our work, a similar contribution to ours appears in Science [4].

Contributions. 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. Compared to prior work [3], 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 [1] is not necessarily a triangle expander and vice versa.
- We provide approximation algorithms for a generalization of the well-studied sparsest cut problem, where the goal now is to minimize the number of triangles cut by a partition. Our *triangle spectral algorithm*, which reweights edges by triangle counts [13] and then runs a spectral clustering algorithm [2], [12] is very practical and performs very well on real data. Also, 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
- We apply our methods to various machine learning tasks, including classification, regression, and clustering. We show that adding triangle weights to k-nearest neighbor graphs typically boosts the performance significantly. We also apply our methods to detecting communities. Using publicly available datasets where groundtruth is available, we verify the effectiveness of our framework, and show it takes orders of magnitude less time and obtains similar performance to Markov clustering (MCL).

Surprisingly, in many real-world networks we find the simple step of reweighting by triangle counts immediately disconnects the graph into numerous non-trivial connected components, that we refer as triangle components. Figure 2 shows the distribution of triangle components for the Amazon, and DBLP networks. Our findings are representative across a wide variety of networks we have experimented with: there exists one giant triangle component and then a large number of triangle components with up to few hundreds of nodes. Note that trivially all degree one nodes in the original graph become isolated components.

These findings agree with the "jellyfish" or "octopus" model [15], 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 [9] that claim that communities have size up to roughly 100 nodes. Our findings show additionally to [9] that no triangles are split between whiskers and the rest of the graph. We generalize this idea for our clustering results and experiments.

References

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

[2] 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.

[3] A. R. Benson, D. F. Gleich, and J. Leskovec. Tensor spectral clustering for partitioning higher-order network structures. In *SIAM International Conference on Data Mining, Vancouver, BC*, pages 118–126. SIAM, 2015.

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

[5] S. Bhattacharya, M. Henzinger, D. Nanongkai, and C. Tsourakakis. Space- and time-efficient algorithm for maintaining dense subgraphs on one-pass dynamic streams. In *Symposium on Theory of Computing (STOC 2015)*, pages 173–182, 2015.

[6] 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, Sydney, NSW, Australia*, pages 2313–2314, 2015.

[7] S. Khuller and B. Saha. On finding dense subgraphs. In *ICALP*, 2009.

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

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

[10] 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.

[11] 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 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 815–824. ACM, 2015.

[12] 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.

[13] R. Pagh and C. E. Tsourakakis. Colorful triangle counting and a mapreduce implementation. *Information Processing Letters*, 2012.

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

[15] 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.

[16] C. E. Tsourakakis. The k-clique densest subgraph problem. In *International conference on World Wide Web*, pages 1122–1132. ACM, 2015.

[17] C. E. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli. Denser than the densest subgraph: Extracting optimal quasi-cliques with quality guarantees. In *19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2013, pages 104–112, 2013.

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

[19] S. Wasserman and K. Faust. Social network analysis: Methods and applications. Cambridge university press, 1994.