Sampling Methods for Counting Temporal Motifs

Paul Liu
Stanford University
Stanford, California
paul.liu@stanford.edu

Austin R. Benson
Cornell University
Ithaca, New York
arb@cs.cornell.edu

Moses Charikar
Stanford University
Stanford, California
moses@cs.stanford.edu

ABSTRACT

Pattern counting in graphs is fundamental to several network science tasks, and there is an abundance of scalable methods for estimating counts of small patterns, often called motifs, in large graphs. However, modern graph datasets now contain richer structure, and incorporating temporal information in particular has become a key part of network analysis. Consequently, temporal motifs, which are generalizations of small subgraph patterns that incorporate temporal ordering on edges, are an emerging part of the network analysis toolbox. However, there are no algorithms for fast estimation of temporal motifs counts; moreover, we show that even counting simple temporal star motifs is NP-complete. Thus, there is a need for fast and approximate algorithms. Here, we present the first frequency estimation algorithms for counting temporal motifs. More specifically, we develop a sampling framework that sits as a layer on top of existing exact counting algorithms and enables fast and accurate memory-efficient estimates of temporal motif counts. Our results show that we can achieve one to two orders of magnitude speedups over existing algorithms with minimal and controllable loss in accuracy on a number of datasets.

ACM Reference Format:
Paul Liu, Austin R. Benson, and Moses Charikar. 2019. Sampling Methods for Counting Temporal Motifs. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3290988

1 SCALABLE PATTERN COUNTING IN TEMPORAL NETWORK DATA

Pattern counting is one of the fundamental problems in data mining [8, 18]. A particularly important case is counting patterns in graph data, which is used within a variety of network analysis tasks such as anomaly detection [42, 57], role discovery [19, 50], and clustering [6, 49, 59]. These methods typically make use of features derived from the frequencies of small graph patterns—usually called motifs [41] or graphlets [47] (we adopt the "motif" terminology in this paper)—and are used across a range of disciplines, including social network analysis [32, 60], neuroscience [5, 22], and computational biology [37, 46]. Furthermore, the counts of motifs have also been used to automatically uncover fundamental design principles in complex systems [37, 40, 41].

The scale of graph datasets has led to a number of algorithms for estimating the frequency of motif counts [2, 7, 12, 24, 63]. For example, just the task of estimating the number of triangles in a graph has garnered a substantial amount of attention [4, 11, 35, 52, 56, 58]. Many of these algorithms are based on sampling procedures amenable to streaming models of graph data [14, 38]. At this point, there is a reasonably mature set of algorithmic and statistical tools available for approximately counting motifs in large graph datasets.

While graphs have become large enough to warrant frequency estimation algorithms, graph datasets have, at the same time, become richer in structure. A particularly important type of rich information is time [13, 15, 21, 27, 51]. Specifically, in this paper, we consider datasets where edges are accompanied by a timestamp, such as the time a transaction was made with a cryptocurrency, the time an email was sent between colleagues, or the time a packet was forwarded from one IP address to another by a router. Accordingly, motifs have been generalized to incorporate temporal information [29, 45, 66] and have already been used in a variety of applications [30, 31, 39, 53]. However, we do not yet have algorithmic tools for estimating frequencies of temporal motifs in these large temporal graphs. This is especially problematic since including timestamps increases the size of the stored data; for example, a traditional email graph would only record if one person *has ever* emailed another person, whereas the temporal version of the same network would record *every time* there is a communication.

To exacerbate the problem, counting temporal motifs turns out to be fundamentally more difficult in a computational complexity sense. In particular, we prove that counting basic temporal star motifs is NP-complete. This contrasts sharply with stars in traditional static graphs, which are generally considered trivial to count (the number of non-induced k-edge stars with center node u is simply $\binom{d}{k}$, where d is the degree of u). Thus, our result highlights how counting problems in temporal graphs involve fundamentally more challenging computations, thus further motivating the need for approximation algorithms.

Here we develop the first frequency estimation algorithms for counting temporal motifs. We focus on the definition of temporal motifs from Paranjape et al. [45], but our methodology is general and could be adapted for other definitions. Our approach is based on sampling that employs as a subroutine any algorithm (satisfying some mild conditions) that *exactly* counts the number of instances of temporal motifs. Thus, our methodology provides a way to accelerate existing algorithms [36, 45], as well as better exact counting algorithms that could be developed in the future.

At a basic level, our sampling framework partitions time into intervals, uses some algorithm to find exact motif counts in a subset of the intervals, and weights these counts to get an estimate of

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3290988

for algorithm designers while simultaneously providing a solution for domain scientists working with large-scale temporal networks.

ACKNOWLEDGEMENTS

This research was supported in part by NSF grant DMS-1830274, NSF grant CCF-1617577, a Google research award, and a Simons Investigator Award. Additionally, we thank Eugene Y.Q. Shen for donating the computing resources that made some of the larger datasets possible.

REFERENCES

[1] The CAIDA UCSD Anonymized Internet Traces – February 17, 2011. http://www.caida.org/data/passive/passive_dataset.xml.

[2] N. K. Ahmed, N. Duffield, J. Neville, and R. Kompella. Graph sample and hold: a framework for big-graph analytics. In *Proceedings of KDD*, 2014.

[3] L. Akoglu, M. McGlohon, and C. Faloutsos. OddBall: Spotting anomalies in weighted graphs. In *Advances in Knowledge Discovery and Data Mining*. 2010.

[4] H. Avron. Counting triangles in large graphs using randomized matrix trace estimation. In *Workshop on Large-scale Data Mining: Theory and Applications*, volume 10, 2010.

[5] F. Battiston, V. Nicosia, M. Chavez, and V. Latora. Multilayer motif analysis of brain networks. *Chaos*, 27(4):047404, 2017.

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

[7] M. Bressan, F. Chierichetti, R. Kumar, S. Leucci, and A. Panconesi. Counting Graphlets: Space vs Time. In *Proceedings WSDM*. ACM Press, 2017.

[8] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang. Data mining curriculum: A proposal. *Intensive Working Group of ACM SIGKDD Curriculum Committee*, 140, 2006.

[9] H. P. Chan, N. R. Zhang, and L. H. Chen. Importance sampling of word patterns in dna and protein sequences. *Journal of Computational Biology*, 2010.

[10] R. Dobrin, Q. Beg, A.-L. Barabási, and Z. Oltvai. Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics*, 2004.

[11] T. Eden, A. Levi, D. Ron, and C. Seshadhri. Approximately counting triangles in sublinear time. *SIAM Journal on Computing*, 46(5):1603–1646, jan 2017.

[12] E. R. Elenberg, K. Shanmugam, M. Borokhovich, and A. G. Dimakis. Distributed estimation of graph 4-profiles. In *Proceedings of WWW*, 2016.

[13] M. Farajtabar, Y. Wang, M. G. Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*, pages 1954–1962, 2015.

[14] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang. On graph problems in a semi-streaming model. *Theoretical Computer Science*, 348(2-3):207–216, 2005.

[15] N. Gaumont, C. Magnien, and M. Latapy. Finding remarkably dense sequences of contacts in link streams. *Social Network Analysis and Mining*, 6(1), sep 2016.

[16] M. Gupta and J. G. Ibrahim. Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association*, 2007.

[17] S. Gurukar, S. Ranu, and B. Ravindran. COMMIT: A Scalable Approach to Mining Communication Motifs from Dynamic Networks. In *Proceedings of SIGMOD*, 2015.

[18] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[19] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. RolX: structural role extraction & mining in large graphs. In *Proceedings of KDD*, 2012.

[20] J. Hessel, C. Tan, and L. Lee. Science, askscience, and badscience: On the coexistence of highly related communities. In *ICWSM*, 2016.

[21] P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 2012.

[22] Y. Hu, J. Trousdale, K. Josić, and E. Shea-Brown. Motif statistics and spike correlations in neuronal networks. *BMC Neuroscience*, 13(Suppl 1):P43, 2012.

[23] Y. Hulovatyy, H. Chen, and T. Milenković. Exploring the structure and function of temporal networks with dynamic graphlets. *Bioinformatics*, 2015.

[24] S. Jain and C. Seshadhri. A fast and provable method for estimating clique counts using Turán's theorem. In *Proceedings of WWW*. ACM Press, 2017.

[25] R. Jin, S. McCallen, and E. Almaas. Trend motif: A graph mining approach for analysis of dynamic complex networks. In *Proceedings of ICDM*, 2007.

[26] D. Kondor, I. Csabai, J. Szüle, M. Pósfai, and G. Vattay. Inferring the interplay between network structure and market effects in Bitcoin. *New J. of Physics*, 2014.

[27] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *Proceeding of KDD*, 2008.

[28] D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos. Summarizing and understanding large graphs. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(3):183–202, may 2015.

[29] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki. Temporal motifs in time-dependent networks. *J. of Stat. Mech.: Theory and Experiment*, 2011.

[30] L. Kovanen, K. Kaski, J. Kertesz, and J. Saramaki. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *PNAS*, 2013.

[31] M. Lahiri and T. Y. Berger-Wolf. Structure prediction in temporal networks using frequent subgraphs. In *IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, 2007.

[32] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings CHI*, 2010.

[33] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *Proceedings of ICWSM*, 2010.

[34] K.-c. Liang, X. Wang, and D. Anastassiou. A sequential monte carlo method for motif discovery. *IEEE transactions on signal processing*, 56(9):4496–4507, 2008.

[35] Y. Lim and U. Kang. MASCOT: Memory-efficient and Accurate Sampling for Counting Local Triangles in Graph Streams. In *Proceedings of KDD*, 2015.

[36] P. Mackey, K. Porterfield, E. Fitzhenry, S. Choudhury, and G. Chin Jr. A chronological edge-driven approach to temporal subgraph isomorphism. *arXiv*, 2018.

[37] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 2003.

[38] A. McGregor. Graph stream algorithms: A survey. *ACM SIGMOD Record*, 2014.

[39] C. Meydan, H. H. Otu, and O. Sezerman. Prediction of peptides binding to MHC class I and II alleles by temporal motif mining. *BMC Bioinformatics*, 2013.

[40] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 2004.

[41] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 2002.

[42] C. C. Noble and D. J. Cook. Graph-based anomaly detection. In *Proceedings of KDD*, 2003.

[43] A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[44] P. Panzarasa, T. Opsahl, and K. M. Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *J. of the American Society for Information Science and Technology*, 60(5):911–932, 2009.

[45] A. Paranjape, A. R. Benson, and J. Leskovec. Motifs in temporal networks. In *Proceedings of WSDM*. ACM Press, 2017.

[46] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, jan 2007.

[47] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, jul 2004.

[48] F. Reid and M. Harrigan. An analysis of anonymity in the bitcoin system. In *Security and Privacy in Social Networks*, pages 197–223. 2012.

[49] K. Rohe and T. Qin. The blessing of transitivity in sparse and stochastic networks. *arXiv*, 2013.

[50] R. A. Rossi and N. K. Ahmed. Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1112–1131, apr 2015.

[51] I. Scholtes. When is a network a network?: Multi-order graphical model selection in pathways and temporal networks. In *Proceedings of KDD*, 2017.

[52] C. Seshadhri, A. Pinar, and T. G. Kolda. Triadic measures on graphs: The power of wedge sampling. In *Proceedings of SDM*, 2013.

[53] H. Shao, M. Marwah, and N. Ramakrishnan. A temporal motif mining approach to unsupervised energy disaggregation: Applications to residential and commercial buildings. In *Proceedings of AAAI*, 2013.

[54] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 2002.

[55] R. Siddharthan. Phylogibbs-mp: module prediction and discriminative motif-finding by gibbs sampling. *PLoS computational biology*, 2008.

[56] L. D. Stefani, A. Epasto, M. Riondato, and E. Upfal. TRIÈST: Counting Local and Global Triangles in Fully Dynamic Streams with Fixed Memory Size. *ACM Transactions on Knowledge Discovery from Data*, 11(4):1–50, jun 2017.

[57] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. GraphScope: Parameter-free Mining of Large Time-evolving Graphs. In *Proceedings of KDD*, 2007.

[58] C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos. DOULION: counting triangles in massive graphs with a coin. In *Proceedings of KDD*, 2009.

[59] C. E. Tsourakakis, J. Pachocki, and M. Mitzenmacher. Scalable motif-aware graph clustering. In *Proceedings of WWW*, 2017.

[60] J. Ugander, L. Backstrom, and J. Kleinberg. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *Proceedings of WWW*, 2013.

[61] T. Viard, M. Latapy, and C. Magnien. Computing maximal cliques in link streams. *Theoretical Computer Science*, 609:245–252, jan 2016.

[62] T. Viard, C. Magnien, and M. Latapy. Enumerating maximal cliques in link streams with durations. *Information Processing Letters*, 133:44–48, may 2018.

[63] P. Wang, J. Zhao, X. Zhang, Z. Li, J. Cheng, J. C. Lui, D. Towsley, J. Tao, and X. Guan. MOSS-5: A fast method of approximating counts of 5-node graphlets in large graphs. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[64] Q.-M. Zhang, L. Lü, W.-Q. Wang, and T. Z. and. Potential theory for directed networks. *PLoS ONE*, 2013.

[65] X. Zhang, S. Shao, H. E. Stanley, and S. Havlin. Dynamic motifs in socio-economic networks. *Europhysics Letters*, 108(5):58001, dec 2014.

[66] Q. Zhao, Y. Tian, Q. He, N. Oliver, R. Jin, and W.-C. Lee. Communication motifs: a tool to characterize social communications. In *Proceedings of CIKM*, 2010.