BLANT: Sampling Graphlets in a Flash

Wayne Hayes[1], Sridevi Maharaj[2]

***Abstract* — BLAST is an invaluable tool to compare and align sequences. Its efficiency comes from *k*-mers—short *k*-length subsequences—that are indexed across a larger corpus, which allows identical *k*-mers from different regions to aid finding longer matches. Analogously, networks could be indexed by *k*-node graphlets. Unfortunately, existing graphlet counting methods enumerate *all* the graphlets, which is infeasible on large networks. We introduce BLANT (*Basic Local Alignment Network Tool*) which randomly samples and indexes graphlets. BLANT samples *millions* of graphlets in seconds, which aids search and local alignment via indexing, and provides a statistical sample of both global graphlet distribution and local orbit degree vectors.**

***Keywords* — biological network alignment, local network alignment, network database, network classification, network search, network function, network topology.**

I. Purpose

Networks are used to represent biological interactions such as protein-protein, gene-uRNA, brain connectomes, and enzymes; their topology (the structure of connectivity between nodes) is related to function [1]. Matching local structures also helps identify similar functional modules in other larger networks. In order to find these modules and understand the details of their structural components, several graph topological features have been studied but none appear to give as robust results as graphlets [2]. Graphlets have been used to quantify the local structure of biological networks via global alignments, alignment-free comparison, analysis of brain connectomes, and in recovering functional and phylogenetic information [3].

Existing graphlet counting methods [4] perform exhaustive enumeration of graphlets and are infeasible on large networks. We propose that statistical sampling [5] can produce a satisfactory approximation. Here, we introduce BLANT, which samples and indexes millions of graphlets in seconds. We show that the sampled distribution agrees with the true graphlet distribution.

II. Method – Node Based Expansion

A *k*-graphlet is an induced, connected subgraph of *k* nodes taken from a graph *G(V, E)*. We construct a sampled *k*-graphlet *g* as follows. Initially, we select an edge (u_1, u_2) uniformly at random from *G* and add u_1 and u_2 to *S*, the set of nodes that will become *g*. We iteratively add nodes to *S* by picking from all nodes outside *S* adjacent to a node inside *S*, until we have *k* nodes.

[1,2]Department of Computer Science, University of California-Irvine, USA. E-mail: {whayes, sridevi.m}@uci.edu

III. Experimental Results

A. We sampled 10,000 *k*-graphlets (taking a fraction of a second) for $k = \{3,4,5\}$ from a total of 1540 synthetic networks of Geometric, Erdös-Rényi, Scale Free, Small World and Sticky graphs of varying sizes (1000, 2000, 4000, 6000 nodes) and densities (0.005, 0.0075, 0.01). We computed their full graphlet counts using ORCA [4], which took weeks of CPU time. Figure 1 shows that the mean relative graphlet frequency from our method agrees well with the true mean relative graphlet frequency both in accuracy and intrinsic variation.

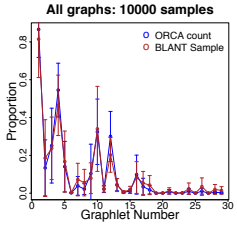

Figure 1

B. We also sampled 10^7 7-graphlets from Enzyme, Brain-ADHD, Gene-µRNA, and Facebook networks. Multi-dimensional scaling on pairwise graphlet correlation distances obtained from our graphlet sample shows that sampling clearly distinguishes between different network types (Figure 2) as well as exhaustive enumeration [6].

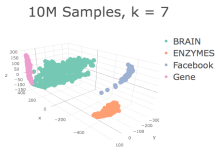

Figure 2

IV. Conclusion

Random graphlet sampling is orders of magnitude faster than existing exhaustive enumeration methods and produces distributions of graphlets that are close to the true distribution. This promises to revolutionize network analysis by allowing graphlet analyses on networks of arbitrary size.

References

[1] Davis, Darren, et al. "Topology-function conservation in protein–protein interaction networks." *Bioinformatics* 31.10 (2015): 1632-1639.

[2] Hayes, Wayne, Kai Sun, and Nataša Pržulj. "Graphlet-based measures are suitable for biological network comparison." *Bioinformatics* 29.4 (2013): 483-491

[3] Kuchaiev, Oleksii, et al. "Topological network alignment uncovers biological function and phylogeny." *Journal of the Royal Society Interface* (2010): rsif20100063.

[4] Hočevar, Tomaž, and Janez Demšar. "Combinatorial algorithm for counting small induced graphs and orbits." *PloS one* 12.2 (2017): e0171428

[5] Hasan, Adib, Po-Chien Chung, and Wayne Hayes. "Graphettes: Constant-time determination of graphlet and orbit identity including (possibly disconnected) graphlets up to size 8." *PloS one* 12.8 (2017): e0181570

[6] Yaveroğlu, Ömer Nebil, et al. "Revealing the hidden language of complex networks." *Scientific reports* 4 (2014): 4547

Nothing should be here on page 2! Please limit your abstract to a single page, and create a one-page .pdf file for submission.