

Implementation of a spanning trees sampling algorithm to detect 4-motifs

Gam A., Yudina M.

Omsk State Technical University, Omsk, Russia
mail@mano.pro, mg-and-all@mail.ru

Key words: motifs, computation method, network analysis, graph theory

Motivation and Aim: Approximate counting all possible subgraphs is the task connected with some important network analysis methodologies like network motifs detection. Path sampling algorithms [1-4] seem to be the most promising among other statistical algorithms. Path sampling algorithms do not use heuristic assumptions, unlike graph generalization algorithms like TNP algorithm [5]. Path sampling algorithms make it possible to estimate subgraphs frequencies, in contrast to algorithms based on random walk estimate only subgraph concentrations. Finally, path sampling algorithms are implementations of the Monte Carlo method and allow one to obtain consistent, unbiased, effective statistical estimates, in contrast well known Rand-ESU algorithm [6] that does not meet the requirements in the general case [7].

Methods and Algorithms: However, the proposed in [1–3] path sampling algorithm does not allow working with directed graphs. As shown in our work [4], statistical experiments using all possible spanning trees are necessary to obtain frequency estimates of all possible subgraphs. The direction of the edges of spanning trees is not taken into account. In a graph, on three vertices there is only one isomorphism class of spanning trees (Fig. 1), on four vertices there are 2 classes, on five there are three, on six there are six. Different values to estimate a subgraph frequency of the same subgraph can be obtained, since the subgraph can be found by the sampling of different spanning trees (Fig. 2).

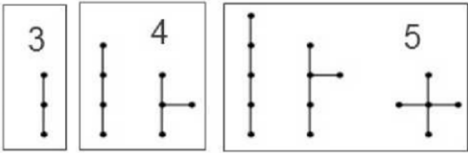


Fig. 1. Spanning trees with three, four and five vertices

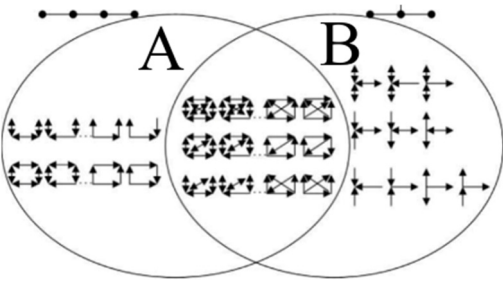


Fig. 2. Venn diagram of the coverage of the spanning trees of subgraphs belonging to different classes of isomorphism; A is the set of subgraphs that are covering by "4-path" spanning tree, B is the set of subgraphs that are covering by "4-branch" spanning tree

Results: The spanning trees sampling algorithm is a development of path sampling algorithms. It allows you to analyze not only undirected graphs, but also directed ones. The spanning trees sampling algorithm is implemented in the MFSViewer program. The program (<https://github.com/MNYudina/MFSView>) allows you to detect network motifs. Moreover, the use 68–95–99.7 rule makes it possible to identify network motifs statistically reliably. The program interface is shown in Fig. 3. The graphical interface, the help and the license of MFSViewer are similar to FANMOD ones [6]. However, MFSViewer differs from FANMOD in speed, accuracy [4] and quality indicators of the obtained statistical estimates (such as unbiased and efficiency).

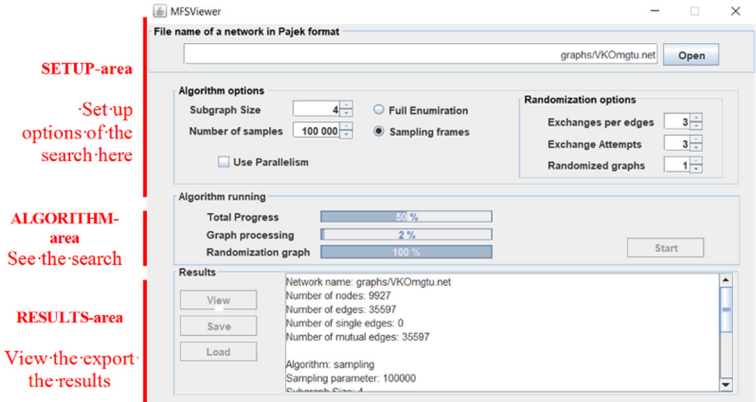


Fig. 3. The three areas of the main interface window of MFSViewer. They correspond to the three working steps when using the tool: setting up the algorithm, running the algorithm, and processing the results

Conclusion: Despite the fact that compared to the analogues we know, such as FANMOD, Kavosh (Cytoscape), etc., our implementation looks competitive, we continue to work on improvements. At the moment, work is underway on colored motifs.

References

1. Seshadhri C. et al. Triadic measures on graphs: The power of wedge sampling. In: Proceedings of the SIAM International Conference on Data Mining. 2013:10-18.
2. Jha M., Seshadhri C., Pinar A. Path sampling: A fast and provable method for estimating 4-vertex subgraph counts. In: Proceedings of the 24th International Conference on World Wide Web. 2015:495-505.
3. Wang P. et al. MOSS-5: A fast method of approximating counts of 5-node graphlets in large graphs. In: IEEE Transactions on Knowledge and Data Engineering. 2018:30(1):73-86.
4. Yudina M.N., Zadorozhnyi V.N., Yudin E.B. Mixed Random Sampling of Frames method for counting number of motifs. *J Physics: Conference Series*. 2019:1260(2):022013.
5. Przulj N. Biological network comparison using graphlet degree distribution. *Bioinformatics*. 2007:23(2):177-183.
6. Wernicke S., Rasche F. FANMOD: A tool for fast network motif detection. *Bioinformatics*. 2006:22(9):1152-1153.
7. Yudina M.N. Assessment of accuracy in calculations of network motif concentration by Rand ESU algorithm. *J Physics: Conference Series*. 2019:1260(2):022012.