The indicator $R'$ of the statistical significance for subgraph for motif discovery task

M N Yudina, A V Gam
Omsk State Technical University, pr. Mira 11, 644050, Omsk, Russia

E-mail: udinev@asoju.com, dean@mano.pro

Abstract. The analysis of the statistical significance of subgraphs is carried out in dealing with solving the problem of detecting network motif in graphs of large networks. A new generalizing indicator of the significance of subgraphs is proposed. The indicator is intended to detect network motif when statistical methods are used to calculate the frequencies of occurrence of subgraphs. In fact that the subgraph in the studied network graph occurs more often than in the randomized versions of this graph (a sign of a network motif) is established statistically reliably.

Key words: frequencies of network motifs, statistical sampling algorithms, motif discovery, efficient unbiased estimations, randomized enumerate subgraphs uniformly

1. Introduction
Studying of the general properties of large networks is engaged in a new branch of science – (Network Science) Twenty years ago much attention was paid to the structural characteristics of the network as a whole (the effect of scale-free [1], small-world phenomenon, the six degrees of separation concept [2]), nowadays a lot of attention is paid to local structural interactions. After all, it could be said about the network, no less.

One of the approaches is developed in Network Science is to identify so-called network motif in network graphs. Network motif on $k$-vertices ($k$- motif) are revealed by comparing the frequencies of occurrence of connected $k$-vertex subgraphs belonging to different classes of isomorphism in the studied network graph and in its randomized versions. Analysis of large networks based on the identification of network motif has led to the discovery of new protein interactions [3], new methods for diagnosing cancer [4], and new approaches to control information exchange in technical networks [5], and to the identification of ‘hidden features’ social networks.

The prevalence of certain network exact motif is a feature of any large network that distinguishes it from other networks, primarily from networks of a different nature. So that, the identification of the FULLY CONNECTED TRIAD- motif is a sign of graphs of information networks of the Internet, BI-FAN- and FEED-FORWARD LOOP- motif are characteristic for graphs of molecular networks of a cell (figure 1), in graphs of trofic webs they find THREE-CHAIN- and BI-PARALLEL- motif, etc.

The main task of finding k-motif in the graph G can be split into three following subtasks.
1. Obtaining the set $G_R$ of randomized versions of graph G.
2. Counting in the graph G and in each graph of the set $G_R$ the frequencies of occurrence of subgraphs belonging to different classes of isomorphism. In this case, the problem of determining the isomorphism classes for the found subgraphs on $k$ vertices is solved.
3. Detecting of network motif based on their occurrence in the graph $G$ and in the graphs of the set $G_R$.
The third subtask is solved by using various indicators of the statistical significance of the subgraphs.
Traditionally [6, 7], researchers of large networks determine network motif based on the indicators of
the statistical significance of the subgraphs $R$ and $Z$-score.

![Examples of network motif](image)

**Figure 1.** Examples of network motif

The indicator of statistical significance $R$ is defined as the ratio of the absolute frequency of occurrence $n_G$ of investigated subgraph in the network graph into the average value of the absolute frequency of occurrence of the same subgraph $M[n_{GR}]$ in a process of randomized versions of the graph:

$$R = \frac{n_G}{M[n_{GR}]}.$$  \hspace{1cm} (1)

The standard score ($Z$-score) is identified from the following equation:

$$Z\text{-score} = \frac{n_G - M[n_{GR}]}{\sigma[n_{GR}]}.$$  \hspace{1cm} (2)

the are $M[n_{GR}]$ and $\sigma[n_{GR}]$ the mean and standard deviation of the absolute frequency of occurrence of the subgraph in randomized versions of the graph.

Network motif are called those subgraphs with belong to isomorphism classes with a $Z$-score exceeding certain threshold value. Subgraphs for each of the $Z$-score $> 2$ and $R > 1.1$ are called network motif [7].

2. Statement of the problem

The presses of the detecting network motif is rather hard, more over the calculating occurrence of subgraphs is the most intensive part of it. An exact calculation of the occurrence of subgraphs was substantiated when this approach [6] was taken shape in 2002, for example, graphs of protein networks containing from several hundred to several thousand nodes were studied. Now the investigated graphs of the networks of protein interactions contain hundreds of thousands of nodes, an exact calculation of the frequencies of occurrence of subgraphs turns out to be impossible. Also, the identification of network motif is used in the analysis of Internet networks, which are often larger in size than the networks of protein interactions. So that, the Table 1 shows us the time of calculating the frequencies of occurrence in some graphs of Internet networks. Data on the graphs Google Plus, Twitter, Email Enron, Email_AU-All are obtained from the Stanford University graph database (http://snap.stanford.edu/data/). The VKOmgtu graph is described in [7]. The absence of data in the corresponding cells of table 1 means that when using the program, it was not possible to get the result in five days of work. The most famous program was used, moreover, suitable for calculating precisely large networks - the Fanmod program, as well as the igraph library for the software environment $R$.

The increase in the size of the graphs of the studied networks led to the development of statistical calculation methods that allows obtaining statistical estimates for the frequencies of occurrence of subgraphs (table 1). It should be noted that most of the statistical methods do not allow the study of mixed graphs, but are designed to carry out calculations only in undirected graphs. In addition, many methods do not allow obtaining estimates of the absolute frequencies of occurrence of subgraphs, but only estimates of their relative frequencies of occurrence.

| Graph    | Library igraph for the software | FANMOD |
|----------|---------------------------------|--------|

**Table 1.** Time of calculation (in seconds) the prevalence episode of subgraphs in graphs of networks
environment $R$

| Subgraph at the three node | Subgraph at the four node | Subgraph at the three node | Subgraph at the four node |
|---------------------------|---------------------------|---------------------------|---------------------------|
| Google Plus               | 39442                     | -                         | -                         |
| Twitter                   | 76                        | 35217                     | 1076                      |
| Email Enron               | 23                        | 10882.83                  | 27                        | 11646                     |
| Email AU-All              | 40                        | 38508                     | 136                       | 272158                     |
| VKOmgtu                  | 0.22                      | 19                        | 1                         | 93                         |

Table 2. Method of the statistical calculation the frequencies of occurrence of subgraphs time

| The Method’s name       | Link for the file’s start | The Year of creation | Limit for the size of subgraph | Calculation frequency distribution like the relatives and absolute($\tilde{C}_{i}^{A}/\tilde{N}_{i}^{A}$) | Possibility of calculation at the graphs’ combined |
|-------------------------|----------------------------|----------------------|--------------------------------|------------------------------------------------------------------------------------------------|-------------------------------------------------|
| ESA [8]                 | [9]                        | 2004                 | No                             | $\tilde{C}_{i}^{A}$                                                             | Yes                                              |
| RAND-ESU [10]           | [11]                       | 2005                 | No                             | $\tilde{N}_{i}^{A}$                                                             | Yes                                              |
| GUISE [12]              | [13]                       | 2012                 | 5                              | $\tilde{C}_{i}^{A}$                                                             | No                                               |
| Wedge Sampling [14]     |                            | 2013                 | 3                              | $\tilde{N}_{i}^{A}$                                                             | No                                               |
| GRAFT [16]              | [17]                       | 2014                 | 5                              | $\tilde{C}_{i}^{A}$                                                             | No                                               |
| $k$-profile sparsifier [18] | [19]                   | 2016                 | 4                              | $\tilde{C}_{i}^{A}$                                                             | No                                               |
| MOSS-5 [20]             | [21]                       | 2018                 | 5                              | $\tilde{N}_{i}^{A}$                                                             | No                                               |
| MFS [22]                | [23]                       | 2019                 | 3, 4                           | $\tilde{N}_{i}^{A}$                                                             | Yes                                              |

Make a note, that the usage of relative frequencies of occurrence of subgraphs instead of absolute frequencies of occurrence in formulas (1) and (2) is incorrect (although it is used without explanation in many software products, such as FANMOD, MFINDER). In this case, another value is calculated. The reason for the significant difference between these values is that the total number of all subgraphs in the graph under study is usually very different from the total number of all subgraphs in the randomized versions of this graph. For most of the networks we study, the number of subgraphs in a randomized graph increases, as a rule, by an order of magnitude.

On the other hand, even for methods that allow one to obtain estimates of the absolute frequencies of occurrence of subgraphs, the substitution of statistical estimates in formulas (1) and (2) is also an approximate operation. After all, statistical methods have a certain margin of error. There is a need to develop such indicators of statistical significance that will allow identifying network motifs using statistical methods to calculate the frequencies of occurrence of subgraphs.

3. Indicator $R'$

We would like to propose a generalization of the indicator $R$, which can be used to obtain statistical estimates for the frequencies of occurrence of subgraphs. This can be done by the MFS method [22] or the RAND-ESU method implemented in the well-known programs FANMOD [11] and igraph [24]. One of the main features of these methods is the ability to obtain not only point, but also interval estimates for the frequencies of occurrence of subgraphs, both in the investigated graph and in its randomized version. We could identify that obtaining a statistical estimate in the investigated and randomized graph has three options and they are possible (figure 2), in which the value of the point estimate for the frequency of occurrence of a subgraph in the investigated graph exceeds the value of the estimate for the frequency of occurrence in the randomized graph (which corresponds to the concept of a network motif):
1) the point estimate \( \hat{n}_R \) for the frequency of occurrence of a subgraph in a randomized graph does not fall into the confidence interval \( \hat{n}_G \pm 3\sigma_G \) for the frequency of occurrence in the investigated graph, but the confidence intervals \( \hat{n}_G \pm 3\sigma_G \) and \( \hat{n}_R \pm 3\sigma_R \) have common points;

2) the confidence intervals for the frequency of occurrence of the subgraph in the randomized graph \( \hat{n}_R \pm 3\sigma_R \) and in the investigated graph \( \hat{n}_G \pm 3\sigma_G \) have no common points;

3) a point estimate for the frequency of occurrence of a subgraph in a randomized graph \( \hat{n}_R \) falls within the confidence interval \( \hat{n}_G \pm 3\sigma_G \) for the frequency of occurrence in the studied graph, but the confidence intervals \( \hat{n}_R \pm 3\sigma_R \) and \( \hat{n}_G \pm 3\sigma_G \) have common points.

Figure 2. Possible identifying point surgical and interval estimate of subgraph into searching graph and its randomized version

We could introduce indicator, let us choose the strictest relation between the obtained statistical estimates for the frequencies of occurrence of the subgraph in the network graph and in its randomized version (case 2). Let us define the significance index \( R' \) of the subgraph by the following formula:

\[
R' = \frac{M[\hat{n}_G - 3\sigma_G]}{M[\hat{n}_R + 3\sigma_R]}
\]

where \( \hat{n}_G, \hat{n}_R \) is the estimates of the absolute frequencies of occurrence of the subgraph, obtained in the calculation in the graph and in its randomized version, respectively; \( \sigma_G, \sigma_R \) are the corresponding standard deviations characterizing the accuracy of the estimates obtained. The proposed estimate generalizes the known indicator \( R \) and shows that the number of subgraphs in the investigated graph ‘statistically significantly’ exceeds this number in its randomized versions.

Identifying network motifs in large graphs we can use the calculation only one estimate for the frequency of occurrence of subgraphs in graph \( G \) and one in its randomized version, then expression (3) will be written as

\[
R' = (\hat{n}_G - 3\sigma_G) / (\hat{n}_R + 3\sigma_R).
\]

4. Experimental results

From the Table 3 we can shows you the results of the study of network motifs in the directed graph GenReg, the calculation of \( R' \) was performed in the MFSView program. The GenReg graph describes the regulatory network of genes involved in the development of adrenal cortex cancer in humans [25].

One calculation of the frequencies of occurrence of subgraphs by the GenReg graph and 100
calculations by its randomized versions were used. Indicators of the statistical significance of the subgraphs corresponding to network motifs are highlighted in gray.

Table 3. Results of the study of network motifs in the GenReg graph

| id | \(\hat{n}_G\)     | \(M[\hat{n}_R]\)  | \(R\)          | \(R'\)         | Z-Score       |
|----|----------------|----------------|--------------|--------------|-------------|
| 3  | 9,1750E+08    | 1,0711E+09    | 0,8566      | 0,7944      | -0,2125     |
| 7  | 2,4150E+09    | 2,5841E+09    | 0,9346      | 0,8934      | 0,1104      |
| 12 | 2,4994E+09    | 2,8199E+09    | 0,8864      | 0,8426      | -0,0296     |
| 13 | 6,9143E+09    | 8,8490E+09    | 0,7814      | 0,7619      | -0,4433     |
| 14 | 2,6494E+08    | 9,1315E+07    | 2,2613      | 1,9320      | 4,9079      |
| 17 | 2,1207E+08    | 8,8373E+07    | 2,3997      | 2,0491      | 5,5049      |
| 19 | 4,1816E+07    | 3,9063E+07    | 1,0705      | 0,6887      | 0,6199      |
| 20 | 8,9347E+06    | 1,2634E+06    | 7,0722      | 2,6975      | 36,8311     |
| 24 | 4,4803E+09    | 4,7995E+09    | 0,9335      | 0,9055      | 0,0953      |
| 29 | 1,2684E+09    | 2,2750E+09    | 0,5575      | 0,5221      | -1,2077     |
| 35 | 1,4025E+08    | 5,1712E+07    | 2,7122      | 2,2189      | 6,1270      |
| 41 | 5,2174E+09    | 6,9174E+09    | 0,7542      | 0,7315      | -0,5620     |
| 42 | 1,3556E+08    | 7,0856E+07    | 1,9132      | 1,5896      | 3,5735      |
| 47 | 3,7075E+08    | 1,7072E+08    | 2,1717      | 1,9336      | 4,6475      |
| 52 | 3,4448E+07    | 3,8433E+07    | 0,8963      | 0,5603      | -0,3249     |
| 58 | 1,0032E+07    | 1,4154E+06    | 7,0874      | 2,8405      | 26,2518     |
| 66 | 3,1451E+07    | 1,8959E+07    | 1,6589      | 0,9470      | 1,9731      |
| 67 | 1,3481E+07    | 2,0614E+06    | 6,5397      | 2,9901      | 43,4739     |
| 76 | 6,8123E+09    | 7,2798E+09    | 0,9358      | 0,9154      | 0,0830      |
| 77 | 6,5371E+08    | 2,4802E+08    | 2,6357      | 2,4031      | 5,7933      |
| 78 | 1,2740E+07    | 1,1991E+06    | 10,6245     | 4,2876      | 38,4280     |
| 80 | 1,5866E+07    | 2,1128E+06    | 7,5096      | 3,5385      | 32,9139     |
| 82 | 2,2976E+07    | 2,5780E+06    | 8,9125      | 4,5779      | 44,1505     |
| 83 | 6,2056E+06    | 5,4260E+04    | 114,3679    | 14,5793     | 31,4693     |

The Table 4 shows us the results of calculating statistical estimates for the frequencies of occurrence of subgraphs in a fragment of the social network Vkontakte [7]. As you can see from Table 4 when using 'standard values' \(R > 1,1, R' > 1,1\) the same network motifs are revealed as when using \(Z\)-score \(> 2\), these are subgraphs with identifiers 142, 203, 205, 217.

The received results show that the proposed indicator of significance \(R'\) generalizes the known indicator of significance \(R\) and can be used to identify network motifs when using statistical methods to calculate the frequencies of occurrence of subgraphs. The peculiarity of the proposed indicator \(R'\) is that it allows you to statistically reliably identify network motifs even by two statistical calculations of the frequencies of occurrence of subgraphs: one is in the investigated network graph, the other is in its randomized version.

5. Conclusion

The article is discussed the well-known statistical indicators of the significance of the subgraphs \(R\) and \(Z\)-Score, which are used to identify the so-called network motifs in the network graph. Significance indicators \(R\) and \(Z\)-Score, by definition, use the results of calculating the absolute frequencies of occurrence of subgraphs in the graph under study and in its randomized versions. However, in a number of programs, developers use these indicators, ‘replacing’ the absolute frequencies of occurrence with relative frequencies of occurrence. In other cases, the exact values of the frequencies of occurrence are ‘substituted’ by their statistical estimates. At the same time, the indicators of statistical significance themselves do not develop, but various meanings of using existing indicators
are determined, which even for specialists can remain ‘hidden’. At this writing investigates the change in the meaning of the significance indicator of the subgraph $R$, if statistical methods are used to calculate the frequency of occurrence instead of exact calculation. The proposed significance index $R'$ generalizes the known significance index $R$ and allows statistically reliable identification of network motifs when using statistical methods for calculating the frequencies of occurrence of subgraphs.

References

[1] Barabasi A L, Albert R 1999 Emergence of scaling in random networks Science 286 509–512.
[2] Newman M E J 2003 The Structure and Function of Complex Networks SIAM Review 45 167–12.
[3] Albert I, Albert R 2004 Conserved network motifs allow protein–protein interaction prediction Bioinformatics 20 3346–6.
[4] Chen L, et al. 2004 Identification of breast cancer patients based on human signaling network motifs. Scientific Reports 2004 3368 1–17.
[5] Zeng T, Semiari O, Saad W 2018 Spatio-temporal motifs for optimized vehicle-to-vehicle (v2v) communications International Conference on Computing, Networking and Communications (ICNC) 789–6.
[6] Milo R 2002 Network motifs: Simple building blocks of complex networks Science 594 824–4.
[7] Yudina M N 2016 Nodes in social networks: central measures and role in network processes Omsk Scientific Bulletin. Ser. Devices, machines and technologies 148 161–6.
[8] Kashtan N, et al. 2004 Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs Bioinformatics. 20 1746–13.
[9] Alon U 2003 Network Motif Software, URL: https://www.weizmann.ac.il/mcb/UriAlon/download/network-motif-software/.
[10] Wernicke S, Rasche F 2006 FANMOD: a tool for fast network motif detection Bioinformatics 22 1152–2.
[11] Wernicke S 2006 FANMOD: a tool for fast network motif detection, URL: http://theinf1.informatik.uni-jena.de/~motifs/.
[12] Bhuiyan M A, Rahman M, Hasan M A 2012 Guise: Uniform sampling of graphlets for large graph analysis Proceedings of the 12th International Conference on Data Mining (ICDM) 91–9.
[13] Bhuiyan M A, Rahman M, Hasan M A 2012 GUISE: Uniform Sampling of Graphlets for Large Graph Analysis, URL: https://github.com/DMGroup-IUPUI/GUISE-Source.
[14] Seshadhri C, Pinar A, Kolda T G 2013 Triadic measures on graphs: The power of wedge sampling Proceedings of the SIAM International Conference on Data Mining SIAM 10–9.
[15] Kolda T, Pinar A, Seshadhri C Triadic Measures on Graphs: The Power of Wedge Sampling, URL: http://www.sandia.gov/~tgkolda/feastpack/.
[16] Rahman M, Bhuiyan M A, Hasan M A 2014 Graft: An efficient graphlet counting method for large graph analysis IEEE Transactions on Knowledge and Data Engineering 26 2466–13.
[17] Rahman M, Bhuiyan M A., Hasan M A GRAFT: an approximate graphlet counting algorithm for large graph analysis, URL: https://github.com/DMGroup-IUPUI/GRaFT-Source.
[18] Elenberg E R, et al. 2016 Distributed estimation of graph 4-profiles Proceedings of the 25th International Conference on World WideWeb 483–11.
[19] Elenberg E R 2016 GraphLab PowerGraph implementation of 4-profile counting, URL: https://github.com/eelenberg/4-profiles.
[20] Wang P, et al. 2018 MOSS-5: A fast method of approximating counts of 5-node graphlets in large graphs IEEE Transactions on Knowledge and Data Engineering 30 73–14.
[21] Wang P 2018 MOSS-5: Fast Method of Approximating Counts of 5-Node Graphlets in Large Graphs, URL: http://nskeylab.xjtu.edu.cn/dataset/phwang/code/mosscode.zip.
[22] Yudina M N, Zadorozhnyi V N, Yudin E B Mixed Random Sampling of Frames method for counting number of motifs Journal of Physics: Conference Series 2019 1260 022013
[23] Yudina M N 2020 MFSView, URL: https://github.com/MNYudina/MFSView.
[24] igraph – Network analysis software, URL: https://igraph.org/