Local Topology of Social Network Based on Motif Analysis

Krzysztof Juszczyszyn, Przemysław Kazienko, Katarzyna Musiał

Wroclaw University of Technology,
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{krzysztof, kazienko, katarzyna.musial}@pwr.wroc.pl

Abstract. Network motifs – small subgraphs that reflect local topology can be used to discover general profile and properties of the network. Analysis of motifs for the large social networks derived from email communication is presented in the paper. The distribution of motifs in all analyzed real social networks is very similar one another and can be treated as the network fingerprint. This property is most distinctive for stronger human relationships.

1 Introduction

When investigating the topological properties and structure of complex networks we must face a number of complexity–related problems. In large social networks, tasks like evaluating the centrality measurements, finding cliques, etc. require significant computing overhead. In this context the methods, which proved to be useful for medium and small networks fail when applied to very large structures. This also refers to complex biological or technology–based networks like computer networks, WWW, gene transcription networks.

The outcomes of the research on network motif detection and analysis in large email–based social network of the Wroclaw University of Technology, consisting of over 5,700 nodes and 140,000 edges (Fig. 1) are presented in this paper. The local structure of this large social network has been investigated by analyzing interconnection patterns between small sets of nodes, called motifs. These small motifs reflect general local topology profile and the properties of the entire network.

2 Related Work

Complex networks, both biological and engineered, were analyzed with respect to so–called *network motifs* [6]. They are small (usually 3 to 7 nodes in size) subgraphs which occur in the given network far more (or less) often then in the equivalent random networks (in terms of the number of nodes, node degree distribution, average path length, clustering, etc). Despite all these structural and statistical similarities, networks from different fields have very different local topological structure. It was recently shown that concentration of network motifs may help to distinguish and classify complex biological, technical and social networks [9]. We can distinguish so–called *superfamilies* of networks [9], which correspond to the specific *significance profiles* (SPs). To create SP for the given network, the concentration of individual motifs is measured and compared to their concentration in a number of random networks. The statistical significance of motif M is defined by its Z–score Z_M:

$$Z_M = \frac{n_M - \left\langle n_M^{rand} \right\rangle}{\sigma_M^{rand}} \qquad (\mathbf{1})$$

where n_M is the frequency of motif M in the given network, $\left\langle n_M^{rand} \right\rangle$ and σ_M^{rand} are the mean and standard deviation of M's occurrences in the set of random networks, respectively [3]. Most algorithms for detecting network motifs assume exhaust enumeration of all subgraphs with a given number of nodes in the network. Their computational cost dramatically increases with the network size. However, it was recently show that it is possible to use random sampling to effectively estimate concentrations of network motifs. The algorithm presented in [4] is asymptotically independent of the network size and enables fast detection of motifs in very large networks with hundreds of thousands of nodes and larger.

The existence of network motifs affects not only topological but also functional properties of the network. For biological networks, it was suggested that network motifs play key information processing roles [11]. For example, so–called FFL motif – Feed–Forward Loop (motif no. 5 in Fig. 2) has been shown both theoretically and experimentally to perform tasks like sign–sensitive filtering, response acceleration and pulse–generation [7]. Such results reveal that, in general, we may conclude about function and properties of very large networks from their basic building blocks [8].

In another work, motif analysis was proved to have ability of fast detection of the small–world and clustering properties of the large network [2]. This result open promising but still unexplored possibilities of reasoning about network's global properties with sampling of local topological structures.

Very little research has been done on motifs in computer science and sociology. SPs for small social networks (<100 nodes) were studied in [9]. A web network counting 3.5x10^5 nodes [1] was used to show the usability of sampling algorithm [4].

3 Motif Analysis Applied to Large Social Network

To discover properties of large social networks using the motif analysis, some important issues should be respected. The key parameter that reflects the significance of motifs is Z–score (Eq. 1). It is based on comparing the actual concentration of subgraphs (motifs) in the considered network with their concentration in a set of random networks. The size of this set should be as small as possible, so we determined what number of random networks is required to detect motifs with the given accuracy. The actual profile of the network is expressed by the set of Z–scores of the motifs. In our case, we checked all the directed three–node subgraphs (triads). Their concentration values for all triads form so–called *Triad Significance Profile* of the network (TSP) [9]. An email-based social network is the directed, weighted graph so it differs from WWW, gene transcription or molecular networks (unweighted graphs). The weights of the edges in the social network depend on the intensity of communication. However, to enable analysis the domain of edge weights need to be discretized – only small set of classes can be analyzed.

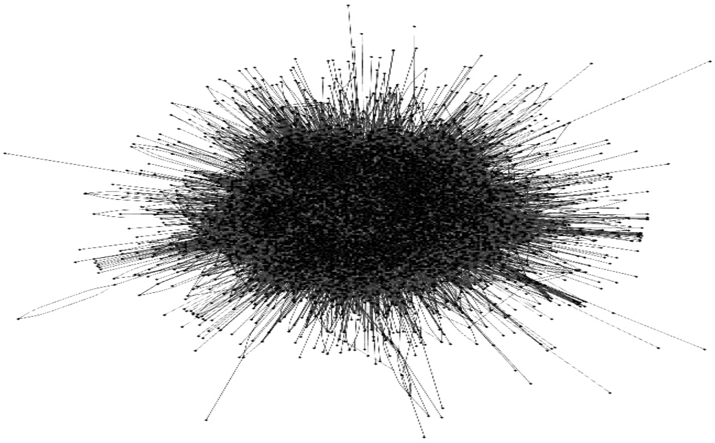

Fig. 1. Social network discovered from the email communication between employees of WUT.

3.1 Extraction of the Social Network from Email Communication

The experiments were carried out on the logs from the Wrocław University of Technology (WUT) mail server, which contain only the emails incoming to the staff members as well as organizational units registered at the university (Fig.1) [5]. All experiments were performed with FANMOD tool [13, 14] dedicated for motif detection in large networks.

First, the data cleansing process was executed. The bad email addresses was removed from the analysis and the duplicated ones were unified. Additionally, only emails from and to the WUT domain were left.

Note that although every single email provides information about the sender activity, it can simultaneously be sent to many recipients. An email sent to only one person reflects strong attention of the sender directed to this recipient. The same email sent to 10 people does not respect each individual recipient so much. For that reason, the strength of email communication $S(x, y)$ from email user x to y has been defined in the following way:

$$S(x,y) = \sum_{i=1}^{card(EM(x,y))} \frac{1}{n_i(x,y)}, \quad \textbf{(2)}$$

where: $EM(x,y)$ – the set of all email messages sent by x to y; $n_i(x,y)$ – the number of all recipients of the ith email sent from x to y. In consequence, every email with more than one recipient is treated as $1/n$ of a regular one (n is the number of its recipients).

The strength of the relationship $RS(x,y)$ between x and y is calculated as follows:

$$RS(x,y) = \frac{S(x,y)}{n(x)}, \qquad (3)$$

where: $n(x)$ – the total number of emails sent by user x, was introduced. The values of this function are from range [0,1]. The similar approach was utilized by Valverde et al. to calculate the strength of relationships. It is established as the number of emails sent by one person to another [12]. However, the authors do not respect the general activity of the given individual. In our approach, this general, local activity exists in the form of denominator in Eq. 3. Based on the *RS* values the email-based social network has been created. The next step of preprocessing is to convert continuous weights of ties, i.e. *RS*(*x*,*y*) into five classes (Table 1). The ranges of *RS* (classes) were established to balance the number of edges in each class.. Note also that every node in the network (except only one node in class 2 and 5) belongs to every class since it is incident to at least one edge from every single class. Having classes extracted, the separate networks were built based on the edges from one or more classes. These networks were used for motifs detection in order to check how the TSP profile changes when different communication intensity is considered.

The general statistics about the extracted classes are presented in Table 1. Note that in the class 1 where the strength of the relations is the lowest, the contribution of mutual edges is the smallest – only 1.2% whereas this rate for class 5 – only the strongest connections (*RS*>0.05) was as much as 16.2%. It means that stronger human relationships tend to be more frequently mutual compared to the weaker ones.

Each of the classes separately as well as their various combinations were utilized in the process of detecting triads within the WUT email social network. There are 13 different motifs that consist of three nodes each (Fig. 2). Their ID=1,2,…,13 are used in the further descriptions interchangeably with the corresponding M1, M2,…, M13.

Table 1. The number of nodes and edges in the particular classes.

	Class 1	Class 2	Class 3	Class 4	Class 5	All
Range of *RS*	(0;0.001]	(0.001;0.005]	(0.005;0.01]	(0.01;0.05]	(0.05;1]	(0;1]
No. of nodes	5,783	5,782	5,783	5,783	5,782	5,783
No. of single edges (% of total in class)	30,788 (98.8%)	33,755 (91.6%)	18,800 (93.6%)	28,249 (85.2%)	13,039 (83.8%)	124,631 (91.1%)
No. of mutual edges (% of total in class)	382 (1.2%)	3,086 (8.4%)	1,283 (6.4%)	4,919 (14.8%)	2,529 (16.2%)	12,199 (8.9%)
Total no. of edges	31,170	36,841	20,083	33,168	15,568	136,830

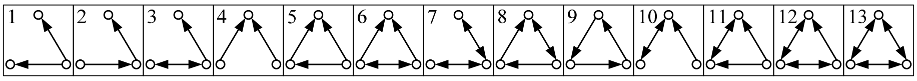

Fig. 2. Directed triads and their IDs that can exist within the social network.

3.2 Triad Significance Profile in Relation to Number of Random Networks

The goal of the first phase of the experiments was to determine the minimum number of random networks which allow to detect the motifs with the required accuracy.

Triad Significance Profiles (TSPs) for all motifs (Eq. 1) were computed separately for the different numbers of random networks (RN), Fig. 3. Note that there are only little differences between obtained Z–scores for the numbers of random networks above 50. It appears that 100 random networks provide sufficient accuracy of calculations. To be sure 500 random networks were used in further research.

We may also conclude, that the considered network reveals the typical property of social networks – the small–world phenomenon. Loosely connected motifs with only 2 edges, like M2, M3, M4, M7, M10 occur less frequently compared to the random networks. It obviously proves the high clustering level, i.e. high probability that two neighboring nodes have connected neighbors. The only exception is M1 which is met relatively often. This reflects specific property of large mail–based social network: there are relatively many broadcasting nodes who spread messages (news, announcements, bulletins) which are never answered.

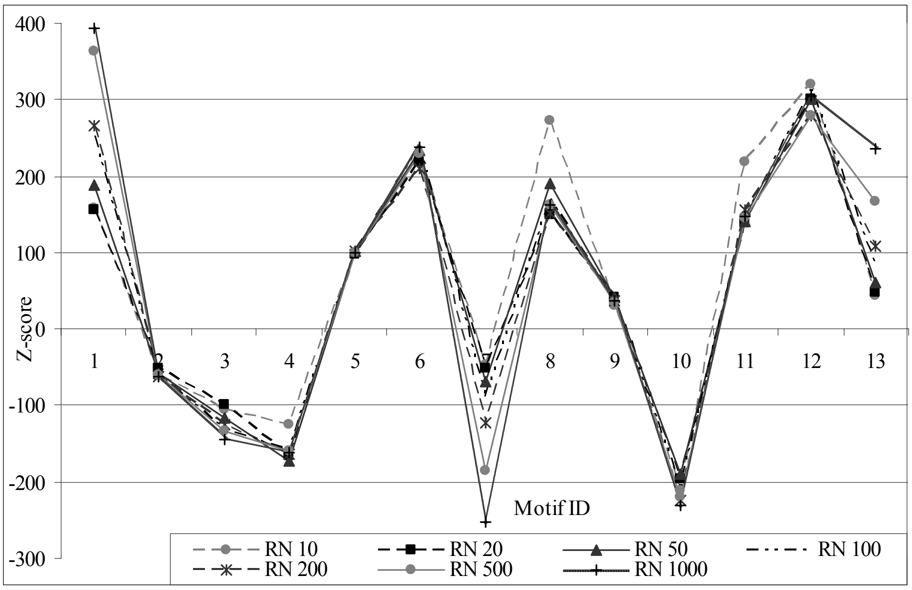

Fig. 3. Triad Significance Profile (Z-score values) of the WUT email-based social network for different numbers of random networks.

3.3 Triad Significance Profile in Relation to Strength of the Ties

Triad Significance Profile for separated and aggregated classes of ties' strengths are presented in Table 2 and Fig. 4. "Class 12345" stands for entire network (all classes merged), while "Class 5" denotes only the subnetwork created by the strongest ties characterized by intense communication (Table 2). We clearly see that the high positive Z-score of M13, big negative Z-score of M7 and decreasing with the growing strength of the ties Z-score of M1 may be called markers of social small-world networks and suggest increasing clustering level (Fig. 4). Also, the change for M1's

Z-score from positive to negative values met when passing to subnetworks composed of stronger ties shows the fading of broadcast-type communication in strongly connected subnetworks. We also see that the subgraphs composed of the stronger ties are “more connected” and clustered then their lightly connected counterparts.

Table 2. *Z*–score values for particular motifs in the WUT social network for different classes extracted based on the relationship strength.

Motif ID	1	2	3	4	5	6	7	8	9	10	11	12	13
Class 12345	363.9	–60.8	–134.3	–158.6	98.8	230.3	–184.5	161.8	35.7	–219.6	144.2	279.8	167.4
Class 2345	210.2	2.0	–234.4	–105.8	143.5	418.7	–236.4	180.1	40.3	–189.5	180.8	360.8	237.4
Class 345	–131.6	–58.6	–496.5	155.6	372.7	533.3	–630.0	233.0	57.4	–256.9	404.8	589.5	651.1
Class 45	–223.8	–109.7	–464.4	288.9	400.3	487.7	–972.3	196.9	35.3	–269.5	427.7	606.4	1279.0
Class 5	–305.3	–160.4	–307.5	500.6	318.1	312.1	–589.5	94.3	13.4	–293.5	382.6	335.1	1754.3
Class 4	–150.6	–207.3	–220.9	–61.9	358.2	247.5	–228.9	257.1	154.6	–219.1	229.5	235.1	541.1
Class 3	6.3	–86.0	–36.0	–96.4	122.2	54.2	–29.5	55.5	65.0	–47.6	45.5	36.0	90.4
Class 2	87.4	–77.4	–29.5	–76.7	68.4	46.8	–44.1	52.6	44.6	–46.9	32.0	35.4	56.6
Class 1	50.7	–32.2	–35.6	–42.0	32.2	37.3	–26.4	13.2	10.6	–14.5	9.4	14.1	90.8

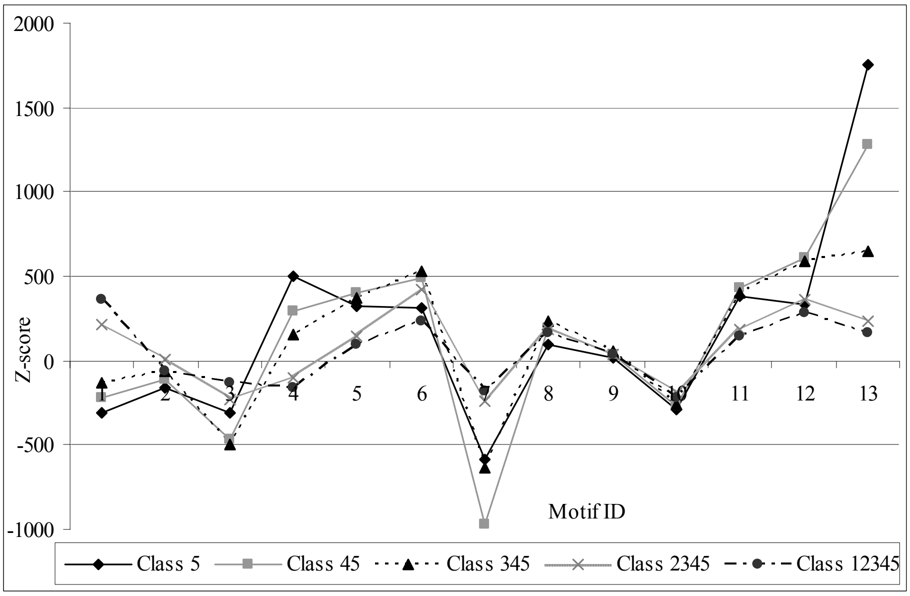

Fig. 4. Triad Significance Profile of the WUT email-based social network for different classes of relationship strength.

The above conclusions are additionally confirmed by TSPs for separate classes of the ties (Fig. 5). They reveal the growing importance of M13 and M7 for subnetworks of strong ties (Class 4 and 5) even in more convincing way. There is also one more observation concerning M4, which is dual-edge triad with two unidirectional edges pointing to the same node (Fig.2). Only for Class 5, we observed positive Z-score of

M4, which is unusual for the investigated social network as a whole – compare with TSP of other classes in Fig. 4 and 5 as well as Table 2. This can be interpreted as above-statistical frequent occurrence of *receiver nodes* – we may treat them as of *executives*. They intensively receive reports while simultaneously being involved in dense clusters and frequent communication activities inside small-world network structures. Note that this effect is characteristic only for intensively communicating subnetwork and it has not been detected in the full communication graph, i.e. based on all classes of ties. It is also absent in TSPs of small social networks and WWW network studied in [9].

Overall, social networks created upon different classes and their various combinations belong to the same *superfamily* of networks [9] – their TSPs have the same shape regardless the range of relationship strengths *RS* (Fig. 3, 4, and 5). However, this general profile is most noticeable for stronger human relationships, especially within Class 5. In other words, Class 5 containing only 11% of all edges (Class 12345) but as much as over 20% of total mutual ties (Table 1) appears to be is the most distinctive representative of the entire social network. Hence, Triad Significance Profile for strong human relationships can be treated as the small, condensed pattern that reflect the character of the entire social network.

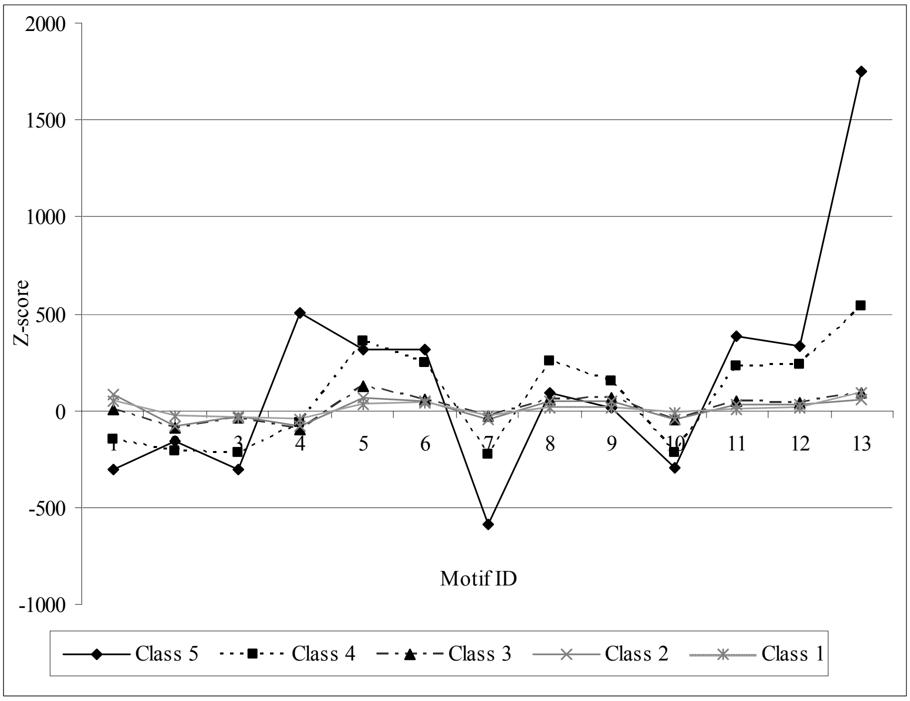

Fig. 5. Triad Significance Profile of the WUT social network for different classes of the relationship strength.

4 Conclusions

Network motif analysis is the fast method to discover general profile of the entire network in the compact form since small motifs reflect patterns of the common local topology. In the email-based social networks, motifs preserved their distribution for all analyzed networks. Stronger relationships in the email-based social network are more mutual. Moreover, the motif-based fingerprint (TSP) is more distinctive for the network created from the stronger relationships. Besides, intensive communication results in greater frequency of more complex motifs and greater clustering level.

Further research will focus on multirelational social networks [10] as well as on dynamical properties of motif analysis in large social networks. New fast algorithms for motif detection may be applied to periodically gathered communication logs in order to discover time-related changes in large, evolving social structures.

Acknowledgments. The work was supported by The Polish Ministry of Science and Higher Education, grant no. N516 037 31/3708.

References

1. Barabasi A.-L. Albert R. (1999) Emergence of scaling in random networks. Science, 286, 509–512.
2. Chung-Yuan H., Chuen-Tsai S., Chia-Ying C., Ji-Lung H. (2007) Bridge and brick motifs in complex networks, Physica A, 377, 340–350.
3. Itzkovitz S., Milo R., Kashtan N., Ziv G., Alon U. (2003) Subgraphs in random networks. Physical Review E., 68, 026127.
4. Kashtan N., S. Itzkovitz S., Milo R., Alon U. (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics, 20 (11), 1746–1758.
5. Kazienko P., Musiał K., Zgrzywa A. (2008) Evaluation of Node Position Based on Email Communication. Control and Cybernetics, to appear.
6. Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U. (2002) Network motifs: simple building blocks of complex networks. Science, 298, 824–827.
7. Mangan S. Alon U. (2003) Structure and function of the feedforward loop network motif. Proc. of the National Academy of Science, USA, 100 (21), 11980–11985.
8. Mangan S., Zaslaver A. Alon U. (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J. Molecular Biology, 334, 197–204.
9. Milo R., Itzkovitz S., Kashtan N., Levitt R., Shen-Orr S., Ayzenshtat I., Sheffer M., Alon U. (2004) Superfamilies of evolved and designed networks. Science 303(5663), 1538–42.
10. Musiał K., Kazienko P., Kajdanowicz T. (2008) Multirelational Social Networks in Multimedia Sharing Systems. Chapter 18 in N.T.Nguyen, G.Kołaczek, B.Gabryś (eds.) Knowledge Processing and Reasoning for Information Society, EXIT, Warsaw, 275-292.
11. Shen-Orr S., Milo R., Mangan S., Alon U. (2002) Network motifs in the transciptional regualtion network of Escherichia coli. Nat. Genet., 31, 64–68.
12. Valverde S., Theraulaz G., Gautrais J., Fourcassie V., Sole R.V. (2006) Self-organization patterns in wasp and open source communities. IEEE Intelligent Systems 21 (2), 36–40.
13. Wernicke.S. (2006) Efficient detection of network motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 3 (4), 347–359.
14. Wernicke S., Rasche F. (2006) FANMOD: a tool for fast network motif detection. Bioinformatics, 22 (9), 1152–1153.