Patterns of Interactions in Complex Social Networks based on Coloured Motifs Analysis

Katarzyna Musial[1], Krzysztof Juszczyszyn[1], Bogdan Gabrys[2], Przemysław Kazienko[1]

Wroclaw University of Technology[1], Bournemouth University[2]
katarzyna.musial@pwr.wroc.pl, krzysztof.juszczyszyn@pwr.wroc.pl,
bgabrys@bournemouth.ac.uk, przemyslaw.kazienko@pwr.wroc.pl

Abstract. Coloured network motifs are small subgraphs that enable to discover and interpret the patterns of interaction within the complex networks. The analysis of three-nodes motifs where the colour of the node reflects its high – white node or low – black node centrality in the social network is presented in the paper. The importance of the vertices is assessed by utilizing two measures: degree prestige and degree centrality. The distribution of motifs in these two cases is compared to mine the interconnection patterns between nodes. The analysis is performed on the social network derived from email communication.

1 Introduction

The investigation of the communication patterns in the complex social networks is a very resource-consuming task. The methods that are quite useful and effective in small and medium sized social networks fail while applying them to the complex networks. In the case of large networks the solution to the complexity problem, which occurs when the models of interaction are studied, can be the analysis of the local network structures, also known as network motifs. Motif analysis stems from bioinformatics and theoretical biology [10],[13], where it was applied to the investigation of huge network structures like transcriptional regulatory networks, gene networks or food webs [7],[8]. Although the global topological organization of metabolic networks is well understood, their local structural organization is still not clear. At the smallest scale, network motifs have been suggested to be the functional building blocks of network biology. So far several interesting properties of large biological network structures were reinterpreted or discovered with help of motif analysis [14],[16].

In this work we apply this biologically-inspired set of methods to the analysis of email-based social network of the size similar to many networks observed in nature. Motif analysis offers low computational overhead and opportunity to gain an insight into the local structure of huge networks which otherwise would require prohibitive computations to investigate. Moreover, we go one step beyond the classic motif analysis and propose distinguishing network nodes with respect to their unique properties, in this case centrality values (as defined in social sciences). The discovered motifs and their numbers enable to assess which patterns of communication appear often in the large social networks and which are rather rare.

The former ones can be seen as these which come into existence in a natural way whereas the latter ones can be treated as the artificial and unnatural. Moreover, the nodes of these subgraphs can be divided into different classes based on the values of various kinds of measures used in social network analysis. These additional information can be used in order to mine new knowledge from the data about interaction between users. The outcomes of the research on two–coloured, three–nodes network motifs (triads) detection and analysis in large email–based social network of the Wroclaw University of Technology (WUT), consisting of over 5,700 nodes and 140,000 edges are presented in this paper. The colours were assigned to the vertices in such a way that the black colour of the node reflects its low centrality whereas the white colour its high centrality. The frequency of occurrence and the distribution of the individual motifs serve as the basis to define the interaction patterns between users. Moreover, depending on the values of centrality measures, we can investigate in what kind of motifs (patterns of relations) people are embedded.

2 Related Work

2.1 Network Motifs

Complex networks, both biological and engineered, have been shown to display so–called network motifs [10]. They are small (usually 3 to 7 nodes) subgraphs which occur in given network far more/less often then in corresponding random networks. In order to evaluate the distribution of motifs their concentration is measured for a set of random networks then compared with the network being investigated. For computational complexity reasons, the size of the random network set should be as small as possible. For example, our former research has revealed that 100 random networks provide sufficient accuracy of calculations in the case of the WUT social network [5]. The statistical significance of a given motif is defined by its Z–score Z_M:

$$Z_M = \frac{n_M - \left\langle n_M^{rand} \right\rangle}{\sigma_M^{rand}}, \tag{1}$$

where: n_M is the number of occurrences of motif M in the network, $\left\langle n_M^{rand} \right\rangle$ and σ_M^{rand} are the mean and standard deviation of its appearances in the set of random networks [4]. The calculated for each motif Z–score measure forms the basis to create the significance profile (*SP*) of the network. It has been recently shown that distribution of network motifs may help to distinguish and classify complex biological, technical and social networks [9]. Each class of these networks has its own specific significance profile. Two methods of motif detection can be utilized. The first one assumes exhaustive enumeration of all subgraphs with a given number of nodes in the network. Note that, their computational cost dramatically increases with the network size. The second one is to use random sampling to effectively estimate concentrations of network motifs. The algorithm presented in [6] is asymptotically

independent of the network size and allows fast detection of motifs in very large networks (hundreds of thousands of nodes and larger).

Network motifs can be used to describe both topological and functional properties of various networks. For biological networks it was suggested that network motifs play key information processing roles [13]. For example, so–called Feed–Forward Loop (motif number 5 on Fig. 1) has been shown both theoretically and experimentally to perform tasks like sign–sensitive filtering, response acceleration and pulse–generation [7]. Such results show that, generally, we may reason about function and properties of very large networks from their basic building blocks [8]. In another work by Chung-Yuan at al. motif analysis was proved to have the ability of fast detection of the small–world and clustering properties of a network [3]. Within the area of computer science and social networks very little work has been done with motifs. In [9] SPs for small (below 100 nodes) social networks were demonstrated. Counting 3.5×10^5 nodes WWW network described in [2] was used to show the effectiveness of sampling algorithm in [6]. First results for large e-mail based network were presented in [5].

2.2 Centrality Measures in Social Network

Both the nodes and the edges of the network motifs can be divided into different classes (colours) based on the freely chosen measure. In this research the nodes will be coloured with respect to their centrality. Two methods – degree centrality and degree prestige – will be utilized to colour the vertices in the e-mail social network in the process of triads detection. Degree centrality $DC(x)$ of a member x takes into account the number of outdegree of member x [11], [12]. It means that a node is more important than another one when it communicates with greater number of network members. It is usually expressed by the number of neighbours that are adjacent to the given person. The degree prestige is based on the indegree number so it takes into account the number of members that are adjacent to a particular member of the community [15]. In other words, more prominent people are those who received more nominations from members of the community [1]. The degree prestige $DP(x)$ of a member x is the number of members from the first level neighbourhood that are adjacent to x.

3 Coloured Motifs Analysis

3.1 Data Preparation and Plan of the Experiments

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. Based on the information about the communication between the employees at the WUT the e-mail based social network was extracted. In such a network the nodes are unique e-mail addresses and the edges reflect the fact that there exists any communication between two distinct e-mail addresses. First, the data cleansing process was executed. Only emails from and to the WUT domain were left. After that the centrality of vertices

was assessed. In order to perform that task, two methods were utilized: degree centrality and degree prestige which were calculated as the number of respectively outgoing and incoming edges from or to the given node. In the next phase, the values of each calculated centrality measure were assigned to one of two classes. These classes were created based on the mean value of the given centrality measure evaluated for WUT social network. In both cases of degree centrality and degree prestige it was 25 (see Table 1). Note that around 70% of WUT employees have the degree centrality lower than the mean value and around 62% have the degree prestige lower than the mean value (Table 1).

Table 1 The number of nodes in each of the created classes.

	Degree Centrality $DC(x)$		Degree Prestige $DP(x)$	
	$DC(x)>25$	$DC(x)\leq 25$	$DP(x)>25$	$DP(x)\leq 25$
No. of nodes	1736	4047	2165	3618
% of all nodes	30.02%	69.98%	37.44%	62.56%

The further stages of experiments were divided into two parts. Each of the parts is concerned with one of the centrality measure. To each network node one of two colours: white or black, was assigned. The black colour is assigned to the vertices that degree centrality is lower or equal 25 whereas the white colour to vertices with degree centrality higher than 25. The analogous procedure was applied in the case of the degree prestige. After that the process of detecting triads within the WUT email social network was performed. Basically, there are 13 different motifs that consist of three nodes each (Fig. 1). Their ID=1,2,…,13 are used in the further descriptions interchangeably with the corresponding abbreviations M1, M2,…, M13. The main part of the experiments took into consideration motifs with two coloured vertices, in consequence for each of the 13 motifs four classes of motifs can be enumerated i.e. motifs with all vertices white, motifs with two vertices black and one white, with two vertices white and one black, and finally three black vertices. As an example see Fig. 2, where M5 is presented together with all possible motifs that have two-coloured vertices and have been built upon the M5. The outcomes of experiments, the goal of which was to detect the described two-colored motifs and determine the role of the nodes with high DC and DP in the social network, are presented in the Sec. 3.3.
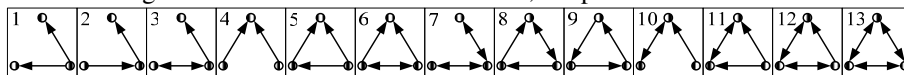

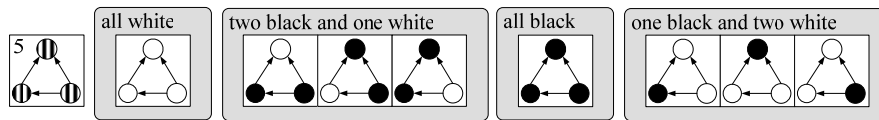Figure 1 Directed triads and their IDs that can exist within the social network.

Figure 2 Classes of the motifs

3.2 Distribution of Network Motifs in WUT Social Network

Triad Significance Profile (TSP) for WUT network presented in Fig. 3 was computed using the set of 1000 random networks. 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 and M10 occur less frequently in comparison to the random networks. As expected, it shows a high clustering level, i.e. high probability that two neighboring nodes have connected neighbors. The only exception is 1 which occurs relatively often. This reflects specific property of large mail–based social network: there are relatively many broadcasting nodes which spread messages (news, announcements, bulletins) which are never answered.
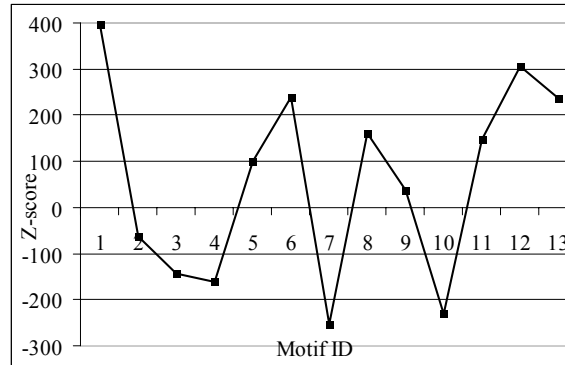

Figure 3 Triad Significance Profile (Z-score values) of the WUT email-based social network for 1000 numbers of random networks.

3.3 Analysis of the Coloured Network Motifs

The very interesting fact can be noticed while observing the motifs where all nodes are black, i.e. all users have low centrality. Their number is minimal (Fig. 4 and 5), although almost 70% of people have low degree centrality and 62% low degree prestige. Based on these results we may infer that there are no sparse structures and loosely connected cliques in the investigated network – nodes connect always with network hubs, direct communication between low-centrality nodes is present, but it does not shape the local structure of the network.

Table 2 Distribution of coloured network motifs for degree centrality

Motif ID	1	2	3	4	5	6	7	8	9	10	11	12	13
3 white	12,8	26,4	26,5	18,9	23,3	18,4	60,6	64,6	59,4	60,3	64,2	69,5	79,0
2 white 1 black	44,6	57,5	57,9	56,1	62,8	73,3	33,0	29,7	31,9	33,0	29,6	26,6	17,9
1 white 2 black	42,4	15,4	15,3	22,3	13,4	7,7	6,1	5,3	7,9	6,4	6,0	3,6	2,9
3 black	0,2	0,7	0,3	2,6	0,5	0,6	0,2	0,5	0,8	0,4	0,2	0,3	0,3

Interesting conclusions come from comparison of topological neighbourhood of the hubs (nodes with high DP and DC). When we compare the motif distribution for DC and DP (Fig. 4 and 5) it can be noticed that in case of DC the mixed motifs, i.e. these with two nodes in one colour and one in another colour, occur more often, especially in the case of motifs containing two edges (i.e. with nodes belonging to different clusters). In this way high-DC nodes ("DC hubs") tend to link low-centrality nodes with the core network, which is generally formed by hubs with high DP – 37% of network nodes with high DP constitute (alone) up to 50% of all network triads the

effect of which is not seen in the case of DC (30% occurrence, forming up to 25% of all triads). This is an important conclusion suggesting that dense cliques in social network are bridged and created via nodes with high DP (more often than in the case of these with high DC), communicating with each other.
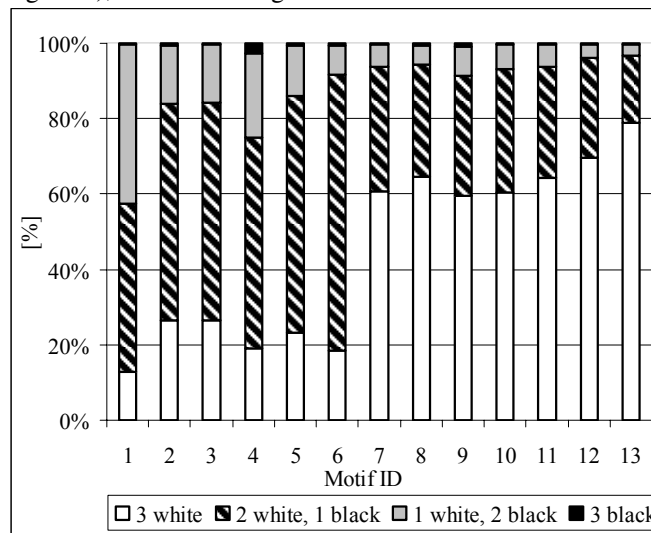

Figure 4 Distribution of coloured network motifs for degree centrality

The bridging effect of DP hubs is indicated by high percentage of white (3 white nodes) motifs. Even taking into account that the sets of nodes with high DC and DP partially overlap (there are obviously nodes which share high DC and DP) the prevailing role of DP hubs in local topology is clearly visible.
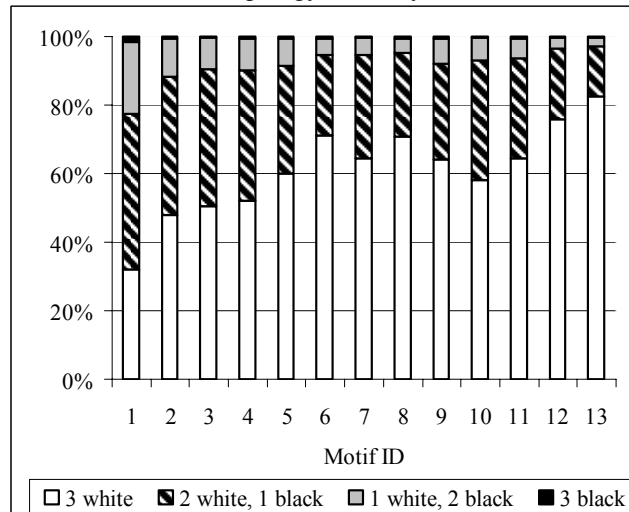

Figure 5 Distribution of coloured network motifs for degree prestige

Table 3 Distribution of coloured network motifs for degree prestige

Motif ID	1	2	3	4	5	6	7	8	9	10	11	12	13
3 white	31,9	48,0	50,4	52,0	60,0	71,0	64,3	70,8	64,2	58,1	64,5	76,0	82,6
2 white 1 black	45,5	40,3	40,0	38,0	31,5	23,8	30,3	24,4	27,9	34,8	29,1	20,4	14,5
1 white 2 black	21,0	10,9	9,1	9,3	7,8	4,6	5,2	4,2	7,3	6,6	5,8	3,2	2,5
3 black	1,6	0,8	0,4	0,7	0,7	0,7	0,3	0,5	0,7	0,5	0,6	0,4	0,4

In the case of DP, for almost all (excluding M1) motifs at least half of them consist of only white nodes, i.e. of people who have the prestige higher than the mean value for the network. Note also that, the more connected the motif is the larger number of motifs with three white nodes occur. This is true for both DP and DC (see Table 3 and 4). For example for both DP and DC three white nodes motifs constitute 80% occurrences of fully connected motif M13. It shows that the important people tend to create cliques within the social networks whereas people who are at the periphery of the network and are less important do not interact with each other.

4   Conclusions and Future Work

The motif analysis enables detection of the communication patterns within complex social networks. The approach presented in this paper extends existing motif analysis techniques by taking into account the properties (centrality) of network nodes. This is a continuation of results presented in [5], where the influence on the communication intensity on the local network topology was investigated. The obtained results revealed unknown properties of local topology of social networks, like differences between the role of DC and DP hubs, invisible for simple TSP analysis. They also form a basis for entirely new set of large dynamic social network analysis methods which build on several unique properties of motif analysis:

- Promising computational cost – even for huge networks TSPs can be obtained with fast sampling algorithms which offers possibility of frequent tracking of social network evolution processes.
- Possible integration with soft computing methods – colouring approach which attributes network nodes and edges with the values of arbitrary parameters coming from social network theory allows association of fuzzy variables (like "average DP") with local network structures/nodes. Future network mining methods may build on this.
- The possibility of developing algorithms for inferring global properties (like clustering coefficient) of the network from the TSP – first promising results were presented in [3] for simple Watts-Strogatz social network structures.

The next research steps will include motif analysis with respect to more sophisticated parameters of social network nodes (like social position and betweenes centrality). Also the dynamics of the network will be addressed through: a) the influence of local topology changes on the network; b) the consequences of attaching/deleting nodes and edges; and c) the analysis of changes in the WUT social

network during academic year – in order to check how known periodic changes in University's business profile and users' activity affect the local topology of its social network.

References

1. Alexander,C.N. (1963) A method for processing sociometric data, Sociometry, 26, 268-269.
2. Barabasi,A.-L., Albert,R. (1999) Emergence of scaling in random networks. Science, 286, 509–512.
3. Chung-Yuan H., Chuen-Tsai S., Chia-Ying C., Ji-Lung H., Bridge and brick motifs in complex networks, Physica A 377 (2007) 340–350.
4. Itzkovitz,S., Milo,R., Kashtan,N., Ziv,G., Alon,U. (2003) Subgraphs in random networks. Phys. Rev. E., 68, 026127.
5. Juszczyszyn K., Musiał K., Kazienko P. (2008), Local Topology of Social Network Based on Motif Analysis, 11th International Conference on Knowledge-Based Intelligent Information & Engineering Systems, KES 2008, Croatia, Springer, LNAI.
6. 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.
7. Mangan,S. Alon,U. (2003) Structure and function of the feedforward loop network motif. Proc. Natl Acad. Sci., USA, 100, 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. Mol. Biol., 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. 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.
11. Proctor,C.H, Loomis,C.P. (1951) Analysis of sociometric data, in: Research Methods in Social Relations, Jahoda,M., Deutch,M., Cok,S. (eds.), Dryden Press, NewYork, 561-586.
12. Shaw,M.E. (1954) Group structure and the behavior of individuals in small groups, Journal of Psychology 38, 139-149.
13. Shen-Orr,S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regualtion network of Escherichia coli. Nat. Genet., 31, 64–68.
14. Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J.-P., Oltvai, Z.N., Barabasi, A., 2004. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc. Natl Acad. Sci. USA 101, 17 940.
15. Wasserman,S., Faust,K. (1994) Social network analysis: Methods and applications, Cambridge University Press, New York.
16. Young-Ho E., Soojin L., Hawoong J., (2006) Exploring local structural organization of metabolic networks using subgraph patterns, Journal of Theoretical Biology 241, 823–829