The Dynamic Structural Patterns of Social Networks Based on Triad Transitions

Krzysztof Juszczyszyn[1], Marcin Budka[2], Katarzyna Musiał[2]
[1]*Institute of Computer Science, Wrocław University of Technology, Poland*
[2]*School of Design, Engineering and Computing, Bournemouth University, UK*
krzysztof@pwr.wroc.pl, mbudka@bournemouth.ac.uk, kmusial@bournemouth.ac.uk

***Abstract*— In modern social networks built from the data collected in various computer systems we observe constant changes corresponding to external events or the evolution of underlying organizations. In this work we present a new approach to the description and quantifying evolutionary patterns of social networks illustrated with the data from the Enron email dataset. We propose the discovery of local network connection patterns (in this case: triads of nodes), measuring their transitions during network evolution and present the preliminary results of this approach. We define the Triad Transition Matrix (TTM) containing the probabilities of transitions between triads, then we show how it can help to discover the dynamic patterns of network evolution. Also, we analyse the roles performed by different triads in the network evolution by the creation of triad transition graph built from the TTM, which allows us to characterize the tendencies of structural changes in the investigated network. The future applications of our approach are also proposed and discussed.**

Social network, network evolution, triad transitions

I. 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 measures, finding cliques, etc. require significant computing overhead. However, the technology-based social networks add a new dimension to the known problems of network analysis [11].

This family of networks (web communities, email social networks, user networks and so on) have two properties which have a significant impact on the analysis. First, the existence of link is a result of a series of discrete events (like emails, phone calls, blog entries) which have some distribution in time. As shown in [9] for various kind of human activities related to communication and information technologies, the probability of inter-event times (periods between the events, like sending an email) may be expressed as: $P(t) \approx t^{-\alpha}$ where typical values of α are from (1.5, 2.5). This distribution inevitably results with series of consecutive events ("activity bursts") divided by longer periods of inactivity.

These phenomena have serious consequences when we try to apply the classic structural network analysis (SNA) to the dynamic networks. The most popular approach is to divide the time period under consideration into time windows, then run SNA methods on the windows separately. This should show us how the measures like node centrality, average path length, group partitions etc. change in time, giving us an insight into the evolutionary patterns of the network.

However, the bursty behavior of the users (long inactivity periods mixed with the bursts) causes dramatic changes of any measure when switching from one time window to another. There is a trade-off: short windows lead to chaotic changes of network measures, while long windows give us no chance of investigation of network dynamics [13][14].

In order to address this problem, a number of methods, designed to predict changes in the structure of dynamic networks. were proposed [15][16]. The special case is so-called *link prediction* problem – the estimation of probability that a link will emerge/disappear during the next time window [12].

In this work we propose a method of characterizing the dynamic evolutionary patterns of the network by the analysis of changes in local topology of connections. This approach stems from our previous experience [20] and will be introduced in Sec. 2. Sec. 3 presents the results on the basis of the Enron e-mail network.

II. LOCAL TOPOLOGY OF ONLINE SOCIAL NETWORKS

A. Triads and network motifs

Standard approaches exploiting network analysis by means of listing several common properties, like the degree distribution, clustering, network diameter or average path lengths often fail when applied to complex networks. In many cases it is possible to construct networks with exactly the same (for example) degree distribution whose structure and function differ substantially. Huge network structures (like social, biological, gene networks) should be investigated with more precise and structure–sensitive methods [1]. During last years we experienced the development of a number of methods investigating complex networks by means of their local structure (especially – frequent patterns of connections between nodes). The simplest, and therefore popular, way to characterize the network in the context of local connections is to examine the links between the smallest non-trivial subgraphs, the triads, consisting of three nodes. If we additionally decide to distinguish between the nodes (which is our case, for in our network they are corporate email addresses) we get 64 patterns of possible connections between any three identifiable nodes (Fig.1).

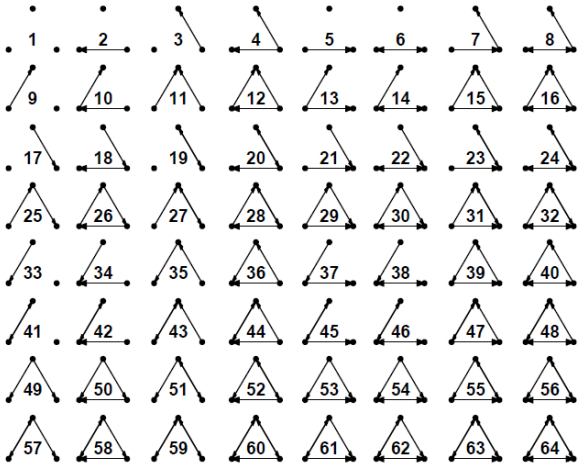

Fig.1. Three-node triads in a directed graph

Please note the triad ID (the number inside the picture of the subgraph) in Fig. 1, as it will be used further on in this paper. Note that there is a correspondence between the IDs and the edit distance between triads – small difference in the ID value *in most cases* suggests small edit distance (the number of link removal/addition operations needed to transform one triad into another).

The basic method utilizing such subgraphs is the well-knows triad census, which is enumeration of all triads in the network and allows to reason about the functional connection patterns of the nodes [18].

Last years have seen the development of more sophisticated approaches, among them *motif analysis* which aims to characterize the network by the difference between its structures and an ensemble of random networks of the same size and degree distribution. A biased distribution of local network structures (subgraphs), a.k.a. *network motifs* is widely observed in complex biological or technology–based networks. Motif analysis stems from bioinformatics and theoretical biology [1][3], where it was applied to the investigation of huge network structures like transcriptional regulatory networks, gene networks or food webs [4][5]. 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 [6][7][8].

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, 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.

In our former research we have investigated the local structure of numerous technology-based networks, among them an e-mail social network of Wroclaw University of Technology (WUT), consisting of more than 5 800 nodes and 140 000 links [2][17].

Our aim was to check if the known properties of local topology in social networks (known on the basis of motif analysis conducted for small social networks [4]) are also present in large email–based social structures, and if there are some distinct features characteristic to the email communication. The most important conclusion from these experiments was that the general motif profile of the network (expressed by so-called *triad significance profile – TSP* – a vector of the Z-score measures of the motifs) is stable over long periods of time. This was confirmed even for periods like summer holidays when the number of links in the university network dropped by 50% [17]. Summing up – the investigated complex network show statistically stable pattern of connections as a whole, despite the fact that stability of a single link is quite low: 59% in our case (which means that 41% of the connections will not be present in the next time window).

These observations led to the idea of characterizing the evolutionary patterns of the network by means of the changes in elementary subgraphs, in this particular case – directed triads.

In the next section we introduce the Triad Transition Matrix (TTM) as a basic structure used in our experiments.

2.2. Triad Transition Matrix

The idea behind the Triad Transition Matrix is to use the data about the history of the network (recorded during past time windows) to derive the probabilities of transitions between triads (patterns of local connections).

The TTM is a matrix of size g_Δ x g_Δ, where g_Δ is the number of considered subgraphs. For directed triads in our experiments $g_\Delta = 64$ (see Fig.1).

The values of TTM entries are defined as follows:

$$TTM_t(i,j) = P(g_i[t] \rightarrow g_j[t+1]) \qquad (1)$$

$TTM_t(i,j)$ is the probability (estimated on the basis of full subgraph enumeration for networks created from data gathered in $[t]$ and $[t+1]$ time windows), that a connection pattern g_i detected during $[t]$ will transit into g_j during $[t+1]$.

The goal was to check if the stability of local network structures (discussed in the former subsection) is followed by the distinguishable evolutionary patterns.

III. Experiments on the Enron email network

For the experiments with the TTM we have chosen the Enron dataset (http://www.cs.cmu.edu/~enron/), one of the popular reference e-mail logs.

3.1. The temporal networks of Enron dataset

First, the data cleansing process was performed (external addresses were removed from the database in order to analyze only the corporate social network). Additionally, only emails from and to the Enron domain were left (we may call the resulting set of nodes and the links between them a corporate social network).

The time period of our experiment was divided into 12 time windows and for each of them a network was created.

The main nodeset in our experiment consists of 150 nodes and up to 1012 links (in a single time window). Fig.2 shows an example of the network derived from Enron email communication for time window no.10.

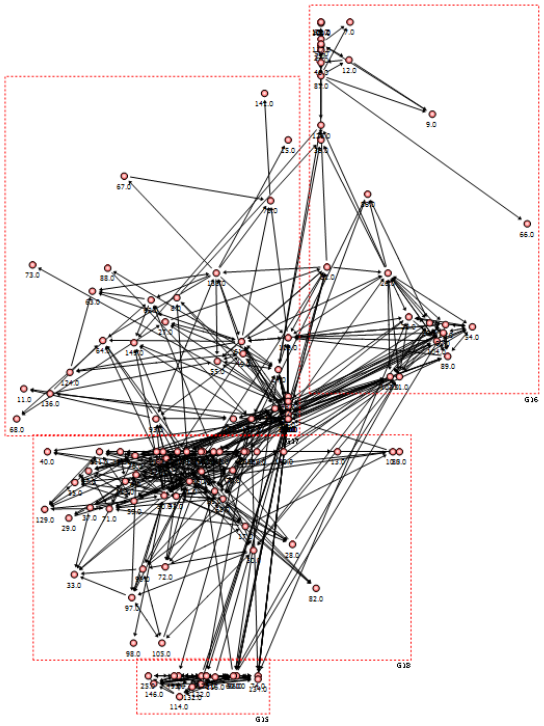

Fig.2. E-mail network (with node groups) generated for time window no.10.

It is also visible, that the values on the diagonal are usually bigger than the rest of the respective matrix row, which may be interpreted as stability of the already-established links.

Table 1. Network size for consecutive time windows

Time:	1	2	3	4	5	6	7	8	9	10	11	12
Edges:	189	207	281	312	334	398	430	462	626	1012	921	480

From Table.1 we see that – despite the equal number of nodes in each time window – the number of edges differs significantly. It is obvious that in terms of the number of links, node centrality, etc. the structure of the network is changing. However, there is some pattern behind this change.

3.2. TTM – the results

In Fig. 4 the TTM derived on the basis of 12 time windows is presented. Despite the changes in network size, all TTMs computed for neighbouring time windows showed similar values, very similar to those in Fig. 3, which contains the mean values of transition probabilities.

We may notice that the distribution of transition probabilities is not flat, and there are distinctive

patterns (the coordinates of TTM correspond to the triad numbers from the Fig.1).

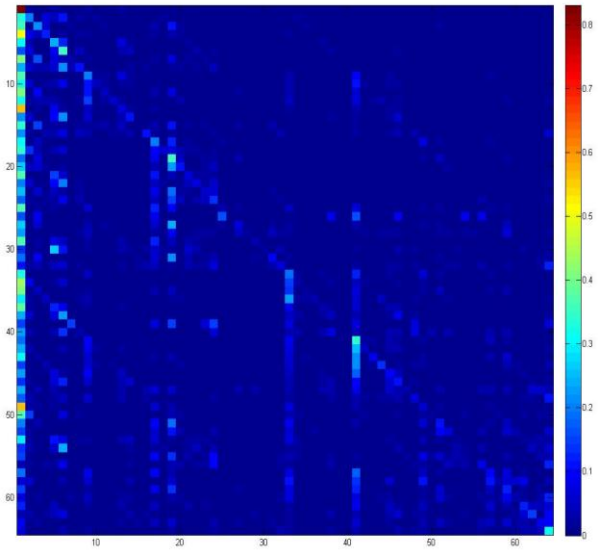

Fig.4. TTM containing the transition probabilities averaged for all 12 time windows.

First of all, the value of TTM(1,1) reflects the fact that the network is sparse (link density below 1%) which means that most of the possible triads contain no edges (in fact this value does not change between time windows). As the result most of the "empty" triads always remain in this state, which gives us a relatively high value of TTM(1,1).

We should also note the high values in the first column of the TTM. This means that when it comes to disappearing of the links, the probability of resetting the entire triad to zero-connection state is relatively high.

From the other hand, it is also visible, that the values on the diagonal of TTM are bigger than most values in their neighborhood, which shows that the already-formed triads tend (in general) to stay in their current state.

The last important observation is that some triads are special, they show clearly bigger values in their columns of TTM, which means that they are "sinks" of the evolution patterns of connections.

All the above observations will be used in the future research on the TTM with an aim to propose a novel approach to link prediction and, consequently disappearing.

3.3. TTM - analysis

On the basis of the TTM discussed in the last section we made an attempt to characterize the roles and behavior of each triad. In order to do this we have proposed an original approach utilizing the classical structural network analysis.

The first step was to treat the TTM as an adjacency matrix, and the transition probabilities as weighted, directed relations between triads. Thus, we got a structure which may be called a Triad Transition Graph (TTG). From Fig.4 we may guess that the TTG was relatively dense. Indeed, there are 2538 (out of 4096) non-zero values in the TTM.

In the second step we have checked the in- and out-degrees of the network nodes (measured as the sum of weights of the incoming/outgoing links). The results are presented in Table 2. In-degrees correspond to the sums of TTM's column values, while the out-degrees – to the sums of rows, with the diagonal values excluded in both cases.

Table 2. Degrees and roles of the triads.

Triad	In-Deg.	Out-Deg.	Node Type (links>0.1)	Triad	In-Deg.	Out-Deg.	Node Type (links>0.1)
1	15.756	0.170	Receiver	**33**	2.322	0.812	Ordinary
2	1.499	0.794	Ordinary	**34**	0.292	0.972	Transmitter
3	1.497	0.789	Ordinary	**35**	0.159	0.974	Transmitter
4	0.552	0.876	Transmitter	**36**	0.083	0.968	Transmitter
5	2.163	0.759	Ordinary	**37**	0.237	0.969	Transmitter
6	2.960	0.640	Ordinary	**38**	0.515	0.949	Transmitter
7	0.361	0.966	Ordinary	**39**	0.008	1.000	Transmitter
8	0.497	0.891	Transmitter	**40**	0.033	1.000	Transmitter
9	1.801	0.799	Ordinary	**41**	3.194	0.659	Ordinary
10	0.172	0.954	Transmitter	**42**	0.256	0.959	Transmitter
11	0.228	0.958	Transmitter	**43**	0.202	0.922	Transmitter
12	0.103	0.940	Transmitter	**44**	0.213	0.868	Transmitter
13	0.827	0.925	Transmitter	**45**	0.738	0.900	Transmitter
14	0.552	0.890	Transmitter	**46**	0.933	0.890	Transmitter
15	0.122	1.000	Transmitter	**47**	0.067	0.974	Transmitter
16	0.230	0.887	Transmitter	**48**	0.441	0.923	Transmitter
17	2.176	0.792	Ordinary	**49**	0.925	0.936	Carrier
18	0.195	0.943	Transmitter	**50**	0.124	0.983	Transmitter
19	3.228	0.649	Ordinary	**51**	0.743	0.887	Ordinary
20	0.476	0.890	Transmitter	**52**	0.344	0.928	Transmitter
21	0.537	0.931	Transmitter	**53**	0.217	0.980	Transmitter
22	0.443	0.927	Transmitter	**54**	0.341	0.979	Transmitter
23	0.488	0.935	Transmitter	**55**	0.099	1.000	Transmitter
24	1.073	0.859	Ordinary	**56**	0.408	0.919	Transmitter
25	0.301	0.964	Ordinary	**57**	0.710	0.896	Ordinary
26	0.003	1.000	Transmitter	**58**	0.076	0.974	Transmitter
27	0.318	0.933	Transmitter	**59**	1.020	0.871	Transmitter
28	0.044	0.974	Transmitter	**60**	0.264	0.962	Transmitter
29	0.166	0.964	Transmitter	**61**	0.480	0.874	Transmitter
30	0.076	0.981	Transmitter	**62**	0.354	0.852	Isolate
31	0.264	0.906	Transmitter	**63**	0.475	0.922	Isolate
32	0.321	0.926	Transmitter	**64**	1.490	0.678	Receiver

The in-degree may be interpreted as "attraction strength" of the triad, in the case of a high value, the other triads will evolve into it more frequently. Out degree corresponds to individual triad "instability", the values close to 1 suggest that it is improbable that this

triad will remain unchanged during the next time window.

The in- and out-degrees are the simple structural properties of network nodes – in our approach we have also checked the structural properties of the TTG.

Due to the density of TTG, in our first experiment we have assumed a cut-off relation strength equal to 0.1 (i.e. in further steps we treated any transition probability less than 0.1 as equal to 0).

For such a reduced TTG, we were able to assess the roles of the triads, also presented in Table 2. We used a standard set of simple node network roles defined in [18]: *isolate* nodes do not have any links. *Transmitter* has only out links and no in links. *Receiver* node has only in links, while *carrier* node has exactly one incoming and one outgoing link. *Ordinary* node does not fall in any of the above categories. The roles allow to classify the triads in the context of dynamic structural patterns of the investigated network.

The last step of our analysis was to deal with the network of transition probabilities (relations between triads). The network is shown on the Fig. 5 (black links have weight close to 1, while fading color corresponds to decreasing link weight value).

From Fig.5 we may see that triad 0 is clearly a network hub, which was already suggested in Table 2.

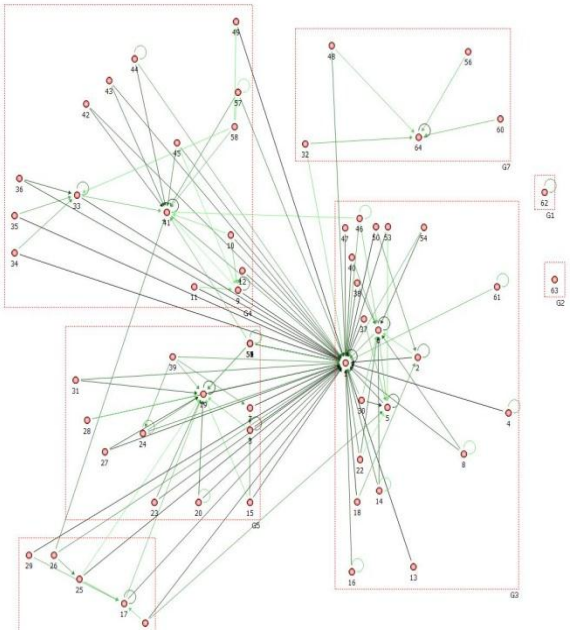

Fig.5. The reduced (only links > 0.1) Triad Transition Graph.

There are also groups of triads (detected with eigenvector community algorithm) in which the inter-group connections are denser. During further research on our method we will use the groups in the TTG to define *transition classes* composed of the triads which tend to evolve inside limited sets (this phenomenon corresponds to the existence of structural groups in TTG).

Summing up, we have shown that in dynamic complex networks there are distinctive patterns which drive the evolution of connections between nodes. Our approach allows to build evolutionary patterns of complex networks, which may form the basis for link and structure (group stability, network connectivity) prediction. In fact it should also allow for classification of evolutionary patterns of complex networks, because there may be different TTMs in the case of networks of different nature and origin. The method will undergo further development in directions briefly listed in the following section.

4. Conclusions and future work

The concept of TTM will be further researched in the following directions:

- Software development in order to deal with large networks.
- The application of TTM concept to the link prediction problem.
- Developing the methods for the analysis and prediction of *triad trajectories* (sequences of triads evolving one into the other). If effective, this approach may result in developing long time prediction methods for dynamic networks.
- Reducing the complexity of the method by using effective algorithms for triad enumeration, for example applying the approach presented in [19], which should allow to analyze large networks.
- The adoption and checking the effectiveness of network sampling algorithms used in motif discovery for the speedup of TTM building process.

Additionally the behavior of the TTMs will be checked for a number of test social networks created from the data gathered in various information systems (social portals, mail servers, blogosphere) in order to check for their distinctive features and tune the method.

5. Acknowledgements

This work was supported by the Polish Ministry of Science and Higher Education, grant no. N N516 518339.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 251617.

REFERENCES

[1] Itzkovitz S., Milo R., Kashtan N., Ziv G., Alon U. (2003) Subgraphs in random networks. Physical Review E., 68, 026127.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] A.-L. Barabási, The origin of bursts and heavy tails in humans dynamics, Nature 435, 207 (2005).

[10] T. Gross, H. Sayama (Eds.): Adaptive networks: Theory, models and applications, Springer: Complexity, Springer-Verlag, Berlin-Heidelberg, 2009.

[11] J. Kleinberg, J. The convergence of social and technological networks. Communications of the ACM Vol. 51, No.11, 66-72, 2008.

[12] D. Lieben-Nowell, J.M. Kleinberg: The link-prediction problem for social networks. JASIST (JASIS) 58(7), pp.1019-1031, 2007.

[13] D.Braha, Y. Bar-Yam, From Centrality to Temporary Fame: Dynamic Centrality in Complex Networks, Complexity, Vol. 12 (2), pp. 59-63, 2006.

[14] D. Kempe, J. Kleinberg, A. Kumar, Connectivity and inference problems for temporal networks. Journal of Computational System Science, 64(4):820–842, 2002.

[15] M. Lahiri, Tanya Y. Berger-Wolf: Mining Periodic Behavior in Dynamic Social Networks. ICDM pp.373-382, 2008.

[16] Lisa Singh, Lise Getoor: Increasing the Predictive Power of Affiliation Networks. IEEE Data Eng. Bull. (DEBU) Vol. 30 No. 2, pp. 41-50, 2007.

[17] K. Juszczyszyn, K. Musial, P. Kazienko, B. Gabrys: Temporal Changes in Local Topology of an Email-Based Social Network. Computing and Informatics 28(6): 763-779 (2009).

[18] S. Wasserman, K. Faust, Social network analysis: Methods and applications, Cambridge University Press, New York, 1994.

[19] Batagelj, V., Mrvar, A., A subquadratic triad census algorithm for large sparse networks with small maximum degree. Social Netw. 23, 237-243, 2001.

[20] K.Juszczyszyn, K.Musial, P.Kazienko, B.Gabrys: Temporal Changes in Local Topology of an Email-Based Social Network. Computing and Informatics 28(6): 763-779 (2009).