A Probabilistic Approach to Structural Change Prediction in Evolving Social Networks

Krzysztof Juszczyszyn, Adam Gonczarek, Jakub M. Tomczak
Institute of Computer Science
Wrocław University of Technology
Wrocław, Poland
{*krzysztof.juszczyszyn, adam.gonczarek, jakub.tomczak*}*@pwr.wroc.pl*

Katarzyna Musial
School of Mathematical and Natural Sciences
King's College London
London, UK
katarzyna.musial@kcl.ac.uk

Marcin Budka
School of Design, Engineering and Computing
Bournemouth University
Bournemouth, UK
mbudka@bournemouth.ac.uk

***Abstract*—We propose a predictive model of structural changes in elementary subgraphs of social network based on Mixture of Markov Chains. The model is trained and verified on a dataset from a large corporate social network analyzed in short, one day-long time windows, and reveals distinctive patterns of evolution of connections on the level of local network topology. We argue that the network investigated in such short timescales is highly dynamic and therefore immune to classic methods of link prediction and structural analysis, and show that in the case of complex networks, the dynamic subgraph mining may lead to better prediction accuracy. The experiments were carried out on the logs from the Wroclaw University of Technology mail server.**

***Keywords*-social networks; mixture of Markov chains; prediction;**

I. Introduction

Network analysis has experienced a rapid development of new methods and algorithms. Our capabilities of gathering and processing data from networked systems lead to many challenges of analysis and change prediction in fast evolving network structures. Classical approaches, based on structural graph theory and using structural measures for characterization of network components, groups and entire networks often fail or, at least, make network analysis error-prone and difficult. When investigating the topological properties and structure of complex networks we face a number of complexity-related problems. In large social networks, tasks like evaluating the centrality measurements or finding cliques require significant computing resources. In this context, methods which proved to be useful for medium and small networks often fail when applied to larger structures. 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). A biased distribution of local network structures is widely observed in complex biological or technology-based networks.

In this work we propose the application of Markov chains to the prediction of local topology changes of dynamic, time-dependent and therefore immune to standard methods of structural analysis e-mail social network. We also show the effectiveness of this approach for analysis of changing social networks in very short time-scales – in our case the network was analyzed in consecutive one-day time windows.

The paper is structured as follows: the next section briefly presents the most important results od structural analysis of dynamic networks, Section III discusses the experimental setup and the properties of the network under investigation, Section IV defines the Markov chain model, and last two sections offer results analysis and outline the most appealing directions of future research.

II. Related Work

In large social networks, evaluating the centrality measures, finding cliques, etc. require significant computing resources. However, the technology-based social networks (like the one used in our experiments) add a new dimension to the known problems of network analysis [15]. In this family of complex networks the existence of a link is a result of a series of discrete events (like email exchanges, phone calls, posting of blog entries) which have some distribution in time. As shown in [2] for various kinds 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 between 1.5 and 2.5. The result of such a distribution are detectable series of consecutive events ('bursts') divided by longer periods of inactivity. These phenomena have serious consequences when coupled with structural network analysis. The standard approach to dynamic complex network is to divide the available time frame into windows to compute the chosen structural network properties for networks created on the basis of data from these windows [5]. This should show how the measures like node centrality, average path length, group partitions etc. change over time. However, the bursty behavior of the users causes dramatic changes of any measure when switching from one time window to another [4]. There is an inevitable trade-off: short windows lead to chaotic and noisy dynamics of network measures, while long windows give us no chance to investigate time evolution of the network [14].

This opens a new research area, which encompasses a number of approaches designed to predict changes in the structure of dynamic networks [16]. The special case of this family of methods is a so-called *link prediction problem* – the estimation of probability that a certain link will emerge/disappear during the next time window [17]. A good survey of link prediction methods is presented in [7]. It should be noted that most methods of the link prediction give rather poor results – the best predictors discussed in [17] can identify $< 10\%$ of emerging links. For big networks, the number of disconnected pairs of nodes increases quadratically (the density of real-world networks is small and the graphs are sparse) while the number of links grows only linearly [8]. There are also link prediction methods which utilize information external from the graph network model itself (like in [1], where the content of Web pages forming the network was used in the prediction).

In this work we propose a method designed for the prediction of elementary network subgraphs – *triads*. Our motivation is that the topology of complex networks is a result of local interactions between the network components [6] and modeling of interactions on this level is a key step to more advanced methods of predictive structural analysis.

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. The basic method utilizing such subgraphs is the well-known triad census, allowing to reason about the functional connection patterns of the nodes [23].

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 [9], [13], where it was applied to the investigation of huge network structures like transcriptional regulatory networks, gene networks or food webs [20], [18]. 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 [19], [22], [24]. The discovered motifs and their numbers enable also to assess which patterns of communication appear often in the large social networks and which are rather rare.

However, in this work, we do not detect biased triad occurrences but propose a method for the prediction of changes in connection patterns in node triads.

III. Dynamic Social Networks - Dataset and Experimental Setup

The experiments were carried out on the logs from the Wroclaw University of Technology (WUT, http://www.portal.pwr.wroc.pl/) mail server, which were pruned to contain only the emails originated from (or: sent to) the staff members registered at the mail server of the university. There are 5834 active email addresses on the server, which implies that even for the shortest time window of 1 day, there were on average approx. 2000 active network nodes. For our experiments we used data from a period of 50 days, starting on the 4^{th} of March 2010. In our former research we have investigated the local structure of numerous technology-based networks, among them the evolution of an e-mail social network of the WUT during the period of two years [11], [12]. We have found that, despite significant changes in networks structure the statistical distribution of the subgraphs remains stable, which led to the idea of characterizing network dynamics by the evolutionary patterns of the subgraphs [11].

It should be noted that the email social networks undergo rapid structural changes when investigated in short time periods. Fig. 1 shows the changes in the number of the links which connect 4560 users active during the timespan assumed for our experiment (100 days out of which first 50 were used to train Markov model, the rest to verify it).

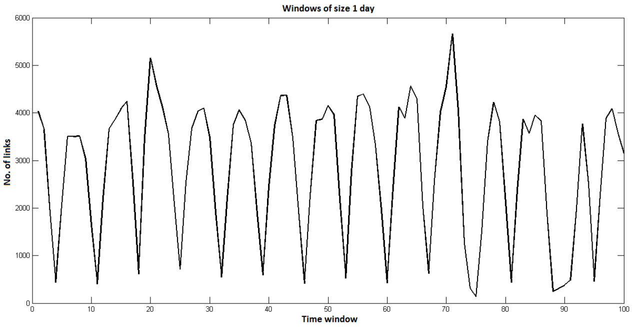

Figure 1. Number of links in the WUT social network.

As we can notice, the visible decrease in the number of links observed every seven days clearly corresponds with Sundays (or, in general, the weekends). One can even recognize the annual student celebration in May which resulted in two free days around the 90^{th} day of the dataset. However such a short timescale (1-day time widows) results in huge variation in all classic structural characteristics of the network (node degree, clustering, betweenness etc.). From our point of view it was interesting that it also affects all known methods of link prediction. For example, we have checked the effectiveness of two methods presented in a classic survey work [4] and got the average accuracy of common neighbors and preferential attachment predictors of 0.9% and 0.06% respectively (these results exactly correspond to the effectiveness of these predictors from [4] where they turned to be approx. 40 times better than random predictor).

Basing on these observations (and results from the works cited in the previous section) we suggest that the accurate predictions for fast-changing social networks observed in short periods of time require the analysis of dependencies and correlations of the activity of the nodes which may be described in terms of temporal patterns of changes in local network topology. In our research we analyze them from the level of the simplest of these patterns – the connections between triples of nodes. There are 64 different connection patterns in a directed network of labeled nodes (Fig. 2).

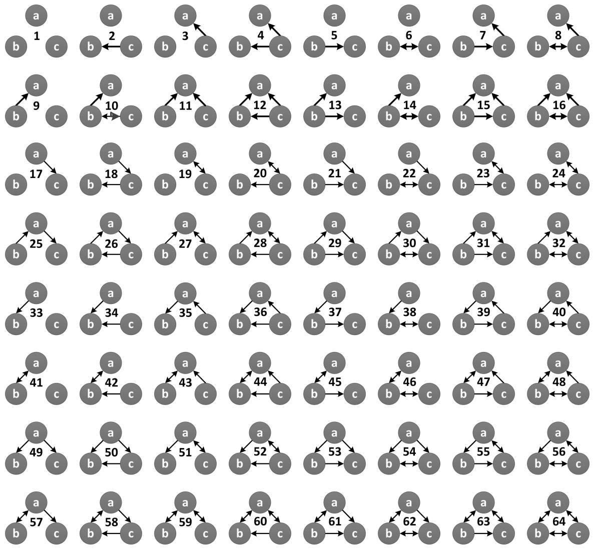

Figure 2. Three-node triads in a directed, labeled graph.

The introductory analysis consisted of the following steps:

1) Creation of social networks from the email communication data. Each network corresponded to the server logs from a single day, and 50 networks were created in total.
2) Determining the connection patterns of any triad of nodes out of which at least two were connected by at least one directed link in any of the 50 networks.
3) The above patterns determine so-called triad trajectories – sequences of 50 numbers from 1 to 64. The i^{th} element of each triad trajectory corresponds to the connections recorded in respective time window between the three nodes being considered (Figure 2). Each trajectory may be interpreted as a sequence of connection patterns emerging between the triples of nodes in the respective time windows.

The result was a test set of $1,280,363$ triad trajectories, out of which $896,255$ (70%) were used for the training of Markov model (presented in the following section) and the remaining 30% served as a test set for evaluating predictions. The average number of non-empty (containing at least one link) triads in one network was $167,170$.

IV. Mixture of Markov Chains

In this section a probability distribution over sequences of triads is presented. A model based on mixture of Markov chains is introduced. Such approach allows to group triad trajectories into clusters of different behaviour types. Hence, let denote the triad at the m^{th} moment in the n^{th} observation as $\mathbf{x}_{nm}$ and encode it as a zero-one sequence of length L, e.g., $\mathbf{x}_{nm} = (0, 1, 0, \ldots, 0)$ if the second triad has appeared. The observed data $\mathbf{X}$ consists of N sequences of triads $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_N]$, where $\mathbf{X}_n = [\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nM}]$. Moreover, let denote a latent random zero-one vector of length K as $\mathbf{z}_n$, e.g., $\mathbf{z}_{n2} = (0, 1, 0, \ldots, 0)$ if the second cluster has occurred. Then the mixture of Markov chains can be represented as a probabilistic graphical model (see Fig. 3; a node with double circles denotes an observable variable, and a node with one – a latent variable). The latent variable $\mathbf{z}$ could be seen as a group of behaviour types of a triad.

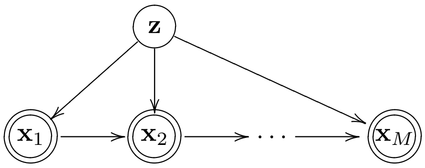

Figure 3. Graphical representation of a mixture of Markov chains.

Furthermore, a multinomial prior distribution over variable $\mathbf{z}_n$ is chosen:

$$p(\mathbf{z}_n) = \prod_{k=1}^{K} \theta_k^{z_{nk}}. \tag{1}$$

Hence, each triad sequence is assumed to be a sample from a first-order Markov chain with a multionomial prior distribution π_k over a first state and a transition matrix $\mathbf{A}_k = [A_{kij}]$, $i, j = 1, 2, \ldots, L$:

$$p(\mathbf{x}_{n1}|\mathbf{z}_n) = \prod_{k=1}^{K}\prod_{l=1}^{L} \pi_{kl}^{x_{n1l}\cdot z_{nk}}, \tag{2}$$

$$p(\mathbf{x}_{n(m+1)}|\mathbf{x}_{nm}, \mathbf{z}_n) = \prod_{k=1}^{K}\prod_{i=1}^{L}\prod_{j=1}^{L} A_{kij}^{x_{n(m+1)j}\cdot x_{nmi}\cdot z_{nk}}. \tag{3}$$

For further simplicity the following notation is introduced:

$$Markov(\mathbf{X}_n|\pi_k, \mathbf{A}_k) = p(\mathbf{x}_{n1}|\mathbf{z}_n)\times \prod_{m=1}^{M-1} p(\mathbf{x}_{n(m+1)}|\mathbf{x}_{nm}, \mathbf{z}_n). \tag{4}$$

A. Learning stage

In order to start the inference procedure, the model should be first trained on data. The goal of the learning algorithm is to obtain parameters π, $\mathbf{A}$, θ that could be accomplish by maximizing the following likelihood:

$$p(\mathbf{X}|\pi, \mathbf{A}, \theta) = \prod_{n=1}^{N}\sum_{k=1}^{K} \theta_k Markov(\mathbf{X}_n|\pi_k, \mathbf{A}_k). \tag{5}$$

However, the likelihood is a mixture distribution and an analytical solution is intractable to be obtained. Therefore, an expectation-maximization procedure [3] is applied.

E-step. *The posterior distribution over the sequence of latent variables is expressed in the form:*

$$p(\mathbf{z}|\mathbf{X}, \pi, \mathbf{A}, \theta) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \{\theta_k Markov(\mathbf{X}_n|\pi_k, \mathbf{A}_k)\}^{z_{nk}}. \tag{6}$$

Thus the following expectations can be obtained due to the fact that $z_{nk} \in \{0, 1\}$:

$$\gamma(z_{nk}) = \mathbb{E}\ [z_{nk}] = \frac{\theta_k Markov(\mathbf{X}_n|\pi_k, \mathbf{A}_k)}{\sum_{h=1}^{K} \theta_h Markov(\mathbf{X}_n|\pi_h, \mathbf{A}_h)} \tag{7}$$

M-step. *Then new values of the parameters can be expressed in a closed form as a result of maximization procedure of the expected value of the joint log-likelihood with respect to the distribution (6). Because of the constraints on the parameters* $\pi, \mathbf{A}, \theta$ *the Lagrange multipliers have to be used in order to obtain the following solution:*

$$\theta_k^{new} = \frac{1}{N}\sum_{n=1}^{N} \gamma(z_{nk}), \tag{8}$$

$$\pi_{kl}^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{n1l}}{\sum_{n=1}^{N} \gamma(z_{nk})}, \tag{9}$$

$$A_{kij}^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \sum_{m=1}^{M-1} x_{nmi} x_{n(m+1)j}}{\sum_{n=1}^{N} \gamma(z_{nk}) \sum_{m=1}^{M-1} x_{nmi}}. \tag{10}$$

STOP: *The EM procedure stops iterating when the change in the likelihood (5) in two consecutive steps is less then given threshold* ε.

B. Prediction stage

The goal of the inference is to predict the next triad $\mathbf{x}_{p+1}$, given the triad sequence $\mathbf{X}_{1:s} = [\mathbf{x}_1, \ldots, \mathbf{x}_s]$. It could be done using following conditional probability:

$$p(\mathbf{x}_{s+1}|\mathbf{X}_{1:s}) = \frac{\sum_{k=1}^{K} \theta_k Markov(\mathbf{X}_{1:s+1}|\pi_k, \mathbf{A}_k)}{\sum_{k=1}^{K} \theta_k Markov(\mathbf{X}_{1:s}|\pi_k, \mathbf{A}_k)}. \tag{11}$$

Then the triad $x_{s+1,m}$ with the highest probability is taken as a predicted value. The extension of this procedure to predict sequences of the triads $\mathbf{x}_s, \ldots, \mathbf{x}_{s+r}$ is straightforward and can be obtained by using the dynamic programming procedure in order to find the most probable sequences.

V. Experimental Results

A. Details

The parameters of the probabilistic model considered in the previous section were calculated due to the EM procedure based on the $896, 255$ triad trajectories and each trajectory consists of 50 moments. A single triad trajectory concerns a single triad that links could disappear and appear in time. An example of the triad trajectory is presented in the Fig. 4 where in the first moment the triad is of the 44^{th} type (see Fig. 2), then in the second moment it evolves to the 42^{nd} type, and next to the 2^{nd} type, and so on. Notice that we are able to observe the triad at each moment and assuming the dependency between two triads at two following moments we get the first order Markov chain.

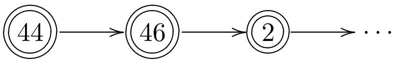

Figure 4. An exemplary triad trajectory.

After the learning stage the model was evaluated based on the $384, 108$ triad trajectories. The one-step-ahead prediction was made for all moments starting from the 3^{rd} to the 50^{th} moment. In the Table I the results concerning the mean value of error for all observations and 50 moments (referred as *Mean*), as well as the standard deviation (*Std. Dev*), the worst and the best case among all moments (*Worst case* and *Best case*, respectively) for mixture of Markov chains with different values of $K = 1, 3, 5$ are presented (results for K = 2, 4 being quite similar). Additionally, below the double line the results for random method (prediction was made due to the random uniform distribution; referred as *Random*), and a method that always returns 1 (the triad that occurs most often in the dataset; referred as *Constant*) are given. Both methods are used as a comparison and a reference. In the experiment no context information was used, e.g., if considered moment is a working day or a weekend.

Table I
RESULTS FOR MIXTURE OF MARKOV CHAINS.

K	*Mean*	*Std. Dev.*	*Worst case*	*Best case*
1	0.079	0.126	0.662	0.016
2	0.081	0.122	0.661	0.015
3	0.080	0.120	0.662	0.018
4	0.079	0.126	0.661	0.019
5	0.082	0.117	0.662	0.017
Random	0.984	0	0.945	0.984
Constant	0.085	0.129	0.669	0.002

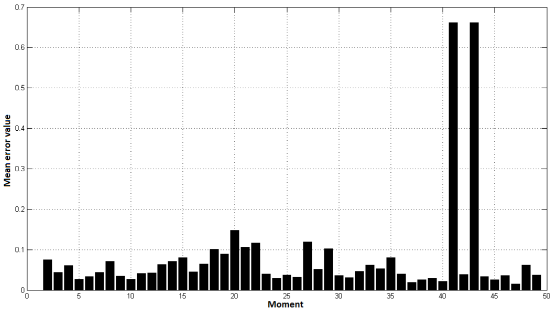

Figure 5. *Mean* error for each moment and $K = 1$.

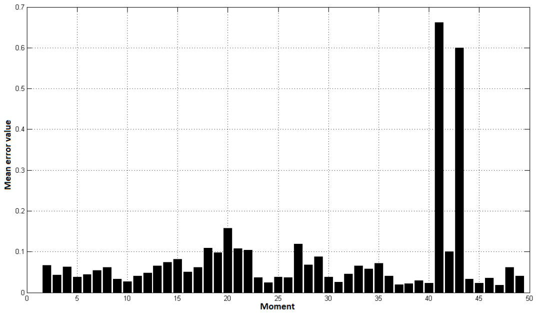

Figure 6. *Mean* error for each moment and $K = 3$.

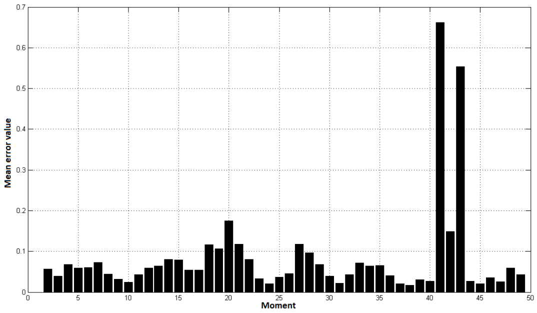

Figure 7. *Mean* error for each moment and $K = 5$.

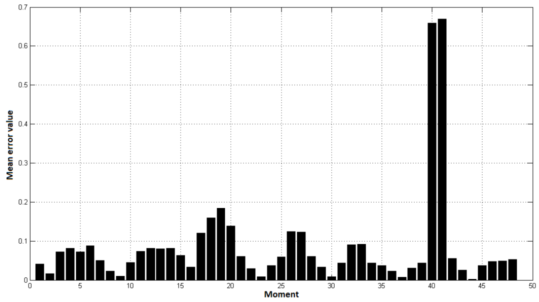

Figure 8. *Mean* error for each moment and the *Constant* method.

B. Discussion

Obtained results indicate that application of the mixture of Markov chains gives very promising outcome (the mean error at the level of approx. 8%). Nevertheless, the *Constant* method that always returns the triad number 1 performed also quite well. However, it could be easily explained because in the dataset only around $170,000$ of triads were other type than No. 1. However, when comparing graphs for the mixture of Markov chains (Fig. 5 - 7) with the one for the *Constant* method (Fig. 8) it could be noticed that during weekends the dynamics of the network is very low and almost all triads have no links. Hence, the *Constant* method made almost no error in prediction. On the other hand, the *Constant* method got worst results during working days than the mixture of Markov chains. Therefore, it could be stated that the proposed probabilistic model performed very good during working days but a little worse during weekends.

Moreover, considering the K value it could be said that the best results were obtained for one group ($K = 1$, see Table I) and four groups ($K = 4$, see Table I). It is an indication that in the considered dataset there could be four groups of trajectories (behaviour types). Nevertheless, such analysis needs further and more detailed research.

Furthermore, quite peculiar is a very bad performance of all methods at 40^{th} and 42^{nd} moments of time. First, triad trajectories included in the test set concern staff members only from several departments at the university. Second, those days (around the 15^{th} of April) special events took place at those departments, e.g., conferences and open days. Therefore, there was an extra activity at the e-mail server. These moments could be treated as anomalies or outliers that are impossible to predict without any additional context knowledge.

The results for the *Random* are not shown since this method returned results exceeding mean error value of 98% for all cases.

VI. CONCLUSIONS

Our experiments have shown that it is possible to predict the evolution of the links in node triads of fast-changing social network with a good accuracy. Is is also interesting that the dynamic network structures built from real-life datasets reflect the influence of external events which may significantly distort the network structure, which was visible in our experiments. This results are preliminary and form the basis of our future experiments which will be carried on in the following directions:

1) The classification of nodes according to their activity patterns.
2) The link prediction method based on prediction of the triad structure.
3) The structure prediction – building the characteristics of network groups from the triad evolutionary patterns.
4) Including link attributes in the analysis. The obvious one is link weight; a link may exist as a consequence of sending one or many messages, and in most cases it is far more stable in the second case. This issue may be used to tune our method.
5) Application of more complex probabilistic models. First of all, instead of first order Markov chains the hidden Markov models [21] should be applied. Further, the nonparametric Bayesian approach [10] could be presumably used to increase the accuracy and automatize the whole process of inference.

Acknowledgments

The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

[1] L. A. Adamic, E. Adar, *Friends and neighbors on the web*, Social Networks, Vol. 25, No. 3, pp. 211-230, 2003

[2] A.-L. Barabsi, *The origin of bursts and heavy tails in humans dynamics*, Nature, Vol. 435, pp. 207, 2005

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, Singapore, 2006

[4] D. Braha, Y. Bar-Yam, *From Centrality to Temporary Fame: Dynamic Centrality in Complex Networks*, Complexity, Vol. 12, No. 2, pp. 59-63, 2006

[5] S. Dynes, P. Gloor, R. Laubacher, Y. Zhao, *Temporal Visualization and Analysis of Social Networks*, NAACSOS Conference, Pittsburgh PA, June 27 - 29, 2004

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

[7] L. Getoor, C. P. Diehl, *Link mining: a survey*, ACM SIGKDD Explorations Newslett., Vol. 7, pp. 3-12, 2005

[8] Z. Huang, D. K. J. Lin, *The Time-Series Link Prediction Problem with Applications in Communication Surveillance*, INFORMS Journal on Computing, Vol. 21, No. 2, pp. 286-303, 2009

[9] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, U. Alon, *Subgraphs in random networks*, Physical Review E., Vol. 68, 2003

[10] M. I. Jordan, *Bayesian Nonparametric Learning: Expressive Priors for Intelligent Systems*, In R. Dechter, H. Geffner, and J. Halpern (Eds.), Heuristics, Probability and Causality: A Tribute to Judea Pearl, College Publications, 2010

[11] K. Juszczyszyn, K. Musial, P. Kazienko, B. Gabrys, *Temporal Changes in Local Topology of an Email-Based Social Network*, Computing and Informatics Vol. 28, No. 6, pp. 763-779, 2009

[12] K. Juszczyszyn, K. Musial, M. Budka. *Link prediction based on subgraph evolution in dynamic social networks*, 2011 IEEE Third International Confernece on Social Computing (SocialCom), pp. 27–34. IEEE, 2011.

[13] N. Kashtan, S. Itzkovitz, R. Milo, U. Alon, *Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs*, Bioinformatics, Vol. 20, No. 11, pp. 1746-1758, 2004

[14] D. Kempe, J. Kleinberg, A. Kumar, *Connectivity and inference problems for temporal networks*, Journal of Computational System Science, Vol. 64, No. 4, pp. 820-842, 2002

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

[16] M. Lahiri, T. Y. Berger-Wolf, *Mining Periodic Behavior in Dynamic Social Networks*, ICDM, pp.373-382, 2008

[17] D. Lieben-Nowell, J. M. Kleinberg, *The link-prediction problem for social networks*, JASIST (JASIS), Vol. 58, No. 7, pp. 1019-1031, 2007

[18] S. Mangan, U. Alon, *Structure and function of the feedforward loop network motif*, Proc. of the National Academy of Science, USA, Vol. 100, No. 21, pp. 11980-11985, 2003

[19] S. Mangan, A. Zaslaver, U. Alon, *The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks*, J. Molecular Biology, Vol. 334, 197-204, 2003

[20] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, *Network motifs: simple building blocks of complex networks*, Science, Vol. 298, pp. 824-827, 2002

[21] L. R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, Vol. 22, Issue 2, pp. 257-286, 1989

[22] A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, A. Barabasi, *The topological relationship between the large-scale attributes and local interaction patterns of complex networks*, Proc. Natl Acad. Sci. USA, Vol. 101, No. 17, pp. 940, 2004

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

[24] E. Young-Ho, L. Soojin, J. Hawoong, *Exploring local structural organization of metabolic networks using subgraph patterns*, Journal of Theoretical Biology, Vol. 241, pp. 823-829, 2006