Different Approaches to Groups and Key Person Identification in Blogosphere

Anna Zygmunt[1], Piotr Bródka[2], Przemysław Kazienko[2], Jarosław Koźlak[1]
[1] Department of Computer Science, AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland
[2] Institute of Informatics, Wrocław University of Technology, Wyb.Wyspiańskiego 27, 50-370 Wrocław, Poland
azygmunt@agh.edu.pl, piotr.brodka@pwr.wroc.pl, kazienko@pwr.wroc.pl, kozlak@agh.edu.pl

***Abstract*—Two approaches to key person identification in the blogosphere-based social network are analysed in the paper: discovery of most important individuals either in persistent or in global social communities existing on web blogs. A new method for separation of stable groups fulfilling given conditions is presented. Additionally, a new concept for extraction of user roles and key persons in such groups is proposed. It has been compared to the general clustering method and structural node position measure applied to rank users in the time-aggregated data. Experimental, comparative studies have been conducted on real blogosphere data gathered over one year.**

Keywords-social network, social network analysis, SNA, social group extraction, key person, persistent role identification, blogosphere, stable group extraction, CPM, fast modularity optimization, node position

I. INTRODUCTION

Blogosphere is a general term denoting all blogs interconnected each other and managed by a single subject. It gathers people who want to share with others their thoughts, remarks and experiences. An important feature of blogs is its ability to bind people via either direct hyperlinks or opinions provided to blog posts or other comments. This interaction of humans is a crucial, social profile of blogosphere, which makes Web 2.0 phenomenon more and more important in recent societies.

Social network analysis (SNA) methods, in turn, enable to study the data registered in the blogosphere in the numerical manner. The main issue analysed in this paper are different approaches to identification of key (most influential) persons active on blogs. This can be done by extraction of social groups, by means of various clustering methods, and discovery of individuals who are most important in these smaller communities. Additionally, the analyses can be carried out either on regular clusters or on persistent groups. Depending on the method partly different and partly coincident results can be achieved.

II. RELATED WORK

A. Social Network Analysis in Blogosphere

During the last decades one can observe enormous changes in the forms of activity in the Internet. Users from passive consumers of information have become their producers too. Internet social media can occur in various forms [21]: blogs (e.g. Blogspot[1]), forum (e.g. Yahoo!answers[2]), media sharing (e.g.YouTube [3]), microblogging (e.g. Twitter [4]), social networking (e.g. Facebook[5]), wikis (e.g. Wikipedia[6]). Internet social media has revolutionized the Internet; many believe that in very near future, Internet will be main or even only global information media.

Blogs play a special role in creating opinion and information propagation. They are some kind of the Internet diary, where an author gives opinions on some themes or describes interesting events and readers comment on these posts. A typical entry on a blog can consist of text, photos, films and links to other blogs or web pages. Posts can be categorized by tags. A very important element of blogs is the possibility of adding comments, which allow discussions. Access to blogs are generally open, so everybody can read the posts and comments.

Basic interactions between bloggers are writing comments in relation to posts or other comments. Blogosphere (space all blogs) are very dynamic, every day thousands of new posts and millions of new comments are written. Thus the relationships between bloggers are very dynamic and temporal: the lifetime of the posts is very short. The huge space of blogosphere constitutes the source of information and is intensively explored [1].

Networks based on blogs, posts and comments, can be analysed by SNA centrality measures due to finding, for example, the most important or influential bloggers. Around such bloggers the groups are forming, sharing similar interests or, for example, politics. SNA measures and their interpretations in relation to blogs are discussed in [25][11].

B. Community Extraction

There is a difficulty to find in literature unequivocal definition of a group, acceptable to everybody [1][22]. Wasserman in [24] points out that this term has been widely used in social science without formal definition. In the literature, a growing interest in research connected with

[1] http://www.blogspot.com
[2] http://answers.yahoo.com
[3] http://www.youtube.com
[4] http://twitter.com
[5] http://www.facebook.com
[6] http://wikipedia.org

The work was supported by: The Polish Ministry of Science and Higher Education, the development project, 2009-2011, The Polish Ministry of Science and Higher Education the research project, 2010-13 and Fellowship co-Financed by European Union within European Social Fund.

identification and understanding groups and communities in social networks has been observed [1][21][14][24]. Usually the terms such as community, cluster, module or coherent subgraph are used interchangeably according to context [21]. In [1] the distinction between groups and communities has been made: a group is a subset of individuals in a population, which of them is identified as belonging to a specific organization, whereas community is a subset of individuals in a social network, which are stronger connected between themselves than others in a network. In this article we will use these terms interchangeably.

The idea of finding groups in a social network is to identify a set of vertices, which communicate to each other more frequently than with vertices outside the group [8][21]. That is, a group can be treated as a dense subset of vertices in a network, which are loosely connected with vertices outside the group. In practice, in complex social networks, groups are not isolated, but individuals can be, in given time, a member of many groups.

Many methods of finding coherent groups has been proposed, most of them are proposed for concrete applications (in [1] there are detail descriptions of more popular methods and algorithms). Interesting approach to systematizing these methods in four categories: node-centric, group-centric, network-centric and hierarchy-centric has been proposed in [21][23]. Methods based on node-centric criteria require each node in a group to satisfy certain properties (such as complete mutuality or reachability). CPM [17] is a good example of these group methods (described in more details in this article). In turn, methods based on group centric criteria considers connections inside a group as a whole. It's acceptable, for example, that some nodes in a group are loosely connected as far as a whole group satisfies certain properties. Group identification using network-centric criteria takes into account global network topology as a whole. Nodes of a network are divided into some number of disjoint sets. Methods based on graph partitioning can be a good example. The last category - hierarchy-centric – consists of methods which built a hierarchical structure of groups based on network structure. Example of this group can be the popular edge betweenness algorithm [13][14].

C. Key Person Extraction

Most of the key person extraction methods relay on various centrality measures, i.e. local (within social communities) or global (for the entire network) structural features [3], [9], [13]. Much research has been conducted in the domain of their application, especially in the context of spread of knowledge or influence [5], [20] as well as terrorist group analysis [15].

III. Group Extraction

Two approaches were considered. The first one concerned the existence span of the groups, expressed by the time regularity of the interactions between group members, the second one identified groups considering the data about interactions from the whole period together.

The first approach concerned the presentation of the society and their detailed changes in subsequent periods whereas the second gives general view of the organisation of the society.

A. Finding Overlapping Stable Communities

Group extraction is based on the algorithm CPM (Clique Percolation Method) [18][19] which is based on finding in a graph *k-cliques*, which means a complete, fully connected subgraph of *k* vertices, where every vertex can be reached directly from all other vertices. Groups are treated as sets of adjacent *k-cliques* (having *k-1* mutual vertices). Therefore one vertex can belong to several groups, which is a good reflection of the real situation, where every blogger can be a member of many groups. For every value of *k* algorithm should be started separately, value of *k*. The basic version of algorithm applies to undirected graph, where between every pairs of vertices there can only be one edge at most. In case of blog analysis we can find groups in directed graphs. For such graphs the algorithm is slightly different. For every clique double edges are eliminated, such that between pairs of vertices there is one edge at most and this edge is directed from the vertex of lower in-degree to vertex of higher in-degree. Additionally, in the directed clique two vertices with the same in-degree or out-degree cannot exist.

The CFinder[7] program, based on CPM algorithm generates different groups according to the value of *k* parameter. By reason of making analysis of groups changing in time, CFinder is running repeatedly with data from consecutive time periods *t*. Such created groups from contiguous periods are then merged in greater communities, according to some conditions (described in more details in IV.A).

B. Fast Modularity Optimization (Blondel)

The growing number of large networks has created a need for very fast group extraction algorithm. Responding to this demand, Blondel, Guillaume, Lambiotte and Lefebvre [4]. have created a method called: *Fast Modularity Optimization* or *Blondel*. Computational complexity of the method is $O(|E|)$, where E is the set of the edges in the networks, so it is very fast and a great problem for it is the disk write speed performance rather than the calculations speed.

The method originates from the modularity of a network that is a measure describing whether a network is well grouped. The modularity Q is defined as follows [12]:

$$Q = \frac{1}{\sum_{x,y \in V} w(x,y)} \cdot \sum_{x,y \in V} \left[\left(w(x,y) + w(y,x) - \frac{DC(x)DC(y)}{\sum_{i,j \in V} w(x,y)} \right) \delta(G_x, G_y) \right]$$

[7] http://cfinder.org

where: V – the set of network nodes; $w(x,y)$ – the weight of the edge from x to y; $DC(x)$ – degree centrality of node x and similarity measure $\delta(G_1,G_2)$ for two groups G_1 and G_2 is:

$$\delta(G_1,G_2)=\begin{cases}1 \ when \ G_1=G_2\\ 0 \ when \ G_1\neq G_2\end{cases}$$

Since the optimization of this measure is NP-complete [2], the approximating algorithms are used for large networks.

Fast optimization algorithm is as follows:

1. Place each node in a separate group
2. For each vertex x remove it from its group, put it in a group G_y of its neighbour y separately for each neighbour y and calculate their modularity increase $\Delta Q(G_y,x)$. Leave neighbour x in the group for which the modularity increase is the highest. If modularity increase $\Delta Q(G_y,x)$ is not positive for all neighbours y ($\Delta Q(G_y,x)\leq 0$) than node x stays in its original group.
3. Repeat step 2 until the modularity can no longer grow, i.e. for all nodes x in the network and all their neighbours y their $\Delta Q(G_y,x)\leq 0$.
4. Build a new network by replacing the separate groups with the super-nodes. The super-nodes are connected if at least one vertex in the two super-nodes are connected. However, the edge weight is the sum of weights of all edges between nodes located in super-nodes.
5. Repeat steps 1-4 until there are no more changes and a maximum of modularity is achieved.

The modularity increase $\Delta Q(G,x)$ is calculated as follows (see [12] for derivation of this formula):

$$\Delta Q(G_y,x)=\left[\frac{D^{in}(G_y)+d^{in}(x)}{2m}-\left(\frac{D(G_y)+DC(x)}{2m}\right)^2\right]-\left[\frac{D^{in}(G_y)}{2m}-\left(\frac{D(G_y)}{2m}\right)^2-\left(\frac{DC(x)}{2m}\right)^2\right]$$

where: $m=\sum_{x,y\in V} w(x,y)$; $D^{in}(G_y)$ – group internal degree; $D(G_y)$ – group degree; $d^{in}(x)$ – node internal degree in the group G_y; $DC(x)$ – node degree centrality in the entire network.

The only downside of this algorithm is the fact that it is dependent on the order of the processed nodes. However, this dependency is not yet fully known.

IV. A Method for Key Person Identification

An important issue in social network analysis of individuals is their role and social position either in relation to the entire population studied (global analysis) or to the selected, smaller community (local analysis). The latter is further considered.

A. Identification of Roles in Persistent Groups (IRPG)

The developed method for the analysis of society needs to partition the analysed period of time, for which the data of interactions was gathered, to subsequent T periods with the same length, for example, the subsequent weeks or months. We assume that T of such periods were distinguished and that they have numbers from 0 to T-1. Overall, one can assume that either these periods are separable or partly overlapped. In the experimental studies, we assumed that they have a length of 30 days and that the neighbouring periods overlaps one with another for 15 days.

For each of these periods the social network was generated and the fundamental SNA measures were calculated. These measures are taken into consideration in the process of the identification of key members of the identified communities.

The algorithm consists of three subsequent steps:

Step 1. Identification of groups and their members for the subsequent time periods. To achieve it, the algorithm CPM described in Section III.A is used. As a result, of the first step for the given time periods t, sets of groups $G_i(t)$ are identified. Each of them consists of nodes $n_i(t)$ having strong connections in the considered time period.

Step 2. Identification of groups which exist for minimal required period of time. It is realised using a following group continuation condition: at least x% of members of the group in the time period t should be members of the group in the time period t+1. In the tests, we assumed that $x>50$. A set of groups G_i, which consists of every member of the groups being their continuations in the subsequent periods j, j+1, ... , j+s is identified and is defined as follows: $G_i=\bigcup_{t=j}^{j+s} G_i(t)$.

Ephemeral groups that do not last for at least t_{req} ($s< t_{req}$) are not taken into consideration in the following analysis. In the experiments is assumed that $t_{req}=3$.

Step 3. The identification of key members of the group. A *core member* of the group has to fulfil the following conditions: (i) be a respected member of blogosphere. i.e. those who are more often commented than make comments themselves and (ii) be present in the group over almost whole time of its existence. An *active members* should belong to the group in a stable way. The guests may have a membership in the group only in some single time periods. Step 3 will be described in more detail in the following section.

B. Identification of Roles in Groups and Key Members (IRGKM)

Each node n_i is described by a set of values of calculated SNA measures and ranking scores obtained as a result of them. These measures are used in two manners:

- measures used for the local comparisons inside the groups, the values of these measures have to be high enough in comparison to values of other members of the group to guarantee that a node may have a given role assigned.

- roles used for building of the score of node-importance – measures where ranking scores are used for building a joint measure – a score.

The third factor used to assign a role in the group is the frequency of the classification of the node *i* to the group, in the considered time horizon *T*:

- k^c_{min} – minimal number of participations in the group in the analysed time horizon *T*, necessary to be assigned to the core (in experiments the value *T*-1 was assumed)
- k^a_{min} – minimal number of participations in the group in the analysed time horizon *T*, necessary to be assigned to the active members (in experiments (*T*+1)/2 was assumed).

For the measures from the groups, (1) and (2) of the following representations are assumed:

- (1) set of values of measures $m_i^{j,t}$ or their ranks in the group -- for given kinds of nodes *i*, kinds of measures j and analysed time periods *t* (from 0 to *T*-1). The used set of measures and conditions may be different for the classification to core or to active members. For all nodes, the following constants are taken into consideration:
  - α_j^c - a minimal value of measure *j* necessary to be assigned to the core,
  - α_j^a - a minimal value of measure *j* necessary to be assigned to the core,.

  The following SNA measures were considered:
  - m^1 - ratio of incoming degree to outgoing degree of connections with members of the group (in experiments: α_1^c =2, α_1^a =0.5),
  - m^2 - a percentage of group members, of which the analysed node should have a higher value of sum of incoming and outgoing degrees of connections with members of the group (in experiments: α_2^c =50%, α_2^a =30%),
- (2) set of scores obtained owing to the ranking positions of measures, i.e. $r_i^{j,t}$ - for given kinds of measures *j* and analysed periods *t* (from *0* to *T*-1). Among different measures, we decided to focus on measures which describe authority of the node in the network and its strong influence on other nodes (number of comments to the posts). The following measures were used in the experiments:
  - r_1- PageRank,
  - r_2 - Authority,
  - r_3 – Incoming degree,

As a result for every nodes *i* a score ranking $score_i$ is calculated which is a sum of the points obtained for position in ranking for given measures, in every considered time period. This measure is compared with measures $score^c_{min}$ and $score^a_{min}$ to check if it fulfils conditions necessary for assigning the node to a core or to active members. Particularly:

- $score^c_{min}$ is a percentage of group members, of which the analysed node should have a higher measure of the score, to be assigned to the core (in experiments: 50%),
- $score^a_{min}$ is a percentage of group members, of which the analysed node should have a higher measure of the score, to be assigned to the core (in experiments: 30%).

C. Node Position

Node position function *NP*(*x*) of individual *x* in the social network can be used to evaluate importance of *x* in community. It respects the values of node positions of *x*'s direct acquaintances as well as their activities towards *x*, in the following way [3][16][9]:

$$NP(x) = (1-\varepsilon) + \varepsilon \cdot \sum_{y \in Y_x} NP(y) \cdot C(y \to x)$$

where Y_x – *x*'s nearest neighbours, i.e. members who are in direct relationship to *x*; $C(y \to x)>0$ is the function that denotes contribution in activity of *y* directed to *x*, see [9] for details on its calculation; ε - the constant coefficient from the range [0;1].

The value of ε denotes the openness of node position measure on external influences, i.e. how much *x*'s node positions is more static and independent (small ε) or more influenced by others (greater ε). Node position is calculated in the iterative way with stop condition (precision), see [3], [16] for details.

In general, the greater node position one possesses the more valuable this member is for the community. The node position of user *x* is inherited from the others but the level of inheritance depends on the activity of the users directed to this person, e.g. intensity of common activities on blogs. Thus, the node position depends both on the number and quality of relationships.

V. Experiments

A. Data Set

The analysed data about blogs was taken from the portal www.salon24.pl, which is dedicated especially to political discussions, but also subjects from different domains may be brought up. The data consists of 9,880 users, 2,632 blogs, 27,189 posts and concerns the discussions as of 2009.

From those data the social network was extracted consisting of 9,807 not isolated nodes and 153,596 edges between them.

The network users are bloggers and registered users who do not keep blogs, but they write comments to others blogs. However, the authors of those blogs that no one commented were rejected, so the number of nodes in the network is lower than the number portal users. Additionally the number of comments is greater than the number of edges, because many comments was between the same authors. Multiple comments between the same pair of users was reflected in the strength of the relationship.

B. Group Extraction

For extracted social network two different community extraction algorithms was applied.

The first one was the CPM method (see Section III.A) for different values of k (from 3 to 9) and for subsequent 30-day timeframes.

The second one was Blondel approach (see Section III.B) which was applied to whole network. Only the first level communities was taken into account in further studies .

The number of communities, with given size, extracted by both methods was presented on Figure 1 and Figure 2 as a percentage distribution.

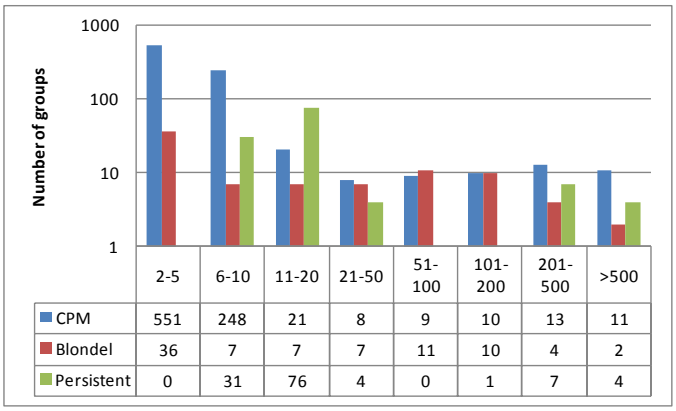

Fig 1. Numbers of communities with given sizes for CPM and Blondel methods, and persistent groups living for two timeframes.

While analysing the results we can see that CPM has extracted ten times more communities (871) than Blondel (84). This is because the CPM was, in fact, used on 12 different networks (12 timeframes) and some of the communities are actually the same group but existing over two or more timeframes, i.e., persistent communities.

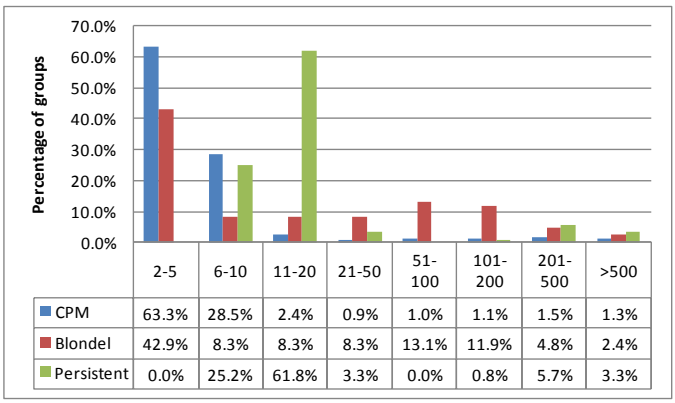

Fig. 2. Percentage distribution of communities with given sizes for CPM and Blondel methods, and persistent groups living for two timeframes

The persistent communities were selected after dropping too weakly connected groups and those groups which did not last long enough. One can notice that the most frequent are medium size groups, between 6 and 20 persons. The number of all persistent communities is 123, so it is quite similar to the number of global communities extracted by Blondel method.

C. Key person identification

The next step was to identify key persons. For all communities extracted by the Blondel method, the key users were extracted using node position approach, see Section III.C, applied separately to groups and to the whole network, see Table I, column no. 2 and 4. Based on these values, the ranking of users was established, see column no. 3 and 6.

Meanwhile, for two selected persistent groups the IRPG and IRGKM approaches where applied. Further analysis have focused on some of the most distinguished groups. Others groups were less stable and less dense.

The identification of the key persons by means of both approaches (IRPG/IRGKM for persistent groups and Blondel/node position) provided slightly different results, see Table I. The table includes only twenty one *Core* and *Active* users from the persistent groups.

TABLE I. THE COMPARISION OF RANKINGS (POSITION OF INDIVIDUAL USERS) RETURNED SEPARATELY BY NODE POSITION MEASURE APPLIED TO THE ENTIRE NETWORK (COL. 2, 3), BY NODE POSITION APPLIED INDEPENDENTLY ONLY TO BLONDEL GROUPS (COL. 4, 5, 6) AND BY IRPG/IRGKM METHOD USED FOR PERSISTENT GROUPS (COL. 7, 8)

User id	Global NP	Rank in the network	Group NP	Blondel group no.	Rank in group	Role	Persistent group no.
1.	2.	3.	4.	5.	6.	7.	8.
5580	55.61	3	39.97	27	1	Core	I
9925	52.26	4	19.37	13	1	Active	I
5423	46.60	5	21.01	32	1	Core	I
4868	35.01	9	10.00	13	5	Active	I
4861	33.30	10	12.23	13	3	Active	I
5153	13.45	60	5.36	13	31	Active	I
8059	13.06	64	5.42	13	29	Active	II
9948	8.62	163	2.09	13	197	Active	I
5458	8.54	167	4.21	17	1	Core	I
6032	8.10	178	4.27	13	60	Core	II
10109	5.36	331	2.34	13	173	Active	I
9929	4.82	377	3.41	13	87	Active	II
7005	4.37	425	2.05	13	200	Active	II
5504	3.29	560	1.64	13	266	Active	II
6451	2.83	668	1.54	16	314	Active	II
6293	2.76	687	1.35	18	29	Core	I
11467	2.73	692	1.86	13	227	Active	II
11437	2.72	695	1.75	13	243	Active	II
6312	1.58	1144	1.12	13	387	Active	II
11227	1.05	1607	0.76	13	548	Active	II
11739	0.75	2060	0.48	13	849	Active	I

While analysing rankings provided by different approaches, one may observe that *Core* and *Active* users (IRPG/IRGKM method) from persistent groups are rather higher in the ranking obtained by regular approach (Blondel/node position). It means that both approaches provide similar knowledge. On the other hand, these rankings not fully correspond each other and it happens that some users, e.g. active user no. 11739 is only 849 in his group. Note that such position is still quite high since group no. 13 is the second largest community; it possesses 2706 members.

Overall, the experimental results have shown that different approaches may be used for different purposes. If one would like to obtain key users in persistent, stable groups that last for

longer period, then IRPG/IRGKM method appears to be reasonable. However, if we want to extract persons who are important for the entire social network or may be influential within smaller communities, typical structural measures like node position may be effective enough.

VI. Conclusions and Future Work

Two separate methods for group extraction have been presented and analysed in the paper: (i) groups provided by Clique Percolation Method (CPM) and filtered to leave only stable ones over a given timeframe as well as (ii) Fast Modularity Optimization (Blondel) method for clustering. Different approaches for key person identification have been applied to the obtained groups: (i) identification of roles in persistent groups (IRPG) and identification of roles in groups and key members (IRGKM) as well as (ii) node position.

Results of experimental studies have revealed that separate approaches provide partly similar and partly different output. (ranking of network/group members) For that reason, decision on what method to use should be undertaken based on the purpose of analyses.

Future work will focus on comparison of other clustering methods, other centrality measures and sociological interpretation of results.

Acknowledgment

The authors are indebted to students from Department of Computer Science AGH-UST, especially L. Krupczak, who participated in the development of the systems used for presented analysis

References

[1] N. Agarwal, H. Liu, Modeling and Data Mining in Blogosphere, Morgan&Claypool, 2009.

[2] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, D. Wagner, Maximizing modularity is hard, 2006 http://arxiv.org/abs/physics/0608255.

[3] P. Bródka, K. Musiał, P. Kazienko, A Performance of Centrality Calculation in Social Networks. CASoN 2009, IEEE Computer Society, 2009, 24-31.

[4] V.D. Blondel, J. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, 2008, p. P10008.

[5] E. Even-Dar, A. Shapira, A note on maximizing the spread of influence in social networks, WINE'07, The 3rd International Conference on Internet and Network Economics, Springer, 2007.

[6] S. Fortunato, Community detection in graphs, Phys. Rep. 486, 2010.

[7] N. Geard, S. Bullock, Group formation and social evolution; a computional model, Proc. of the Eleventh International Conference on the Simulation and Synthesis of Living Systems, MIT Press, uSA, 2008.

[8] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci., USA, 2002.

[9] P. Kazienko, K. Musiał, On Utilising Social Networks to Discover Representatives of Human Communities, International Journal of Intelligent Information and Database Systems, 1 (3/4), 2007, 293-310.

[10] P. Kazienko, K. Musiał, A. Zgrzywa, Evaluation of Node Position Based on Email Communication. Control and Cybernetics, 38 (1), 2009, 67-86.

[11] J. Koźlak, A. Zygmunt, E. Nawarecki, Modelling and analysing relations between entities using the multi-agent and social network approaches, MCSS 2010, IEEE International Conference, Kraków, 2010

[12] M. E. J. Newman, Analysis of weighted networks, Physical Review E, 70, 056131, 2004.

[13] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E, 69, 026113, 2004.

[14] M.E.J. Newman, Networks: An Introduction, Oxford University Press, 2010.

[15] N. Memon, H.L. Larsen, D.L. Hicks , N. Harkiolakis, Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies, Intelligence and Security Informatics, LNCS 5075, Springer, 2008, 477-489

[16] K. Musiał, P. Kazienko, P. Bródka, User Position Measures in Social Networks. SNA-KDD at KDD 2009, ACM Press, Article no. 6, 2009.

[17] G. Palla, D. Abel, I. Derényi, I. Farkas, P. Pollner, T. Vicsek, K-clique percolation and clustering in directed and weighted networks, Bolayai Society Mathematical Studies, 2005.

[18] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435, 814–818, 2005.

[19] G. Palla, P. Pollner, A.-L. Barabasi, T. Vicsek, Social group dynamics in networks, in Adaptive networks, ed. T. Gross, H. Sayama, Springer Berlin/Heidelberg, 2009.

[20] J. Tang, J. Sun, C. Wang, Z. Yang, Social influence analysis in large-scale networks, KDD '09, The 15th ACM SIGKDD Int. Conference on Knowledge discovery and Data Mining, ACM, 2009.

[21] L. Tang, H. Liu, Community detection and data mining in social media, Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan&Claypool Publishers, 2010.

[22] L. Tang, H. Liu, Graph mining applications to social network analysis, in Managing and Mining Graph Data, ed. C. Aggarwal, X. Wang, 2010.

[23] L. Tang, Learning with large-scale social media network, PhD thesis, Arizona State University, 2010.

[24] S. Wasserman, K. Faust, Social network analysis: methods and applications, Cambridge University Press, 1994.

[25] A. Zygmunt, J. Koźlak, Ł. Krupczak, Identifying the influential individuals in blogosphere, Studia Informatica, vol. 31, no 2A(89), 2010 (in Polish).