*Journal of Universal Computer Science, vol. 18, no. 4 (2012), 577-597*
*submitted: 29/9/11, accepted: 14/12/11, appeared: 28/2/12 © J.UCS*

# Key Person Analysis in Social Communities within the Blogosphere

**Anna Zygmunt**
(AGH University of Science and Technology, Kraków, Poland
azygmunt@agh.edu.pl)

**Piotr Bródka**
(Wrocław University of Technology, Wrocław, Poland
piotr.brodka@pwr.wroc.pl)

**Przemysław Kazienko**
(Wrocław University of Technology, Wrocław, Poland
kazienko@pwr.wroc.pl)

**Jarosław Koźlak**
(AGH University of Science and Technology, Kraków, Poland
kozlak@agh.edu.pl)

**Abstract:** Identifying key persons active in social groups in the blogosphere is performed by means of social network analysis. Two main independent approaches are considered in the paper: (i) discovery of the most important individuals in persistent social communities and (ii) regular centrality measures applied either to social groups or the entire network. A new method for separating of groups stable over time, fulfilling given conditions of activity level of their members is proposed. Furthermore, a new concept for extracting user roles and key persons in such groups is also presented. This new approach was compared to the typical clustering method and the structural node position measure applied to rank users. The experimental studies have been carried out on real two-year blogosphere data.

**Keywords:** social network, social network analysis, SNA, social community extraction, social group, persistent group, stable social group, key person, role identification, blogosphere, CPM, fast modularity optimization, node position
**Categories:** E.m, F.2.m, G.2.2, G.2.3, H.3.1, H.3.5, H.4.3, I.5.1, I.5.3, I.5.4, I.7.4, I.7.5, J.4, K.4.2, L.6.1, L.6.2, M.0

## 1 Introduction

The general term 'blogosphere' denotes all web-based blogs interconnected with each other and managed by a single subject, usually a company. It gathers users who want to share their experiences, remarks, and thoughts with other people. Typically, blogs enable linking of humans via either direct hyperlinks or by means of opinions or comments provided to blog posts. This human interaction is simultaneously a crucial, social phenomenon of blogosphere, making blogs, a typical Web 2.0 service, very important for recent societies.

Social network analysis (SNA) methods, in turn, allow us to study the data registered in the blogosphere in a numerical manner. The main issue analysed in this paper are different approaches to identifying key (most influential) persons active on blogs. This can be done by extraction of social groups, by means of various clustering methods, and the 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.

The work presented in this paper significantly extends the research published in proceedings of the ASONAM 2011 conference [Zygmunt, 11].

## 2 Overview of Research

### 2.1 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 in turn producers of it. Internet social media can occur in various forms [Tang, 10]: blogs (e.g. Blogspot), forum (e.g. Yahoo!answers), media sharing (e.g. YouTube), microblogging (e.g. Twitter), social networking (e.g. Facebook), wikis (e.g. Wikipedia). Internet social media has revolutionized the Internet; many believe that in the very near future, Internet will be the main or even the only global information media.

Blogs play a special role in creating opinion, information propagation and knowledge sharing [Tang, 10b], [Dolińska, 10], [Bross, 11], [Jung, 08]. They are some kind of an Internet diary, where an author gives opinions (called posts) 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. Posts are arranged in reverse chronological order. A very important element of blogs is the possibility of adding comments, which allow discussions [Jung, 09]. Access to blogs are generally open, so everybody can read the posts and comments.

The basic interactions between bloggers are writing comments in relation to posts or other comments. Blogosphere (space of 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 posts is very short. The huge space of blogosphere constitutes the source of information and is intensively explored [Agarwal, 09], [Jung, 10a].

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, groups form, sharing similar interests or, for example, politics. SNA measures and their interpretations in relation to blogs are discussed in [Macskassy, 11], [Zygmunt, 10], [Jung, 10b], [Koźlak, 10], [Koźlak, 11].

There are many kind of blogs: personal diaries, research, political, technological etc.. In our research we have been analysing political blogosphere. Adamic in [Adamic, 05] first applied social network analysis to the political blogosphere, analysed 2004 U.S. presidential election (first in which blogging played an important roles) and discovered groups of conservative and liberal blogs. They found that their internal structure differed significantly and there were interesting patterns of communication between different political groups. In [Yano, 10], authors defined the popularity in blogosphere as a great number of received comments and tried to find the relationship between the content of a political blog post and the number of comments post could receive. On that basis, they attempted to predict which posts would be popular.

### 2.2 Community Extraction

The existence of groups (communities) in social networks is intuitively obvious [Porter, 09] and has been studied for a long time in sociology and anthropology.

There is a difficulty to find in literature an unequivocal definition of a group, acceptable to everybody [Wasserman, 94], [Agarwal, 09], [Tang, 10a], [Jung, 11], so the term has been widely used without formal definition. Group is generally considered to be a set of nodes that are connected "more densely" to each other than to the nodes outside the group [Evans, 09], [Fortunato, 10], [Porter, 09]. But others try to define a group as a set of closely interrelated links rather than a set of nodes[Evans, 09], [Ahn, 10], [Jung, 12a].

In the literature, a growing interest in research related to identification and understanding groups and communities in social networks has been observed [Agarwal, 09], [Tang, 10], [Newman, 10], [Wasserman, 94], [Jung, 12b]. A major breakthrough was done in 2002 as a result of the paper by Girvan and Newman with a proposition for a graph partitioning algorithm [Girvan, 02] which became very attractive for a broad group of researchers, especially physicists and mathematicians.

Discussions whether the groups are disjointed or overlap has been ongoing [Palla, 05a], and whether such partitions are at one level or form some kind of hierarchical structure (each partition could be divided recursively) [Fortunato, 10], [Girvan, 02], [Porter, 09]. The last approach better reflects the hierarchical nature of many real networks [Ahn, 10], [Lancichinetti, 09], [Evans, 09]. In [Lancichinetti, 09] the method of finding simultaneously both hierarchical and overlapping groups was proposed. That method finds local maxima of a fitness function by local, iterative searching and a group is recognised as a peak in a fitness histogram.

Many methods of finding coherent groups have been proposed, most of them are proposed for specific applications (in [Agarwal, 09] there are detailed descriptions of the more popular methods and algorithms). An interesting approach to systematize these methods into four categories: node-centric, group-centric, network-centric and hierarchy-centric has been proposed in [Tang, 10b], [Tang, 10a]. Methods based on node-centric criteria require each node in a group to satisfy certain properties (such as complete mutuality or reachability). CPM [Palla, 05] is a good example of these group methods (described in more detail in this article). In turn, methods based on group centric criteria consider connections inside a group as a whole. It is 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 [Newman, 04a], [Newman, 10]. Recently, in [Yang, 11], authors tried to compare and evaluate several community detection algorithms on different small data sets. They came to the conclusion that different algorithms have different performance on different social networks and the quality of communities detected by algorithms is hard to evaluate.

### 2.3 Key Person Extraction

Two separate approaches to key person extraction in social networks may be enumerated: based on context roles played by individuals and based on structural network measures. The former approach is described in details in Sec. 4.1, while the latter in Sec. 4.2. In this paper, we will make use of both of these concepts.

The most common methods for key person extraction rely on various centrality measures calculated separately for individual nodes. These structural features can be either more local (reflect the position of the node within social community, commonly with respect to its neighbourhood) or more global (corresponding to overall node position for the entire network). The examples of the former are degree prestige, degree centrality, whereas global measures are represented by proximity prestige, rank prestige, node position, eccentricity, closeness centrality, betweenness, etc.[Bródka, 09], [Carrington, 05], [Fazeen, 11], [Kazienko, 07], [Musiał, 09], [Newman, 04], [Wasserman, 94]. Much research has been conducted in the domain of their application, for example in the context of spread of knowledge or influence [Even-Dar, 07], [Tang, 09] as well as terrorist group analysis [Memon, 08].

## 3 Social Community 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 method respects stability of user activities within the social group and its detailed changes in subsequent periods whereas the second one provides a general view of the society organization.

### 3.1 Finding Overlapping Stable Communities

The proposed method of finding stable groups is based on the algorithm CPM (Clique Percolation Method) [Palla, 05], [Palla, 05a] which finds in a graph *k-cliques*. *K-clique* means a complete, fully connected subgraph of *k* nodes, where every node can be reached directly from all other nodes. The method is based on the observation that communities consist of several small cliques that share many of their nodes with other cliques in the same community. Two cliques are adjacent if they share *k-1* nodes. A *k-clique* community is defined as a maximal union of *k-cliques* that can be reached from one to another through a sequence of adjacent *k-cliques*, so they share *k-1* nodes. The algorithm is described as a some kind of rolling a *k-clique* template from any *k-clique* in the graph to any adjacent *k-clique* by relocating only one of its nodes [Palla, 05], [Porter, 09].

Some nodes might never belong to any clique, but others can be part of several communities and it is a good reflection of the real situation, where every blogger can be a member of many groups (or any). For every value of *k* the algorithm should be started separately, with the increase of *k* the smaller and more disintegrated communities arise [Palla, 05a]. There is a suggestion that values of *k* = 3,...,6 seems to be the most appropriate [Porter, 09]. In [Ye, 11 ] indicated that in very dense networks we can find too many overlapping communities but in sparse networks it could be difficult to find enough numbers of connected cliques.

The basic version of the algorithm applies to an undirected graph, where between every pair of vertices there can only be one edge at most. In cases 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 node of lower in-degree to the node of higher in-degree. Additionally, in the directed clique, two nodes with the same in-degree or out-degree cannot exist.

The CFinder program[1], based on CPM algorithm generates different groups according to the value of *k* parameter. Due to the temporal nature of group changes, analysis of groups changing in time, CFinder is running repeatedly with data from consecutive periods *t*. That way created groups from contiguous periods are then merged in greater communities, according to some conditions (described in details in Sec. 4.1).

### 3.2 Fast Modularity Optimization (Blondel)

The growing number of large networks has created a need for a very fast group extraction algorithm. Responding to this demand, Blondel, Guillaume, Lambiotte and Lefebvre [Blondel, 08] 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 calculation 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 0:

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

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:

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

[1] http://cfinder.org

Since the optimization of this measure is NP-complete [Brandes, 06], 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 [Newman, 04] 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.

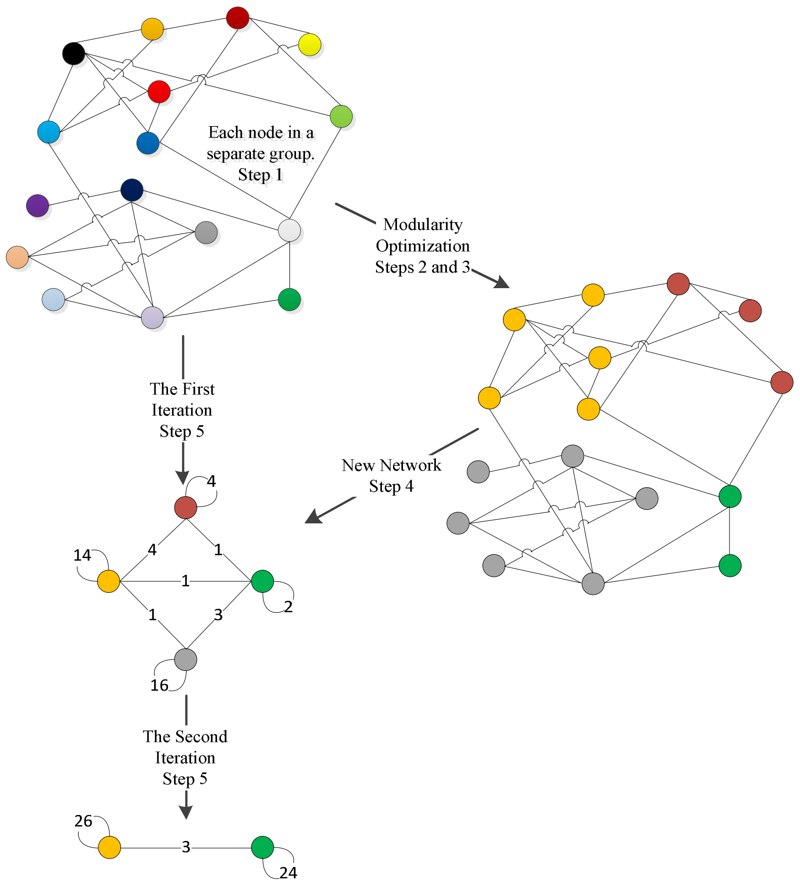
Each node in a separate group. Step 1
Modularity Optimization Steps 2 and 3
The First Iteration Step 5
New Network Step 4
4
14
4
1
1
2
1
3
16
The Second Iteration Step 5
26
3
24

*Figure 1: The example of Fast Modularity Optimization [Blondel, 08]*

## 4 Analysis of Key Persons in Persistent Social Groups within Blogosphere

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.

### 4.1 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, see Sec. 5, we assumed that they have the length of 30 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 identifying key members of the identified communities.

The idea of the algorithm for identifying the stable group and key group members is presented in Fig. 2. In the first step, the data about interactions is partitioned into the subsets which contain interactions from defined, subsequent periods. Then, CPM algorithm is used to identify the groups in the graphs constructed from data from each partition. In the next step, the algorithm tries to associate one group with another group from the neighbouring period partitions, which fulfil the criterion of having a sufficient number of common members. On the basis of such groups, the persistent groups which exist for the sufficient number of periods are defined. In these groups, the most influential members, called key persons, are identified.

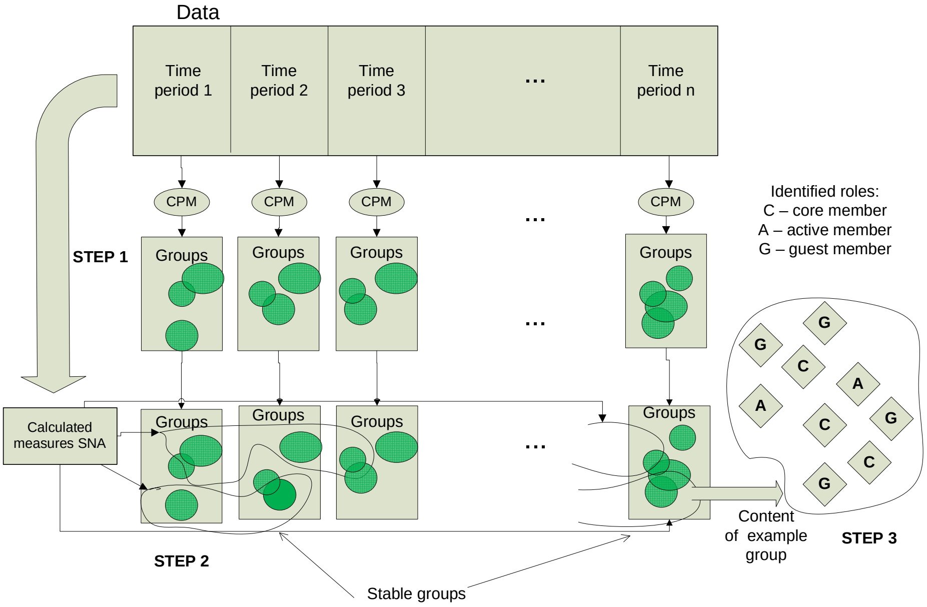

*Figure 2: Stable group and key group member identification algorithm*

The algorithm consists of three subsequent steps:

*Step 1. Identification of groups and their members for the subsequent periods*. To achieve it, the algorithm CPM described in section 3.1 is used. As a result, of the first step for a given period *t*, sets of groups $G_i(t)$ are identified. Each of them consists of nodes $n_i(t)$ having strong connections in the considered period.

*Step 2. Identification of groups which exist for a minimal required period of time*. It is realised using the following group continuation condition: at least *x%* of members of the group in period *t* should be members of the group in 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 it is assumed that $t_{req} = 3$.

*Step 3. The identification of key members of the group.* To be identified as a key person the following conditions should be fulfilled

- The presence in a given group should exceed a given percentage of periods of the whole group duration time ($k_{min}$).
- High enough sum of weights of incoming and outgoing arrows connected with the other members of the group (larger then $m_1$%**).**
- High enough value of ratio of sums of arrows incoming to outgoing, considering the links to the members of the group (larger then $m_2$)
- High enough ranking calculated on the basis of points assigned to a given node for the high positions of values of SNA measures considering the whole network. The node should have the ranking value higher than a given percentage of the group members $score_{min}$%).
- Having values of the selected measures larger than given percentages ($\alpha_j$) of the members of the considered group.

We distinguish two kinds of key persons in the groups. A *core member* of the group has to fulfil the following conditions: (i) be a respected member of blogosphere, i.e. those whose statements receive more comments than they make comments themselves and (ii) be present in the group over almost the whole time of its existence. An *active member* should belong to the group in a stable way. The remaining members of groups are assigned the role of guest.

In the analysis carried out, we assumed the following values of the parameters mentioned above, different for core member and active member roles:

- core member $k_{min}^c = T$-1, where $T$ – lifespan of the group, $m_1^c$ = 50%, $m_2^c$=2, $score^c_{min}$= 50%;
- active member $k_{min}^a$ = [$T$+1/2], $m_1^a$ =30%, $m_2^c$=0,5, $score^c_{min}$=30%.

To calculate the score value, we were taking into consideration the measures obtained in given periods during the presence of the node in the group, and calculated a sum on the basis of its position in the ranking of values of the measures, considering every node in the network. We tested different choices of measures, finally, in the performed experiments, we took into consideration PageRank, Authority and incoming degree.

## 4.2 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 [Bródka, 09], [Kazienko, 09], [Musiał, 09]:

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

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 0 for details on its calculation; $\varepsilon$ - the constant coefficient from the range [0;1].

The value of $\varepsilon$ denotes the openness of node position measure on external influences, i.e. how much $x$'s node positions is more static and independent (small $\varepsilon$) or more influenced by others (greater $\varepsilon$). Node position is calculated in the iterative way with stop condition (precision), see [Kazienko, 09], [Musiał, 09] 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. For example, node (A) in Fig. 3a has *NP*(A)=0.9 because it possesses one significant neighbour with *NP*=0.9, whose contribution *C*(B→A) towards (A) is relatively high – 0.6. Node (C) in Fig. 3b, in turn, inherits medium node position *NP*(C)=0.4, due to relatively small node position of its neighbours (*NP*(D)=0.25 and *NP*(E)=0.2), even though these neighbours are strongly engaged towards (C), with contribution *C*(D→C)=0.8 and *C*(E→C)=1.

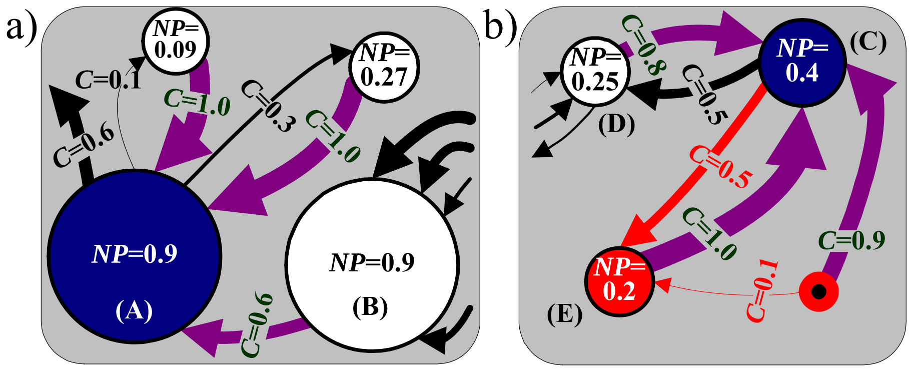

*Figure 3: Two samples of social network fragments with calculated node position centrality measures*

## 5 Experiments

The goal of conducted research was to apply two proposed methods for the same large data set and compare the obtained results.

### 5.1 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 19 966 users (among which 8 340 have blogs), 118 225 posts, 1 710 150 comments to posts and concerns the discussions as of 2009 and 2010.

Among all bloggers, 4 573 of them wrote at least 1 post in analysed period 2009-2010, 3 504 wrote at least 2 posts and 2 406 at least 5 posts. Taking into consideration the bloggers who wrote at least 1 post, the average number of posts for each blogger was about 25.85 posts.

The record post has 1 103 comments and 88 638 posts has more than 1 comment. In years 2009 and 2010, 12 145 users wrote at least one comment, 8900 wrote more than 1 comment and record holder wrote 1 297 wrote comments.

The average number of comments for one post (considering only posts having at least one comment) amounts about 19.29. The average number of commentaries for one blogger equals 205.05, it is an average value for every maintained blog. The popularity of the analysed portal is regularly increasing. In 2009 the users wrote 37 084 posts and 527 111 comments (14.2 comments/post), while in 2010 – 81 141 posts and 1 183 039 comments (14.5 comments/post).

### 5.2 Identification of Groups Existing in Given Periods

The groups with given ranges of member numbers in given periods (duration 1 month) were identified for the years 2009 and 2010. It was obtained using CPM algorithm, started with different *k* values (from 3 to 7), see Sec. 3.1. In Fig. 4, the numbers of groups calculated for given size ranges, for periods 2009, 2010, and 2009 and 2010 together were shown. It is possible to observe that for the year 2010 an important increase of the group number took place.

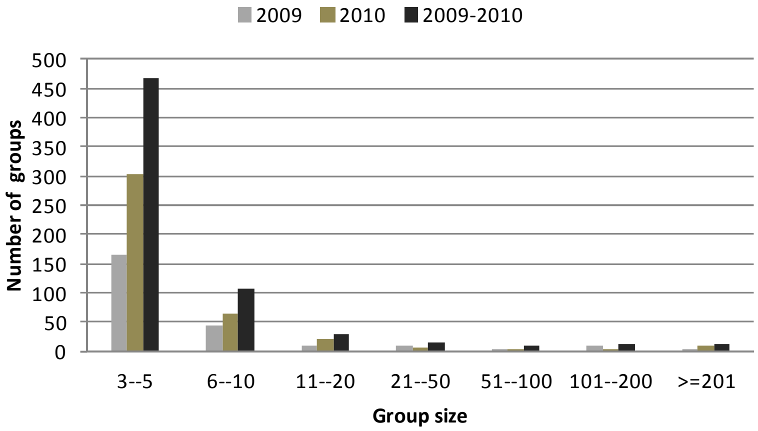

*Figure 4: Number of groups identified by CPM in 2009, 2010 and in total*

In the [Palla, 05], [Palla, 05a] (section 3.1) the use of *k*=3-6 was proposed. In this work, the groups for *k*=3-7 were generated, for *k*=7 only one group having size of 9 was found.

While analysing the results obtained by means of the second clustering method, we can see that CPM has extracted around ten times more communities than the Blondel method for each year and twenty times more for both years aggregated,

Fig. 5. This is because CPM was, in fact, used separately to 12 different networks (12 time frames) for each year and 24 for both years, and some of the communities are actually the same group but existing over two or more periods – persistent communities.

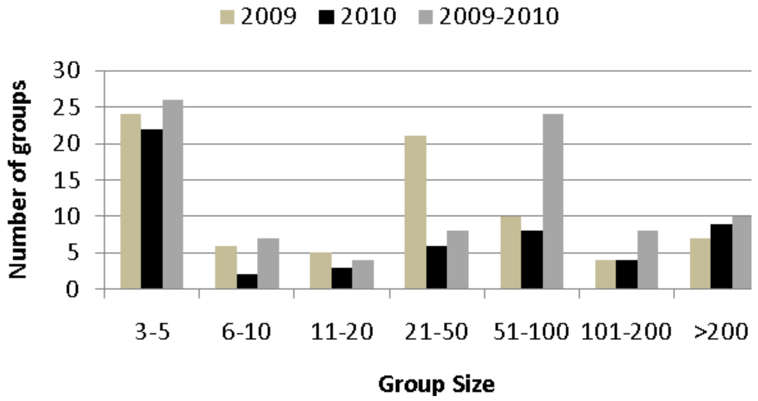

*Figure 5: Number of groups identified by Blondel in 2009, 2010 and in total*

### 5.3 Sizes of Identified Stable Groups

The subsequent analysis concerned the identification of stable groups achieved by the CPM method, whose duration was at least equal to or exceeded a given minimum period length.

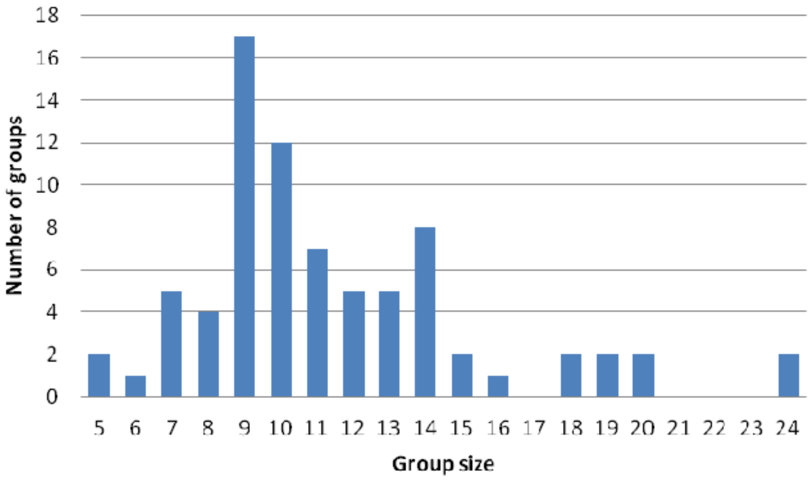

*Figure 6: Numbers of stable social groups of given sizes identified by CPM*

We identified 77 such groups, 173 bloggers were members of these groups, which equals about 2% of the all bloggers 3.78% of bloggers who wrote at least one post and 7.19% of bloggers who wrote at least 5 posts (and for which we could assume that, to some degree, they checked in (appeared) their presence in the blogosphere). One can notice, that the proposed method indicates strongly shaped centres, the bloggers functioning irregularly are not assigned to any stable groups.

Fig. 6 shows the numbers of stable groups having a given number of members. One can notice, that the sizes of stable groups range from 5 to 24 members, with the most frequent size of stable group being 9.

### 5.4 Analysis of Persistent Group Membership and Roles of Users

The next analysis concerned the membership of users in the persistent groups identified in Sec. 5.3. The applied algorithm allows a member to belong to multiple groups in the same or to many groups in different periods. It became apparent that a significant majority of users did not qualify for the stable groups. Those who qualified for stable groups, in a significant majority of cases, belonged to 1, 2 or 3 groups (Fig. 7) – they constituted 58.18% of all bloggers which appeared in stable groups. Only unique bloggers were counted among more than 7 groups, the record-holder belonged to 32 groups. More detailed analysis of the history of activities of this user shows that they actively participated in discussions concerning different subjects (politics, tourism, social issues, and everyday live).

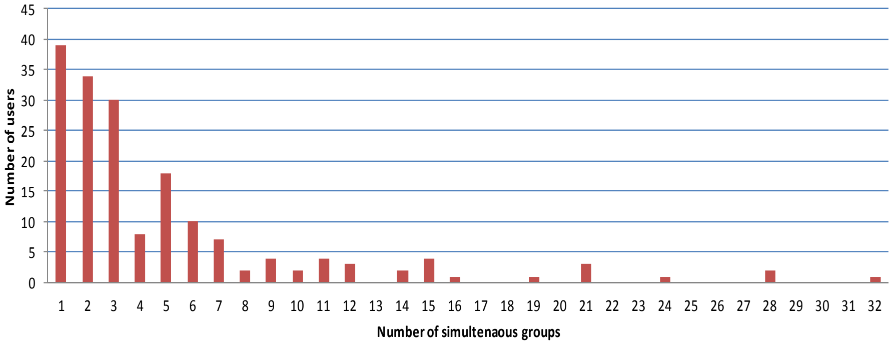

*Figure 7: Number of users which are simultaneously members of a given group number*

The important part of the analysis was the identification of key members in the stable groups and to verify if they played important roles in higher numbers of groups or if being an important member of one group they did not play important roles in other groups (Fig. 8). The analysis showed that the majority of key members played important roles in more than one group and especially a high number of users played important roles in two groups (8 users were active members in two groups). The record holder played the role of core member of 15 groups and was an active member of 7 other groups, it was a journalist, who was writing frequent posts and comments and often was discussing with supporters of different political options.

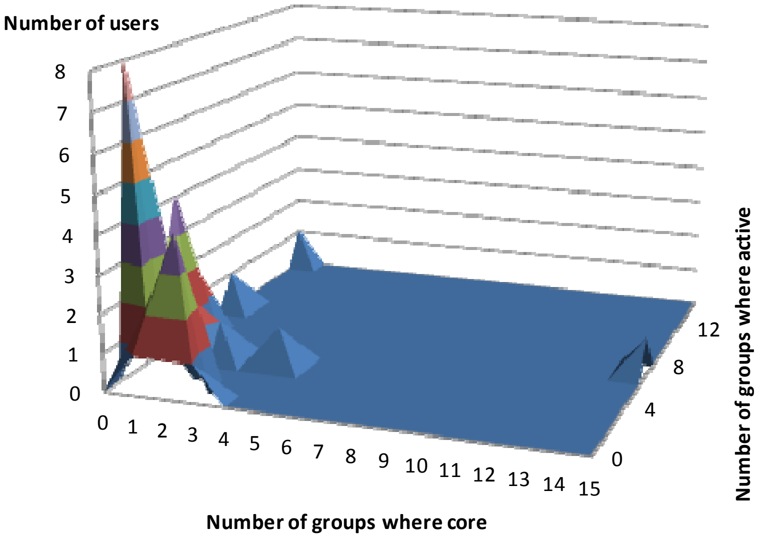

*Figure 8: Number of users in groups with different role configurations*

## 5.5 Analysis of Duration Time of Stable Groups

The lengths of duration for stable groups were calculated for the whole two-year period (Fig. 9). The significant majority of them lasted for 3 periods (3 months) which is the minimum duration necessary to consider the group as a stable group in our analysis. There were 51 such stable groups (which makes 66% of every identified stable group). 20 groups lasted exactly 4 periods and the identified group with the longest duration lasted 8 periods.

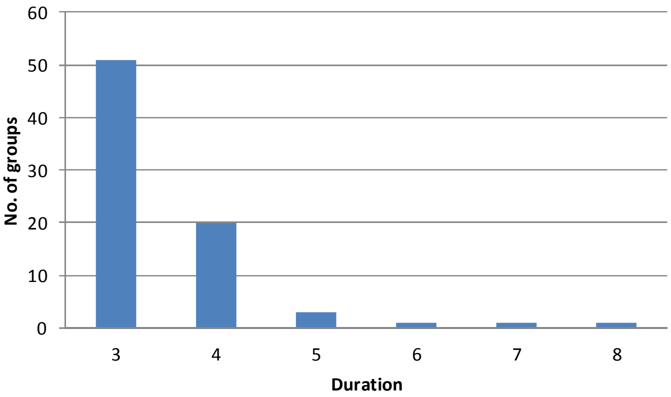

*Figure 9: Number of stable groups having a given time duration*

Durations of stable groups are presented in Fig. 10. The concentration of groups took place especially in the second part of the analysed period. It is associated with a high intensity of activity of bloggers, which was a response to important political events taking place in Poland (23 co-existing groups in April 2010 and 18 co-existing groups in September 2010).

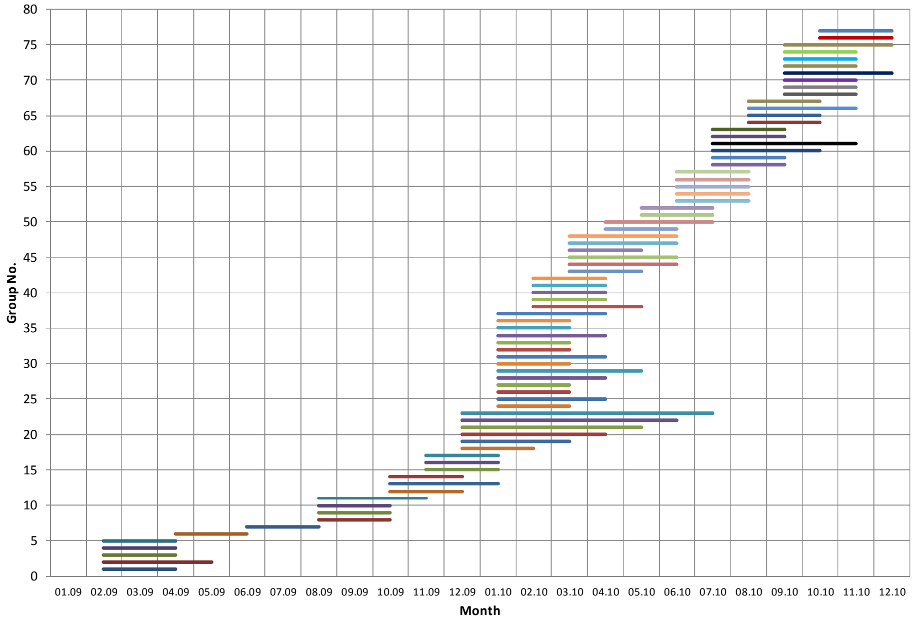

*Figure 10: Durations of stable groups*

### 5.6 Analysis of Group Overlapping

The goal on the next analysis was to find common members for every possible pair of stable groups, i.e. for 77 stable groups, we have 77 x 77 – 77 = 5 852 pairs (the pairs produced by the group with itself were omitted, for such cases the common part equals 100%). We calculated what percentage of the first group membership constitutes the members of the second group in the pair (Fig. 11). It can be observed that almost 60% of pairs of groups do not have any common members, while only 0.24% of pairs overlap in 100%. Besides, 11.6% of pairs overlap within the range (0; 10%], 10.9% pairs – within the range (10%; 20%], and 7.2% pairs – (20%; 30%].

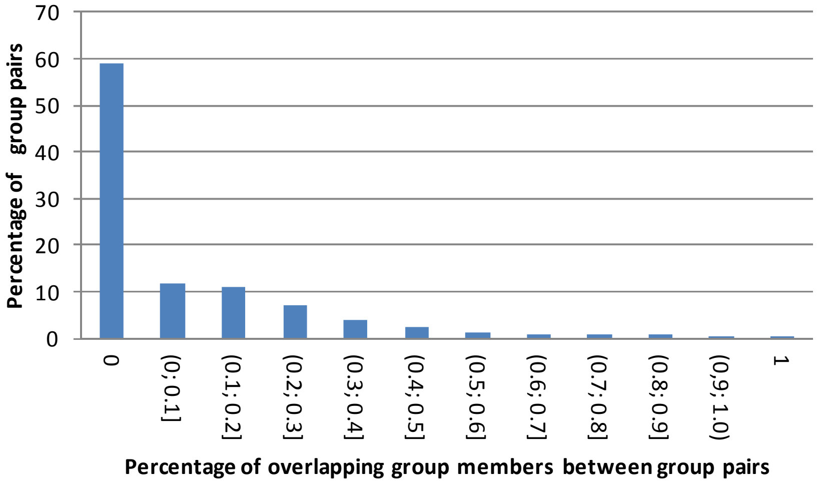

*Figure 11: Overlapping percentage of members for every possible pair of stable groups*

One can distinguish that even though a relatively small number of different bloggers belonged to the stable groups, the obtained group are characterised by a high degree of diversity of their sets of members.

### 5.7 Validation of Key Persons Extracted by IRPG/IRGKM vs. Blondel/Node Position Rankings

While analysing rankings provided by different approaches (see Table 1), one may observe that *core* and *active* users identified by IRPG/IRGKM method, see Sec. 4.1, from persistent groups, in general match very well the ranking obtained by regular clustering and centrality-based approach (Blondel/node position). It means that both approaches provide pretty similar knowledge. On the other hand, these rankings not fully correspond each other and it happens that some users, e.g. core user no. *16688* is only the 196th in his group. Note that such a position is still quite high since group no. 12 is very large community, i.e. it consists of as much as 3,398 members.

| User id | *Core* in group id | *Active* in group id | NP in the network | Rank in the network | NP in the group | Blondel group number | Rank in the group | Blondel group size |
|---|---|---|---|---|---|---|---|---|
| 11210 | 55 | 13,56 | 33.78 | 17 | 28.75 | 10 | 1 | 75 |
| 5192 | | 27,28,29,31 | 33.13 | 18 | 32.90 | 12 | 2 | 3398 |
| 1196 | | 12,13,39,51,52 | 11.93 | 137 | 17.07 | 81 | 2 | 795 |
| 14190 | 69 | 27,28,29 | 17.36 | 65 | 7.84 | 15 | 32 | 2580 |
| 20640 | 72,76 | | 20.67 | 46 | 23.71 | 80 | 1 | 63 |
| 21306 | 39 | 12,13,51 | 0.34 | 5862 | 0.47 | 81 | 316 | 795 |
| 4852 | 62,63 | 5 | 39.41 | 12 | 41.36 | 5 | 1 | 1107 |
| 14056 | | 12,13 | 3.82 | 574 | 8.45 | 81 | 15 | 795 |
| 18774 | | 12,13,39,51,52 | 7.81 | 256 | 12.20 | 81 | 7 | 795 |
| 16688 | 7 | | 6.44 | 327 | 3.62 | 12 | 196 | 3398 |
| 1357 | 10 | 11 | 13.76 | 105 | 12.90 | 12 | 22 | 3398 |
| 5439 | | 5,15,18,19,20,21,22,23,30,31,37,42,56,67 | 36.29 | 15 | 15.43 | 15 | 8 | 2580 |
| 22 | 8,9,12,13,18,30,31,32,33,35,37,40,41,42,69 | 5,19,20,21,22,23,34 | 74.92 | 5 | 77.27 | 77 | 1 | 191 |
| 1765 | | 30,31,71,75 | 10.79 | 162 | 8.93 | 81 | 14 | 795 |
| 17754 | | 72A,76 | 6.01 | 355 | 2.57 | 12 | 275 | 3398 |
| 7170 | | 38 | 13.46 | 110 | 7.85 | 12 | 62 | 3398 |
| 20144 | | 30,31,42,67 | 18.13 | 62 | 14.33 | 81 | 4 | 795 |
| 5951 | 15,17,30,31,42 | 32,37 | 13.34 | 113 | 10.78 | 12 | 34 | 3398 |
| 8759 | | 27,28,29,54,69 | 24.55 | 33 | 12.08 | 81 | 8 | 795 |
| 15686 | | 5 | 19.77 | 50 | 14.80 | 12 | 15 | 3398 |
| 4707 | | 30,31 | 7.591444857 | 271 | 6.169591347 | 81 | 28 | 795 |
| 5709 | | 7,21,22,23,49,50 | 8.61 | 230 | 12.14 | 6 | 1 | 120 |
| 39657 | | 15,17,30,31,39,42 | 4.40 | 503 | 3.41 | 15 | 133 | 2580 |
| 6282 | 15,17 | | 6.68 | 313 | 3.33 | 81 | 42 | 795 |
| 198 | 40 | 41 | 10.20 | 180 | 7.36 | 12 | 73 | 3398 |
| 16182 | | 72,76 | 7.69 | 264 | 3.07 | 12 | 233 | 3398 |
| 2739 | 7,49,50,67 | 21,22,23,66 | 7.61 | 270 | 10.12 | 6 | 2 | 120 |
| 9694 | | 2,26,43,44,45,46,47,48,68 | 7.61 | 269 | 9.89 | 5 | 17 | 1107 |
| 1375 | 55,56 | 54 | 21.51 | 43 | 23.68 | 0 | 1 | 89 |
| 2946 | 30,31,42 | | 10.67 | 166 | 12.27 | 81 | 6 | 795 |
| 1844 | 35,38 | 55 | 50.36 | 8 | 21.31 | 15 | 3 | 2580 |
| 21403 | | 27,28,29 | 7.68 | 266 | 8.24 | 81 | 16 | 795 |
| 13055 | | 62,63 | 43.50 | 11 | 21.04 | 15 | 4 | 2580 |
| 1256 | | 24,25 | 22.31 | 41 | 16.68 | 12 | 10 | 3398 |
| 18169 | | 30,31 | 12.89 | 118 | 10.68 | 81 | 9 | 795 |
| 19989 | 27,28 | 29 | 11.71 | 140 | 6.43 | 15 | 48 | 2580 |

*Table 1: The comparison of rankings (position of individual users) returned separately by (i) IRPG/IRGKM method used for persistent groups, (ii) node position measure applied to the entire network, (iii) node position applied independently only to Blondel groups*

## 6 Conclusion and Future Work

Two separate methods for key person identification and analysis have been presented and considered in the paper: (i) a completely new method based on identification of overlapping groups by means of Clique Percolation Method (CPM), verified towards their stability over time (persistent groups) and identification of user roles within these stable groups (IRPG), see Sec. 3.1 and 4.1, as well as (ii) a typical approach to extraction of social communities - Fast Modularity Optimization (Blondel) with the application of the regular centrality measure to discover key group members, see Sec. 3.2. and 4.2.

Both methods were applied to the data from the Polish blogosphere, see Sec. 5. The experiment results have revealed that most stable groups are similar to each other with respect to their duration, usually 3 months – it refers to over 66% of stable groups, see Fig. 9 and also Fig. 10. Moreover, when some groups disappear, others are established, which shows a more or less stable concept drift in the studied blogosphere, Fig. 10. Besides this, most of groups are animated by a small number of *core* and very *active* users, who however play simultaneously the same role in rather few groups, see Fig. 8.

The presented concept can be extended with sociological interpretation of results, in particular why users are active only for limited periods.

**Acknowledgements**

The work was supported by: The Polish Ministry of Science and Higher Education the research project, 2010-13 and Fellowship co-Financed by European Union within European Social Fund. The publication has been prepared as part of the project of the City of Wrocław, entitled "Green Transfer" - academia-to-business knowledge transfer project co-financed by the European Union under the European Social Fund, under the Operational Programme Human Capital (OP HC): sub-measure 8.2.1.

We thank and confirm the contribution of the students of the Department of Computer Science AGH University of Science and Technology, especially Ł. Krupczak, D. Kiełbasa and A. Mika, who have participated in the development of the system for identification and analysis of stable groups and in gathering the data about blogs.

## References

[Adamic, 05] Adamic L. and Glance N.: The Political Blogosphere and the 2004 U.S. Election: Divided They Blog, Proc. of the 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2005.

[Ahn, 10] Ahn, Y.-Y., Bagrow, J.P. and Lehman, S.: Link Communities Reveal Multiscale Complexity in Networks, Nature, 466, 2010.

[Agarwal, 09] Agarwal, N. and Liu, H.: Modeling and Data Mining in Blogosphere: Morgan & Claypool, 2009.

[Brandes, 06] Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z. and Wagner D.: Maximizing modularity is hard, 2006 http://arxiv.org/abs/physics/0608255.

[Bross, 11] Bross J., Richly K., Kohnen M. and Meinel C.: Christoph Identifying the top-dogs of the blogosphere , Social Network Analysis and Mining, , Volume 1, Number 1, 1-25, Springer, 2011, DOI: 10.1007/s13278-011-0027-7

[Bródka, 09] Bródka, P., Musiał, K. and Kazienko, P.: A Performance of Centrality Calculation in Social Networks. CASoN 2009, IEEE Computer Society, 2009, 24-31.

[Blondel, 08] Blondel, V.D., Guillaume, J., Lambiotte, R. and Lefebvre, E.: Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, 2008, p. P10008.

[Carrington, 05] Carrington, P., Scott, J. and Wasserman, S.: *Models and methods in Social Network Analysis*. Cambrige: University Press, 2005.

[Dolińska, 10] Dolińska I.: Simple Blog Searching Framework Based on Social Network Analysis, Proc. of the Int. Multiconference on Computer Science and Information Technology, vol. 5, Poland, 2010.

[Evans, 09] Evans, T.S. and Lambiotte, R.: Line Graphs, Link Partitions and Overlapping Communities, Phys.Rev.E, 2009.

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

[Fazeen, 11] Fazeen M., Dantu R., and Guturu P.: Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches, Social Network Analysis and Mining, Volume 1, Number 3, 241-254, Springer, 2011, DOI: 10.1007/s13278-011-0017-9

[Fortunato, 10] Fortunato, S.: Community Detection in Graphs, Phys. Rep. 486, 2010.

[Girvan, 02] Girvan, M. and Newman, M.E.J.: Community Structure in Social and Biological Networks, Proc. Natl. Acad. Sci., USA, 2002.

[Jung, 08] Jung, J.J.: Ontology-based Context Synchronization for Ad Hoc Social Collaborations, Knowledge-Based Systems, 21 (7), 2008, 573-580.

[Jung, 09] Jung, J.J.: Contextualized mobile recommendation service based on interactive social network discovered from mobile users, Expert Systems with Applications, 36 (9), 2009, 11950-11956.

[Jung, 10a] Jung, J.J.: Ontology Mapping Composition for Query Transformation on Distributed Environments, Expert Systems with Applications, 37 (12), 2010, 8401-8405.

[Jung, 10b] Jung, J.J.: Integrating Social Networks for Context Fusion in Mobile Service Platforms, Journal of Universal Computer Science, 16 (15), 2010, 2099-2110.

[Jung, 11] Jung, J.J.: Service Chain-based Business Alliance Formation in Service-oriented Architecture, Expert Systems with Applications, 38 (3), 2011, 2206-2211.

[Jung, 12a] Jung, J.J.: Attribute selection-based recommendation framework for short-head user group: An empirical study by MovieLens and IMDB, Expert Systems with Applications, 39 (4), 2012, 4049-4054.

[Jung, 12b] Jung, J.J.: Evolutionary Approach for Semantic-based Query Sampling in Large-scale Information Sources, Information Sciences, 182 (1), 2012, 30-39.

[Kazienko, 07] Kazienko, P. and Musiał, K.: On Utilising Social Networks to Discover Representatives of Human Communities, International Journal of Intelligent Information and Database Systems, 1 (3/4), 2007, 293-310.

[Kazienko, 09] Kazienko, P., Musiał, K. and Zgrzywa, A.: Evaluation of Node Position Based on Email Communication. Control and Cybernetics, 38 (1), 2009, 67-86.

[Koźlak, 10] Koźlak, J., Zygmunt, A. and Nawarecki, E.: Modelling and Analysing Relations Between Entities Using the Multi-agent and Social Network Approaches, MCSS 2010, IEEE International Conference, Kraków, 2010.

[Koźlak, 11] Koźlak, J., Zygmunt, A.: Agent-based Modelling of Social Organisations, CISIS, International Conference on Complex, Intelligent and Software Intensive Systems, Seul, Korea, 2011.

[Lancichinetti, 09] Lancichinetti, A, Fortunato, S. and Kertesz, J.: Detecting the Overlapping and Hierarchical Community Structure in Complex Networks, New Journal of Physics, 11, 2009.

[Macskassy, 11] Macskassy S.A.: Contextual linking behavior of bloggers: leveraging text mining to enable topic-based analysis, Social Network Analysis and Mining, Volume 1, Number 4, 355-375, Springer, 2011, DOI: 10.1007/s13278-011-0026-8

[Newman, 04] Newman M. E. J.: Analysis of Weighted Networks, Physical Review E, 70, 056131, 2004.

[Newman, 04a] Newman, M.E.J. and Girvan, M.: Finding and Evaluating Community Structure in Networks, Physical Review E, 69, 026113, 2004.

[Newman, 10] Newman, M.E.J.: Networks: An Introduction, Oxford University Press, 2010.

[Memon, 08] Memon, N., Larsen, H.L., Hicks, D.L. and Harkiolakis, N.: Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies, Intelligence and Security Informatics, LNCS 5075, Springer, 2008, 477-489.

[Musiał, 09] Musiał, K., Kazienko, P. and Bródka, P.: User Position Measures in Social Networks. SNA-KDD at KDD 2009, ACM Press, Article no. 6, 2009.

[Palla, 05] Palla, G., Abel, D., Derényi, I., Farkas, I., Pollner, P. and Vicsek, T.: K-clique Percolation and Clustering in Directed and Weighted Networks, Bolayai Society Mathematical Studies, 2005.

[Palla, 05a] Palla, G., Derényi, I., Farkas, I. and Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society, Nature 435, 814–818, 2005.

[Palla, 09] Palla, G., Pollner, P., Barabasi, A.-L. and Vicsek, T.: Social Group Dynamics in Networks, in Adaptive Networks, ed. T. Gross, H. Sayama, Springer Berlin/Heidelberg, 2009.

[Porter, 09] Porter, M.A., Onnela, J.-P. and Mucha, J.: Communities in Networks, Notices of the American Mathematical Society, Vol. 56, No. 9, 2009.

[Tang, 09] Tang, J., Sun, J., Wang, C. and Yang, Z.: Social Influence Analysis in Large-scale Networks, KDD '09, The 15th ACM SIGKDD Int. Conference on Knowledge discovery and Data Mining, ACM, 2009.

[Tang, 10b] Tang, L. and Liu, H.: Community Detection and Data Mining in Social Media, Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan&Claypool Publishers, 2010.

[Tang, 10] Tang, L. and Liu, H.: Graph Mining Applications to Social Network Analysis, Managing and Mining Graph Data, ed. C. Aggarwal, X. Wang, 2010.

[Tang, 10a] Tang, L. and Liu H.: Learning with Large-scale Social Media Network, PhD thesis, Arizona State University, 2010.

[Wasserman, 94] Wasserman, S. and Faust K.: Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.

[Yang, 11] Yang, Y., Sun, Y., Pandit, S., Chawla, N.V. and Han, J.: Is Objective Function the Silver Bullet?, Int. Conf. on Advances in Social Networks Analysis and Mining, Taiwan, 2011.

[Yano, 10] Yano T. and Smith N.S.: What's Worthy of Comment? Content and Comment Volume in Political Blogs, Proc. of the Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[Ye, 11] Ye, Q., Wu, Q., Zhao, Z. and Wang, B.: Detecting Link Communities in Massive Networks, Int, Conf. on Advances in Social Networks Analysis and Mining, Taiwan, 2011.

[Zygmunt, 10] Zygmunt, A., Koźlak, J. and Krupczak, Ł: Identifying the Influential Individuals in Blogosphere, Studia Informatica, vol. 31, no 2A(89), 2010 (in Polish).

[Zygmunt, 11] Zygmunt, A., Bródka, P., Kazienko, P. and Koźlak J.: Different Approaches to Groups and Key Person Identification in Blogosphere. ASONAM 2011, The 2011 International Conference on Advances in Social Network Analysis and Mining, IEEE Computer Society, 2011, 593-598.