A Method for Group Extraction in Complex Social Networks

Piotr Bródka[1], Katarzyna Musial[2], Przemysław Kazienko[1]

[1] Institute of Informatics, Wrocław University of Technology
Wyb.Wyspiańskiego 27, 50-370 Wrocław, Poland
[2] School of Design, Engineering & Computing, Bournemouth University, Poole, Dorset,
BH12 5BB, United Kingdom
piotr.brodka@pwr.wroc.pl, kmusial@bournemouth.ac.uk, kazienko@pwr.wroc.pl

Abstract. The extraction of social groups from social networks existing among employees in the company, its customers or users of various computer systems became one of the research areas of growing importance. Once we have discovered the groups, we can utilise them, in different kinds of recommender systems or in the analysis of the team structure and communication within a given population.

The shortcomings of the existing methods for community discovery and lack of their applicability in multi-layered social networks were the inspiration to create a new group extraction method in complex multi-layered social networks. The main idea that stands behind this new concept is to utilise the modified version of a measure called by authors multi-layered clustering coefficient.

Keywords: multi-layered social network, groups discovery in social network, multi-layered clustering coefficient, social network analysis

1 Introduction

In the recent few decades, the area of complex networks has attracted more and more scientists from different research fields. All complex networked systems have some common features such as: (i) skewed distribution of connections, (ii) small degree of separation between vertices, (iii) high clustering rate, (iv) non-trivial temporal evolution and last but not the least (v) presence of motifs, hierarchies and communities [3], [10]. The feature that is investigated by authors in this paper is the last enumerated one, i.e. the existence of communities whereas the subset of complex systems that is analysed are social networks.

A social network (SN) is one of the type of complex networks in which nodes are people (social entities) and the edges denote the relationships between various people [13]. The concept of SN, first coined in 1954 by J. A. Barnes [1], has been in a field of study of modern sociology, geography, social psychology, organizational studies and computer science for the last few decades. Social networks and social network analysis supported by computer science provide the opportunity to expand other branches of knowledge. Lately, we have experienced the rapid growth of social structures supported by communication technologies and the variety of Internet- and Web-based services. This article focuses on discovering the communities within the complex multi-layered social networks (CMSN) extracted from different systems based on communication technologies, in which users interact or cooperate with each other by means of various dedicated services. The main characteristic of CMSN is that they consist of many layers, corresponding to different kinds of relationship.

The extraction and analysis of groups in social networks is not a new concept and as presented in Related Work section has been investigated by many scientists. However, none of the research addresses this issue for networks where more than one type of relationship exists and this is a goal of the presented research.

The next, second section of the article includes the description of the most commonly utilised methods to group extraction. In section 3, the concept of multi-layered social network is presented. After that the new method and its characteristics together with preliminary experiments are described. Finally, the conclusions and future work in the area of group extraction are depicted.

2 Related Work

Many approaches to the problem of community identification in social networks which consist of one type of connections have been developed. The existing methods origin both from the graph theory and data mining techniques. In the former, the notion of a group is formalised by the general property of cohesion among community members and the evaluation of this feature determines whether the set of people can be seen as a group or not. The assessment of cohesion can be made based on the complete mutuality, reachability, diameter and nodal degree [13]. Other methods that are used in extracting communities are fuzzy clustering approaches and specifically clique percolation methods that will allow the groups to overlap [8], [7].

Girvan and Newman analyzed a network of scientific collaboration [4]. Scientific collaboration is associated to co-authorship: two scientists are connected if they have written at least one paper together. The authors invented and used a new method on a collaboration network of scientists working at the Santa Fe Institute. Obtained groups reflect research divisions at the Santa Fe Institute. The community structure of scientific collaboration networks has been investigated by many other authors. Radicchi et al analyzed network of scientific collaborations and network of college football teams to test improved version of Girvan and Newman method [9]. Some other types of collaboration networks have been studied as well. Gleiser and Danon investigated a collaboration network of jazz musicians [5]. Musicians are connected if they have played in the same band. Extracted, by Girvan and Newman method, communities reflect both racial segregation and geographical separation.

Tyler et al. also used modified version of Girvan and Newman algorithm to study a network of e-mail exchanges between workers of HP Labs [12]. The method enables to measure the degree of membership of each member within a community and allows communities to overlap. The extracted groups matched quite closely the organization of the Labs in departments and project groups. The same method have been used to find communities of related genes [15].

Blondel et al. developed a fast hierarchical modularity optimization technique and used it to analyze a huge network of mobile phone communications between 2.6 million users of a Belgian phone operator [2]. The group extraction and analysis, performed, reveals six hierarchical levels. The 1st level consists of 261 groups with more than 100 vertices. Users are split in two main groups which reflect the linguistic division of Belgian population.

Traud et al. used data from Facebook and created a network of friendships between students of different American universities; students were connected if they were friends on Facebook [11]. Newman's spectral optimization of modularity was used to detect the communities. The results were compared to demographic information on the students' populations, one finds that communities are organized by class year or by House (dormitory).

However, the domain analysis has indicated that there is no group extraction method dedicated to the multi-layered social networks. Moreover, there is lack of research that would point out whether the existing methods for community discovery in regular social networks can be utilized in multi-layer social networks.

3 Complex Multi-layered Social Networks

The structure that will be analysed in this research is *a network*, i.e. set of interconnected nodes. In this paper, the networks extracted from different systems based on communication technologies will be investigated. The units (nodes) in such CMSNs are digital representations of people who use email services, telecommunication systems, multimedia sharing systems, access blogosphere etc. The node is also called identifier (*id*). Based on interactions between users their mutual relationships are extracted and in the next step the communities can be identified. Due to diversity of communication channels the analyzed networks are *multi-layered,* i.e. they consist of more than one type of relationship. Different relations can emerge from different communication channels, i.e. based on each communication channel separate relation that can be also called a layer of a network is created. These various relations between two users can be grouped into tie. The concept of multi-layered complex social network is presented in Fig. 1.

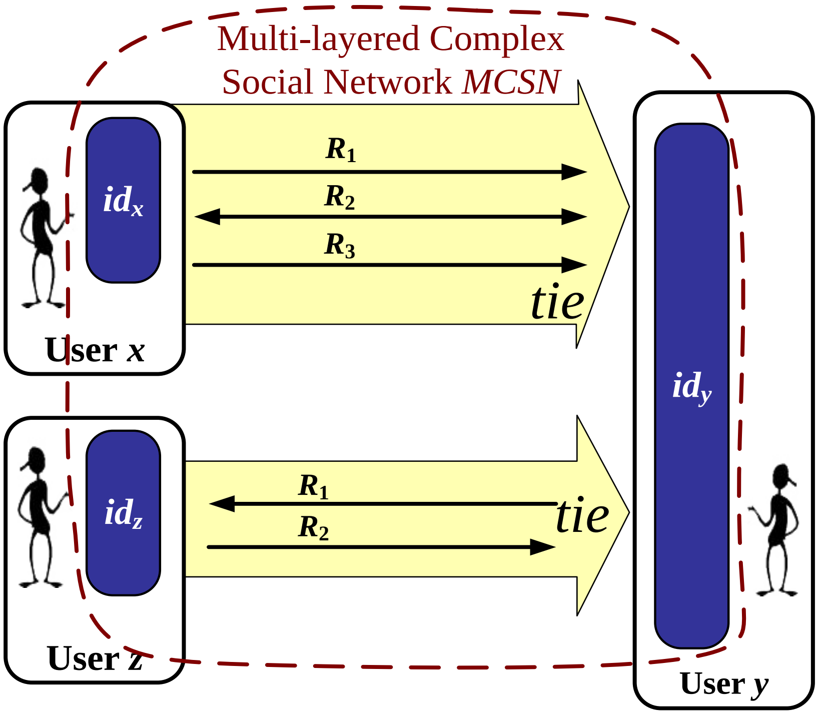

Fig. 1. The concept of multi-layered complex social network extracted from systems where users are represented by their digital identities

4 Multi-layered Clustering Coefficient in the Complex Multi-layered Social Network

Before the general concept of the method can be presented, three new measures need to be described: local clustering coefficient tailored for the needs of CMSN – multi-layered clustering coefficient (*MCC*) and two measures that respect either extended nearest neighbourhoods (*MCCEN*) or reduced nearest neighbourhoods (*MCCRN*). All these concepts were developed to be utilized in the group extraction within CMSN, see Sec. 5. Their detailed description is presented in this section.

4.1 Local Clustering Coefficient in the 1-layered Social Network

The regular local clustering coefficient is a measure of degree to which nodes in the network structure tend to cluster together. It quantifies how close the node's neighbours are to being a fully connected graph. Local clustering coefficient was introduced by Duncan J. Watts and Steven Strogatz who utilized it in order to determine whether a graph is a small-world network [14].

The local clustering coefficient $CC_l(x)$ for node x in the network that contains single layer l in which the relationships are weighted and directed, is calculated as follows:

$$CC_l(x) = \frac{\sum_{y \in N_l(x)} \left(in_l\left(y, N_l(x)\right) + out_l\left(y, N_l(x)\right) \right)}{2 \cdot card(N_l(x))}, \tag{1}$$

where:

$N_l(x)$ – the 1–level neighbourhood of node x (the set of nearest neighbours) in the network containing one layer l,

$in^l(y,N_l(x))$ – the weighted in-degree of a node y that belongs to the 1-level neighbourhood of x in the network containing one layer l,

$out_l(y,N_l(x))$ – the weighted out-degree of a node y that belongs to the 1-level neighbourhood of x in the network containing one layer l.

The weighted in-degree $in_l(y,N(x))$ for a given node y in the network containing one layer l is the sum of all weights $w(z,y)$ of edges $z \rightarrow y$ from network containing one layer l that income to a given node y from other nodes z from the neighbourhood $N_l(x)$.

$$in_l\left(y, N_l(x)\right) = \sum_{z \in N_l(x)} w(z, y), \tag{2}$$

Similarly, the weighted out-degree $out_l(y)$ for a given node y is the sum of all weights $w(y,z)$ of the outgoing edges $y \rightarrow z$ that come from x's neighbours z.

Note that if the sum of weights of outgoing edges for a given node is always 1, then $CC_l(x)$ is from the range (0;1]. It reaches 1, if the neighbours $y \in N_l(x)$ have outgoing relationships only towards other nodes $z \in N_l(x)$. It means that there does not exist edges $y \rightarrow x$ or $y \rightarrow k$, where k is a node outside $N_l(x)$.

4.2 Multi-layered Local Clustering Coefficient for Complex Multi-layered Social Networks

A measure called the multi-layered local clustering coefficient $MCC(x)$ needs to be introduced as not the simple one-layered social networks but the multi-layered social networks are investigated in this paper. The measure $MCC(x)$ in the multi-layered environment, is the average of clustering coefficients $CC_l(x)$, see Eq. 1, from all component layers l. Thus, each layer is treated here as a separate network containing one, lth layer. The value of $MCC(x)$ is computed in the following way:

$$MCC(x) = \frac{\sum_{l \in L} CC_l(x)}{card(L)}, \tag{3}$$

where: l is the index of the lth layer from the set L of all layers in a given CMSN.

Note that $CC_l(x)$ is calculated separately for each layer l, i.e. it takes into account only the neighbourhood of x that exists only in layer l.

4.3 Multi-layered Clustering Coefficient in the Extended Neighbourhood

The second measure – multi-layered clustering coefficient in the extended neighbourhood (*MCCEN*) is in its concept similar to the multi-layered clustering coefficient *MCC* but the neighbourhoods used by *MCCEN* are defined differently. The measure *MCCEN*(x) denotes, to which extent the nearest neighbourhoods existing in different layers for a given user x overlap each other. First, we need to define the extended, multi-layered neighbourhood *EN*(x) for a given node x in the set L of layers. It is a union of neighbourhoods from all layers:

$$EN(x) = \bigcup_{l \in L} N_l(x)\,, \tag{4}$$

where: $N_l(x)$ – the nearest neighbours of node x in the layer l.

The multi-layered clustering coefficient in the extended neighbourhood *MCCEN*(x) for a given node x, respects the extended neighbourhoods *EN*(x) instead of neighbourhoods from only one layer $N_l(x)$, compare Eq. 1, as follows:

$$MCCEN(x) = \frac{\sum\limits_{l \in L} \sum\limits_{y \in EN(x)} \left(in_l\left(y, EN(x)\right) + out_l\left(y, EN(x)\right) \right)}{2 \cdot card(EN(x)) \cdot card(L)}\,. \tag{5}$$

4.4 Multi-layered Clustering Coefficient in the Reduced Neighbourhood

Similarly to multi-layered clustering coefficient in the extended neighbourhood (*MCCEN*), we can define multi-layered clustering coefficient in the reduced neighbourhood (*MCCRN*). It takes into account yet another reduced neighbourhood *RN*(x) of a given node x, i.e. the set of neighbours who occur in every layer. Hence, the reduced neighbourhood *RN*(x) is an intersection of sets – neighbourhoods from all layers:

$$RN(x) = \bigcap_{l \in L} N_l(x)\,. \tag{6}$$

The multi-layered clustering coefficient in the reduced neighbourhood *MCCRN*(x) for a given node x, is computed in the following way:

$$MCCRN(x) = \frac{\sum\limits_{l \in L} \sum\limits_{y \in EN(x)} \left(in_l\left(y, RN(x)\right) + out_l\left(y, RN(x)\right) \right)}{2 \cdot card(RN(x)) \cdot card(L)}\,. \tag{7}$$

5 A New Method for Group Extraction in the Complex Multi-layered Social Network based on Multi-layered Clustering Coefficients

5.1 Why Do We Need Yet Another Method for Group Extraction

The concept of group and methods for its extraction is not fully addressed even in the networks where only one type of relationship exists as the definition of a term ‘group’ in the literature is inconsistent and sometimes even researchers use a group concept without giving it a precise formal definition [13]. This task is even more challenging in CMSN as it is very hard to establish which types of relationships determine that a set of users and their connections can be called ‘a group’. Thus, the main contribution of the research within this topic to the field of community extraction and analysis would be to create a definition of ‘a group’ within the multi-layered environment and to develop the methods that enable to extract these groups from the gathered data about users and their interactions. On the other hand, the groups can be discovered and investigated in each of the layer separately (each layer is created based on one type of relation).

5.2 General Concept of Method for Group Extraction

There are two general approaches to extract groups in multi-layered environment: (i) extracting groups in each layer separately and then merge the communities throughout the layers, (ii) first flatten the network into one layer and then discover groups within it. In the new method proposed in this article, the communities are extracted from the network as a whole not from each layer separately.

The general concept of clustering is selection of “strong nodes”, i.e. nodes with strong enough multi-layered clustering coefficients, create a group from their neighbourhoods and join to this group the neighbourhoods of “strong neighbours”. The process proceeds as follow:

STEP 1: Calculate all three multi-layered clustering coefficients for multi-layered social network, namely, $MCC(x)$, $MCCEN(x)$, and $MCCRN(x)$ separately for each node x in the network, see Section 4.

STEP 2: Extract set A of “strong nodes x”, for which all three multi-layered clustering coefficients exceed appropriate thresholds: α, β, γ, i.e. $MCC(x)>\alpha$, $MCCEN(x)>\beta$, and $MCCRN(x)>\gamma$. For all nodes in A, identify their 1-level extended neighbourhoods $EN(x)$, see Eq. 4. Neighbourhoods $EN(x)$ are achieved as by-products while computing $MCCEN(x)$, see Sec. 4.3. Initialize an empty set B of already processed nodes.

STEP 3: The entire ith group is created within this step, starting from the extended neighbourhoods of a strong node from set A, next going throughout the neighbourhoods of the neighbours.

a) Create a new empty group $G_i = \varnothing$ and initialize set T_i with node x taken from set A: $T_i=\{x\}$. T_i is a set of strong and leading nodes to be processed for G_i.

b) Take the first/next node y from T_i and fill group G_i with the extended neighbourhood of y: $G_i=G_i \cup EN(y)$ and add node y to already processed nodes: $B=B\cup\{y\}$. Remove node y from set A: $A=A\backslash\{y\}$. At the first run $y=x$.

c) Identify not yet processed leading nodes z within $EN(y)$ using the following criteria: $MCC(x)>\alpha'$, $MCCEN(x)>\beta'$, $MCCRN(x)>\gamma'$, $\alpha'\leq\alpha$, $\beta'\leq\beta$ and $\gamma'\leq\gamma$. Create set S from these nodes.

d) Remove from S all already processed nodes: $S=S\backslash B$ and add new set S to T_i for further processing: $T_i=T_i\cup S$. The neighbourhoods of S members will be joint to G_i.

e) Remove the just processed node y from T_i: $T_i=T_i\backslash\{y\}$.

f) Go to step 3b unless set T_i is empty.

STEP 4: Repeat entire step 3 until set A of "strong nodes" for processing is empty.

STEP 5: Create a separate group of outliers from all nodes that do not belong to any group G_i.

Note that a new group is created with each iteration of step 3. Finally, we achieve as many groups as many times step 3 is invoked, plus eventually one group of outliers created in step 5.

6 Experiments

The main goal of the experiments was to investigate the characteristics of proposed in the paper metrics that serve to assess the extent to which the neighbourhoods of user are clustered within the complex multi-layered social networks.

The experiments were performed on 1000 users from the Flickr system where eleven different layers have been identified. These layers include: tags used by more than one user R^t, user groups R^g, photos added by users to their favourites R^{fa}, R^{af}, R^{ff}, opinions about photos created by users R^{oa}, R^{ao}, R^{oo}, and the relations derived from the contact lists R^c, R^{ac}, R^{cac}. The detailed information about the data set can be found in [6].

The outcomes of the experiments have shown that the value of *MCC* coefficient varies from 0 to 0.53, the value of *MCCEN* is from the range 0 to 0.61 and the value of *MCCRN* from 0 to 0.19. There were only 7 users whose *MCCRN* was greater than 0. All the results are presented in Fig. 2, 3, and 4.

It can be noticed that the multi-layered clustering coefficient for reduced neighbourhood equals 0 for most users and it means that there are only few users who have similar neighbourhoods on all layers. This implicates that a user maintains the relationships with different neighbourhood on different layers. This is also confirmed by the relatively high *MCCEN* value (average is 0.48).

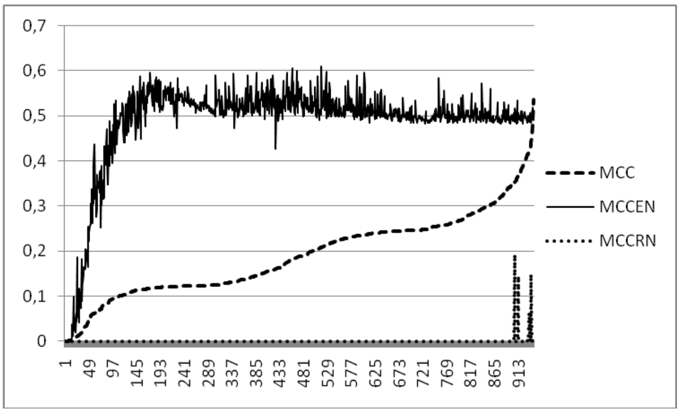

Fig. 2. Distribution of multi-layered clustering coefficients; order by value of *MCC*

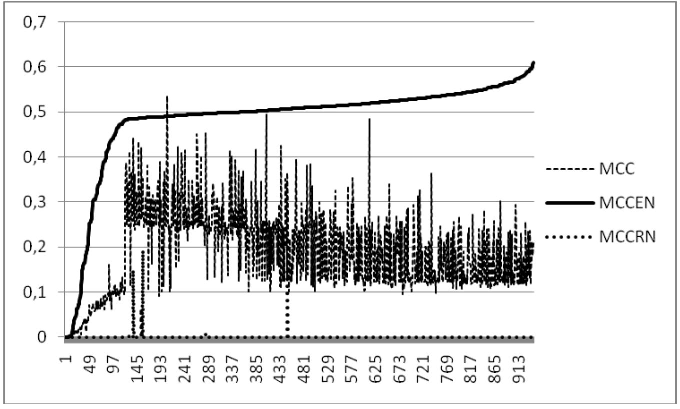

Fig. 3. Distribution of multi-layered clustering coefficients ; order by values of *MCCEN*

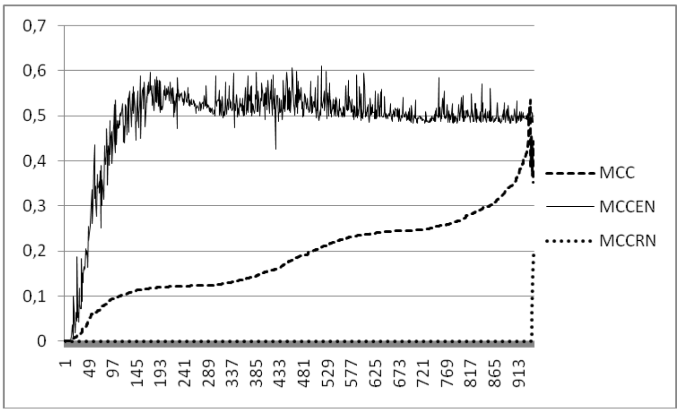

Fig. 4. Distribution of multi-layered clustering coefficients; order by values of *MCCRN* and *MCC*

7 Conclusions

The proposed method of group identification facilitates to extract groups in the social network that contains more than one type of relationship. The main contribution of this paper is to develop some new measures that enable to investigate the user neighbourhood in the complex multi-layered social networks. Additionally, the methods to create and evaluate the clustering coefficient in both extended and reduced neighbourhood were proposed. The preliminary experiments were performed on the Flickr system. They have revealed that the multi-layered clustering coefficient in the extended neighbourhood *MCCEN* is relatively high and reaches in average the level of 0.61, whereas the multi-layered clustering coefficient in reduced neighbourhood *MCCRN* equals 0 for most users. The values of multi-layered local clustering coefficient *MCC* vary from 0 to 0.53.

The interesting extension of current work will be to investigate the influence of these coefficients on the size and structure of the groups discovered. Future work will also focus on the comparison of the communities extracted using the proposed method and the investigation of correlation between these groups. This will enable to define which types of relations are utilized in the process of group formation and based on which of them the more or less sustainable groups emerge.

Acknowledgments. This work has been supported by the Polish Ministry of Science and Higher Education, the development project, 2009-11.

References

1 Barnes J. A., Class and Committees in a Norwegian Island Parish, Human Relations 7, 1954, 39–58.

2 Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E., Fast unfolding of communities in large networks, J. Stat. Mech. P10008 (2008).

3 Caldarelli, G., Vespignani, A. (eds.): Large Scale Structure and Dynamics of Complex Networks, From Information Technology to Finance and Natural Science, Complex Systems and Interdisciplinary Science, vol. 2, World Scientific Publishing Co. Pte. Ltd., 2007.

4 Girvan M., Newman M.E.J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.

5 Gleiser P., Danon L., Community structure in jazz, Adv. Complex Syst. 6 (2003) 565.

6 Kazienko P., Musial K., Kajdanowicz T.: Multidimensional Social Network and Its Application to the Social Recommender System. IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans, 2010, in press.

7 Palla G., Barabasi A.-L., Vicsek T., Quantifying social group evolution, Nature, vol. 446, 2007, 664–667.

8 Palla G., Derenyi I., Farkas I., Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society, Nature, 435, 2005, 814818.

9 Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D.: Defining and identifying communities in networks, PNAS, vol. 101 , 2 March 2004, 2658–2663

10 Strogatz S. H. Exploring complex networks, Nature, 410(6825), March 1998, 268–276.

11 Traud A.L., Kelsic E.D., Mucha P.J., Porter M.A., Community structure in online collegiate social networks, eprint arXiv:0809.0690.

12 Tyler J.R., Wilkinson D.M., Huberman B.A., Email as spectroscopy: Automated discovery of community structure within organizations, in: Communities and Technologies, Kluwer, B.V., Deventer, The Netherlands, 2003, 81–96.

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

14 Watts D.J., Strogatz S., Collective dynamics of 'small-world' networks. Nature 393, June 1998, 440–444.

15 Wilkinson D.M., Huberman B.A., A method for finding communities of related genes, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 5241–5248.