Identifying Overlapping Communities in Directed Networks Via Triangles

Qingyu Zou, Fu Liu*, Tao Hou, Yihan Jiang
College of Communication Engineering, Jilin University, Changchun 130000, Jilin, China
Corresponding author, e-mail: qyzou11@mails.jlu.edu.cn, liufu@jlu.edu.cn

Abstract

A lot of complex systems in nature and society can be represented as the form of network. The small-scale subnets topological features are vital to understand the dynamics and function of the networks. Triangles comprised of three nodes are the simplest subnet in the network. Based on the triangle distribution of the complex network, we present a novel approach to detect overlapping community structure in directed networks. Different from previous studies focused on grouping nodes, our method defines communities as groups of links rather than nodes so that nodes naturally belong to more than one community. It can identify a suitable number of overlapping communities without any prior knowledge about the community. We evaluated our approach on several real-networks. Experimental results prove that the algorithm proposed is efficient for detecting overlapping communities in directed networks.

Keywords*: directed network, triangles, overlapping communities, link similarity, partition density*

1. Introduction

In order to shed light on the structure, dynamic and robust of complex systems in nature and society they have been represented as networks, in which nodes symbolize the components of system and links connecting nodes denote the relationships between them [1-3]. Community structure is one of the most important feature of many complex networks [4-6], with which can reveal topological relationships between system elements and represent function [7, 8]. Therefore detecting the community structure in the network has been attracted much attention in recent years [9-15].

The clustering algorithm is a class of pattern recognition method widely used in many fields [16-18]. By now there are mainly two kinds of clustering algorithms have been proposed to detect communities in complex networks, one is optimization algorithm, and the other is the hierarchical clustering method [19]. One approach of the first scheme is based on a measure called betweenness. It calculates one of several measures [20, 21] of the flow of traffic across the links of a network and then removes the most traffic links from the network. Two other related algorithms used to identify links for removal are fluid-flow and current-flow analogies [22]. A different class of optimization algorithms is the methods based on information-theoretic ideas, such as the minimum description length methods of Rosvall and Bergstrom [23]. The basic idea is to define a quantity that is high for good divisions of a network and low for bad ones, and then search the division with the highest score through all the possible cases. Various different measures for calculating scores have been proposed, such as the likelihood-based measures and others [24], but the most widely used measure is the modularity [15]. The hierarchical clustering algorithms include agglomerative and divisive methods to find community structure in networks. They first compute the strength of link between each pair nodes based on different properties, such as link betweenness [25], link clustering coefficient [26], information centrality [27], similarity based on random walks [28], clustering centrality [29], and so on. Then, merging the two nodes with the highest strength of link repeatedly (agglomerative method), or removing the link with the lowest strength repeatedly (divisive methods), the partition results of the networks are obtained. Nearly all of these methods are based on the properties of nodes and assumed each node belongs to only one community. Yong-Yeol Ahn et al [30] and T. S. Evans et al [31] reinvent communities as groups of links in undirected networks and show that the quality of a link partition can be evaluated by the modularity of its corresponding line graph. However, many of the networks that we would like to study are directed, and a node may belong

Received April 15, 2013; Revised July 1, 2013; Accepted July 15, 2013

to several communities, including the World Wide Web, food webs, many biological networks, and even some social networks. The commonest approach to detecting communities in directed networks has been simply to ignore the link directions and apply algorithms designed for undirected networks [32]. It is clear that we are throwing away a good deal of information about our network's structure information that could allow us to make a more accurate determination of the communities if discarding the directions of links.

In this paper, a new algorithm based on triangle distribution is proposed to detect the overlapping community structure in directed networks. We consider a community to be a set of closely interrelated links rather than a set of nodes with many links between them. Then, using hierarchical clustering with a similarity between links to build a dendrogram where each leaf is a link from the original network and branches represent link communities. The link dendrogram provides a rich hierarchy of structure, but to obtain the most relevant communities it is necessary to determine the best level at which to cut the tree. For this purpose, we introduce a new partition density based on link density inside communities. Computing partition density at each level of the link dendrogram allows us to pick the best level to cut. We compared the performance of our algorithm with three successful methods: clique percolation [33], link partition [31], and modularity spectral optimization [34] with three real-networks including Gene network, Email network and Metabolic gene network. Clique percolation is the most prominent overlapping communities identifying algorithm in undirected networks, link partition is the first detecting overlapping communities algorithm based on link property and modularity maximization can be generalized in a principled fashion to incorporate information contained in link directions. The application to real-networks show that our method works effectively in detecting overlapping communities in directed networks.

2. Research Method

2.1. Community

A community consisted of nodes and links between these nodes is part of the network with a few ties with the rest of the system. Although no common definition has been agreed upon, it is widely accepted that a community should have more internal than external connections [21, 35]. The nodes in the same community often have common properties and densely interconnected compared with the rest of the network. It is noted that two communities may overlap each other while a node can connect with different communities simultaneously [4]. In Figure 1, an example of a directed network with communities is shown. There are three communities in this network, denoted by circle, square, pentagon and triangle, respectively. Node of pentagon is a common node since it should belong to the circle community as well as the triangle community.

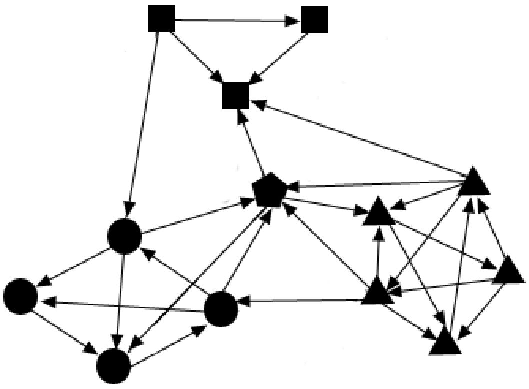

Figure 1. Example network showing community structure. The nodes of this network are divided into three groups, node pentagon is the common node of both the circle and triangle communities

2.2. Triangle Vertexes Weightiness

In general, the simple building blocks of complex networks is not a link but a small structure of several nodes called motif [36]. Network motifs are small subgraphs that can be found in a network statistically significantly more often than in randomized networks. Among the possible motifs, the simplest one is the triangle which represents the basic unit of transitivity and redundancy in a graph, see Figure 2.

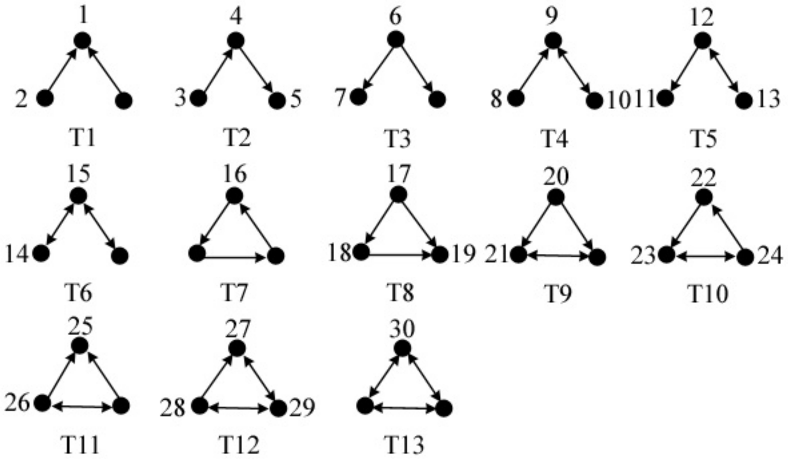

Figure 2. List of all 13 types of triangles

As shown in Figure 2, there are 13 triangle cases at most, including 39 vertexes, in an arbitrary directed network. We compare all three vertexes one another for each triangle T_i and merge the code of vertexes had the same place. Then, there are 30 special vertexes for triangles, encoded from 1 to 30 in Figure 2. We assign different weights w_i to different vertexes i, because some complex triangles contain the simple triangles, such as triangle 11 contain triangle 1. We assign higher weights to the vertexes whose are not affected by other vertexes, and lower weights to depend on other vertexes. The w_i is calculated using a function as follows:

$$w_i = \frac{TC_i}{\max(TC_i)} \tag{1}$$

where TC_i means the number of vertexes affected by vertex *i*. We consider that each vertex affects itself. For instance, for vertexes 1, TC_1=2, since it affects vertexes 25 and itself; similarly, TC_6=3, since vertex 6 affected vertexes 17, 20 and itself. The weights of 30 vertexes as shown in Table 1.

Table 1. The weights of 30 vertexes

Vertex	Weight	Vertex	Weight	Vertex	Weight	Vertex	Weight	Vertex	Weight	Vertex	Weight
No.1	0.5	No.6	0.75	No.11	1	No.16	0.25	No.21	0.25	No.26	0.25
No.2	0.5	No.7	1	No.12	1	No.17	0.25	No.22	0.25	No.27	0.25
No.3	0.75	No.8	0.75	No.13	1	No.18	0.25	No.23	0.25	No.28	0.25
No.4	0.75	No.9	1	No.14	1	No.19	0.25	No.24	0.25	No.29	0.25
No.5	0.75	No.10	0.75	No.15	0.75	No.20	0.25	No.25	0.25	No.30	0.25

2.3. Triangle Degree

The number of triangles that the node touches is triangle degree of it. For a node *u*, as shown Figure 3, the triangle degree values of the 30 vertexes of node *u*, are presented in the Table 2.

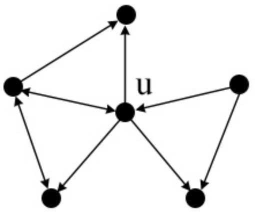

Figure 3. An illustration of a small network containing node *u*. The values of the 30 vertexes of the triangle degree are shown in Table 2

Table 2. Values of the 30 vertexes of the triangle degree of node *u* in Fig.3

Vertex	Degree	Vertex	Degree	Vertex	Degree	Vertex	Degree	Vertex	Degree	Vertex	Degree
No.1	0	No.6	3	No.11	0	No.16	0	No.21	0	No.26	1
No.2	2	No.7	1	No.12	2	No.17	0	No.22	0	No.27	0
No.3	0	No.8	2	No.13	0	No.18	0	No.23	0	No.28	1
No.4	3	No.9	1	No.14	1	No.19	0	No.24	0	No.29	0
No.5	0	No.10	3	No.15	0	No.20	0	No.25	0	No.30	0

2.4. Link Similarity

The link similarity is a measure of closeness between a pair of links. We limit ourselves to only connected pairs of links (i.e. sharing a node) since it is unlikely that a pair of disjoint links are more similar to each other than a pair of links that share a node; at the same time this choice is much more efficient. For a connected pair of links e_{ik} and e_{jk}, we call the shared node *k* a share node and *i* and *j* equal nodes. The similarity between links is defined as:

$$S(e_{ik}, e_{jk}) = \frac{\sum_{l=0}^{30} D_l(i,j)}{\sum_{l=0}^{30} w_l} \tag{2}$$

where $S(e_{ik},e_{jk})$ means link similarity value, $D_l(i,j)$ is distance between the vertex *l* of nodes *i* and *j* as:

$$D_l(i,j) = w_l \times \frac{\left|n_l(i) \cap n_l(j)\right|}{\left|n_l(i) \cup n_l(j)\right|} \tag{3}$$

where $n_l(i)$ is the triangle degree of vertex *l* of node *i*, $n_l(i) \cap n_l(j)$ means the number of *l* triangles shared by node *i* and *j*, $n_l(i) \cap n_l(j)$ means the number of all *l* triangles connected node *i* and *j*.

2.5. Hierarchical Clustering

For a given network, we calculate the similarities for all connected link-pairs at first, and then use average-linkage hierarchical clustering [37] to find hierarchical community structure. The finding processes are described in the following three steps.

Stage 1: calculate the link similarities $S(e_{ik},e_{jk})$ for link e_{ik} and e_{jk}, and each link is initially assigned to a single cluster.

Stage 2: merge clusters iteratively if their similarity is highest using the average linkage function and ties, which are common, are agglomerated simultaneously.

Stage 3: stop merging when all links belong to a unique cluster.

The history of the clustering process is then stored in a dendrogram, which contains all the information of the hierarchical community organization. The similarity value at which two clusters merge is considered as the strength of the merged community, and is encoded as the height of the relevant dendrogram branch to provide additional information.

2.6. Dendrogram Partition

Hierarchical clustering methods repeatedly merge groups until all elements are members of a single cluster. This eventually forces highly disparate regions of the network into single clusters. To find meaningful communities rather than just the hierarchical organization pattern of communities, it is important to know where to partition the dendrogram. Modularity has been widely used for similar purposes, but is not easily defined for overlapping communities. Thus, we introduced a new quantity, the partition density *D*, which measures the quality of a link partition.

For a network with *M* links and *N* nodes, P=[P_1,…, P_c,…, P_k] is a partition of the links into *k* subsets. Then we define the density, D_c, of community *C* is

$$D_c = \frac{2m_c}{n_c(n_c-1)}(1-\frac{1}{k}\sum_{i=1}^{k}\frac{n_{ci}}{n_i}) \tag{4}$$

Where m_c is the number of links in subset P_c, *k* is the number of subset of network, n_c is the number of nodes which links of P_c touch, and n_{ci} is the number of common nodes between P_c and P_i. The partition density, *D*, is the average of D_c:

$$D = \frac{1}{k}\sum_{c=1}^{k} D_c \tag{5}$$

The maximum of *D* is 1, when every community is a fully connected clique and each community is independent.

3. Results and Analysis

To evaluate the performance of the proposed method three real-networks containing Gene network, Email network and Metabolic gene network are used to be the test networks. The main properties of them such as average degree, average shortest path length and average clustering coefficient are shown in Table 3.

Table 3. Properties of real-networks

Network name	Nodes	Links	Degree	Shortest path length	Clustering coefficient
Gene network	1624	3212	3.960	2.070	0.221
Email network	1133	5451	19.245	3.606	0.297
Metabolic gene network	962	2724	5.437	3.798	0.241

Mycobacterium tuberculosis is an extraordinarily successful pathogen that currently infects approximately one-third of the global population [38]. In order to evaluate our method, we use a new gene regulatory network (GRN) of Mycobacterium tuberculosis constructed by Sanz et al [39]. Removing duplicate interactions, the resulting TRN involving 1624 nodes and 3212 interactions, with 83 regulatory genes controlling the expression of 1598 genes, some main parameters are listed in Table 3. A GRN model represents the molecular regulation process by which genes regulate transcription of other genes. A gene X directly regulates a gene Y, if protein encoded by X is a transcriptional factor for Y.

The email communication network [40] covers all the email communications within a data set of around half a million emails. The nodes of the network are email addresses, and there is a link between two nodes if at least one email exists between them. Lastly, this network consists of 1133 nodes and 5451 links.

E. coli is considered the most complete available prokaryotic and the TRN of E. coli is the best characterized of all prokaryotic organisms. The metabolic functional gene transcriptional regulatory network (TRN) of Escherichia coli is a newly version of the TRN of E.coli, which was downloaded from the RegulonDB [41], controlling metabolism based on functional annotations from GeneProtEC [42] and Gene Ontology (GO) [43].

In order to evaluate algorithm quality we must be assessed in a different way. The most common method is modularity, which measures the relative number of intercommunity and intracommunity links. A high modularity indicates that there are more intracommunity links than would be expected by chance. However the modularity measure, *Q*, is defined only for non-intersect communities. Nicosia et al. [44] proposed proposed a new modularity measure, Q_{ov}, which is defined for directed networks with overlapping communities structures. In a network including *n* nodes and *m* links, k_i and k_j is the the number of links of *i* and *j*, respectively. Modularity, Q_{ov}, was defined as:

$$Q_{OV} = \frac{1}{m}\sum_{c\in C}\sum_{i,j\in V}\left[r_{ijc}A_{ij} - s_{ijc}\frac{k_i^{out}k_j^{in}}{m}\right] \qquad (5)$$

where r_{ijc} and s_{ijc} are the portion of the contribution to modularity given by community *C* because of link *l*(*i*, *j*) and A_{ij} are the terms of the adjacency matrix. Q_{ov}=0 when all vertices belong to one community, and higher values of Q_{ov} indicate stronger community structure. We use modularity, Q_{ov}, here to evaluate some well-known algorithms and our algorithm on real-world networks. Figure 4 shows the modularity, Q_{ov}, of the networks listed in Table 3.

4. Conclusion

In this paper, we presented a new algorithm for detecting overlapping communities in directed networks, which partition communities with links instead of nodes. A new measure of link similarity based on triangle distribution has been introduced. Using link similarity values, a dendrogram of link has been constructed by hierarchical clustering method. To determine the best cut level, a new partition density has been introduced. Mainly contribution of the algorithm is that it can successfully reveal overlapping communities and hierarchies simultaneously in directed network. The algorithm has been applied to server real-network compared with several popular community structure identify algorithms. The results show that it is rather efficient to discover the community structure in directed networks. However its full potential remains unexplored. Our work has primarily focused on the highly overlapping community structure of complex networks, but an existing limitation of our algorithm is the relationship between the overlaps and hierarchical. Therefore, the hierarchy that organizes these overlapping communities keep up great promise for further study.

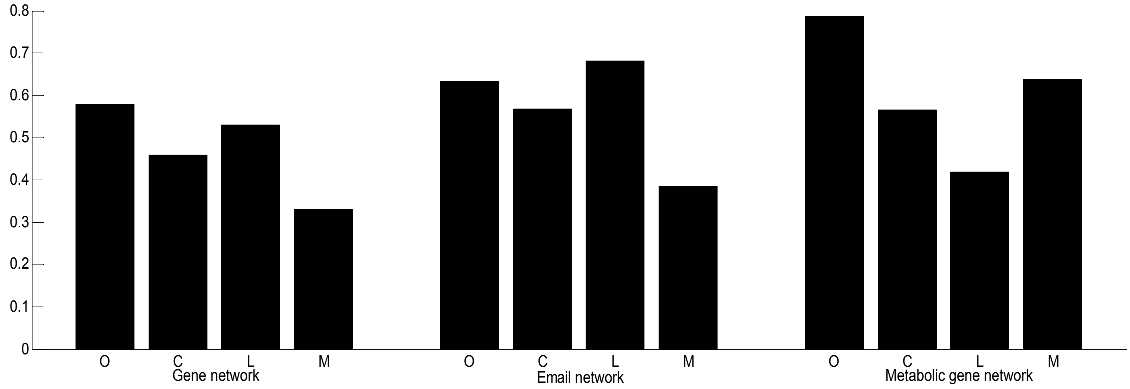

Figure 4. Q_{ov} of each real-network calculated by four community detecting algorithm. O: Our algorithm; C: Clique percolation algorithm; L: Link partition algorithm; M: Modularity spectral optimization algorithm

References

[1] Rotundo G, Ausloos M. Organization of networks with tagged nodes and biased links: A priori distinct communities The case of intelligent design proponents and Darwinian evolution defenders. *Physica a-Statistical Mechanics and Its Applications.* 2010; 389(23): 5479-5494.

[2] Mason O, Verwoerd M. Graph theory and networks in Biology. *Iet Systems Biology*. 2007; 1(2): 89-119.

[3] Leicht E A, Clarkson G, Shedden K, et al. Large-scale structure of time evolving citation networks. *European Physical Journal B.* 2007; 59(1): 75-83.

[4] Newman M E J. Communities, modules and large-scale structure in networks. *Nature Physics.* 2012; 8(1): 25-31.

[5] Newman M E J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103(23): 8577-8582.
[6] Dorogovtsev S N, Goltsev A V, Mendes J F F. Critical phenomena in complex networks. *Reviews of Modern Physics*. 2008; 80(4): 1275-1335.
[7] Kaltenbach H M, Stelling J. Modular analysis of biological networks. *Adv Exp Med Biol.* 2012; 736: 3-17.
[8] Fortunato S. Community detection in graphs. *Physics Reports-Review Section of Physics Letters*. 2010; 486(3-5): 75-174.
[9] Yang B, Jin D, Liu J M, et al. Hierarchical community detection with applications to real-world network analysis. *Data & Knowledge Engineering*. 2013; 83: 20-38.
[10] Ochab J K, Burda Z. Maximal entropy random walk in community detection. *European Physical Journal-Special Topics.* 2013; 216(1): 73-81.
[11] Lancichinetti A, Radicchi F, Ramasco J J, et al. Finding statistically significant communities in networks. *Plos One*. 2011; 6(4): e18961.
[12] Ball B, Karrer B, Newman M E J. Efficient and principled method for detecting communities in networks. *Phys Rev E*. 2011; 84(3): 036103-1-036103-13.
[13] Gao Z K, Jin N D, Ieee. Detecting community structure in complex networks based on K-means clustering and data field theory. 20th Chinese Control and Decision Conference. Yantai. 2008; 4411-4416.
[14] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E*. 2006; 74(3): 1-22.
[15] Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E*. 2004; 69(2): 1-16.
[16] Kashef R, Kamel M S. Cooperative clustering. *Pattern Recognition*. 2010; 43(6): 2315-2329.
[17] Newman M E J. Random Graphs with Clustering. *Physical Review Letters*. 2009; 103(5).
[18] Xu R, Wunsch D. Survey of clustering algorithms. *Ieee Transactions on Neural Networks*. 2005; 16(3): 645-678.
[19] Fagiolo G. Clustering in complex directed networks. *Phys Rev E*. 2007; 76(2).
[20] Wilkinson D M, Huberman B A. A method for finding communities of related genes. *Proc Natl Acad Sci U S A.* 2004; 101 Suppl 1: 5241-5248.
[21] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101(9): 2658-2663.
[22] Zanjani A A H, Darooneh A H. Finding communities in linear time by developing the seeds. *Phys Rev E*. 2011; 84(3).
[23] Rosvall M, Bergstrom C T. An information-theoretic framework for resolving community structure in complex networks. *Proc Natl Acad Sci U S A*. 2007; 104(18): 7327-7331.
[24] Li Z, Zhang S, Wang R-S, et al. Quantitative function for community detection. *Phys Rev E*. 2008; 77(3): 036109-1-036109-9.
[25] Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(12): 7821-7826.
[26] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks. *Proc Natl Acad Sci U S A*. 2004; 101(9): 2658-2663.
[27] Fortunato S, Latora V, Marchiori M. Method to find community structures based on information centrality. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2004; 70(5 Pt 2): 056104.
[28] Pons P, Latapy M. Computing communities in large networks using random walks. *Lect Notes Comput Sc*. 2005; 3733: 284-293.
[29] Yang B, Liu J M. Discovering Global Network Communities Based on Local Centralities. *Acm Transactions on the Web*. 2008; 2(1).
[30] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*. 2010; 466(7307): 761-U711.
[31] Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities. *Phys Rev E*. 2009; 80(1).
[32] Resendis-Antonio O, Freyre-Gonzalez J A, Menchaca-Mendez R, et al. Modular analysis of the transcriptional regulatory network of E-coli. *Trends in Genetics*. 2005; 21(1): 16-20.
[33] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*. 2005; 435(7043): 814-818.
[34] Leicht E A, Newman M E J. Community structure in directed networks. *Physical Review Letters*. 2008; 100(11): 118703-1-118703-4.
[35] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*. 2009; 11: 1-20.
[36] Milo R, Shen-Orr S, Itzkovitz S, et al. Network motifs: Simple building blocks of complex networks. *Science*. 2002; 298(5594): 824-827.

[37] Day W E, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*. 1984; 1(1): 7-24.

[38] World Health Organization. Global tuberculosis control. Geneva. Report number: WHO/CDS/TB/2000.275:1179. 2000.

[39] Sanz J, Navarro J, Arbues A, et al. The Transcriptional Regulatory Network of Mycobacterium tuberculosis. *Plos One*. 2011; 6(7).

[40] Guimera R, Danon L, Diaz-Guilera A, et al. Self-similar community structure in a network of human interactions. *Phys Rev E*. 2003; 68(6): 065103-1-065103-4.

[41] Serres M H, Goswami S, Riley M. GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Research*. 2004; 32: D300-D302.

[42] Consortium T G O. The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.* 2012; 40(Database issue): D559-564.

[43] Gama-Castro S, Salgado H, Peralta-Gil M, et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research.* 2011; 39: D98-D105.

[44] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics-Theory and Experiment.* 2009; P03024.