TELKOMNIKA, Vol. 11, No. 9, September 2013, pp. 5441~5447
ISSN: 2302-4046

Overlapping Communities Detection based on Link Partition in Directed Networks

Qingyu Zou*, Fu Liu, Tao Hou, Yihan Jiang
College of Communication Engineering, Jilin University, Changchun 130000, Jilin, China
*Corresponding author, e-mail: 37008574@qq.com

Abstract

Many complex systems can be described as networks to comprehend both the structure and the function. Community structure is one of the most important properties of complex networks. Detecting overlapping communities in networks have been more attention in recent years, but the most of approaches to this problem have been applied to the undirected networks. This paper presents a novel approach based on link partition to detect overlapping communities structure in directed networks. In contrast to previous researches focused on grouping nodes, our algorithm defines communities as groups of directed links rather than nodes with the purpose of nodes naturally belong to more than one community. This approach can identify a suitable number of overlapping communities without any prior knowledge about the community in directed networks. We evaluate our algorithm on a simple artificial network and several real-networks. Experimental results demonstrate that the algorithm proposed is efficient for detecting overlapping communities in directed networks.

Keywords: *directed network, overlapping communities, transform, link, modularity*

1. Introduction

Recently many complex systems in nature and society have been represented as networks to understand structure, dynamic, robust, and evolution [1-3]. Community structure is one of the most important features of many complex networks [1, 4, 5]. The detection and analysis of community structures have been attracted much attention in many applications [6-11], because communities reveal topological relationships between system elements and represent function [12, 13].

So far, 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. One approaches of the first scheme are based on a measure named betweenness. It calculates one of several measures [14, 15] of the flow of traffic across the links of a network and then removes those links with the most traffic from the network. Two other related approaches are the use of fluid-flow and current-flow analogies [16] to identify links for removal. A different class of optimization methods is whose based on information-theoretic ideas, such as the minimum description length methods of Rosvall and Bergstrom [17]. The basic idea is to define a quantity that is high for `good' divisions of a network and low for `bad' ones, and then to search through possible divisions for the one with the highest score. Various different measures for assigning scores have been proposed, such as the likelihood-based measures [18] and others [19], but the most widely used approach is the modularity [20]. 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 methods, such as link betweenness [21], link clustering coefficient [22], information centrality [23], similarity based on random walks [24], clustering centrality [25], 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 [26] and T. S. Evans et al [27] 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

Received February 5, 2013; Revised June 15, 2013; Accepted June 25, 2013

of the networks that we would like to study are directed, and a node may belong 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 [28]. 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 links similarity rather than the node's property is proposed to detect the overlapping communities structures in directed networks. On the basis of links similarity of a directed network, we transformed an unweighted directed network into a weighted undirected network, whose nodes are the links of the original network and the weight of links is the links similarity of the original network. Then, we used hierarchical clustering with the shortest path between nodes in the transformed network to identify community structure. In order to measure the strength of the community structure and obtain the most relevant communities, a popular algorithm Newman-Girvan modularity Q [29] was used.

We compared the performance of our algorithm with three successful methods: clique percolation [30], link partition [27], and modularity spectral optimization [31] with a simple artificial network and three real-world networks including Gene network, Email network and Amazon.com 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. To measure the effectiveness of the community detecting algorithm, the extending modularity Q_{ov} [32] had been used on real-world networks.

2. Research Method

2.1. Community

A community, also called cluster or module, consists of nodes and links between these nodes. Although no common definition has been agreed upon, it is widely accepted that a community should have more internal than external connections [15, 33]. The nodes in the same community often have common properties and densely interconnected compared to the rest of the network. It is noted that two communities may overlap each other while a node can connect with different communities simultaneously [1]. 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.

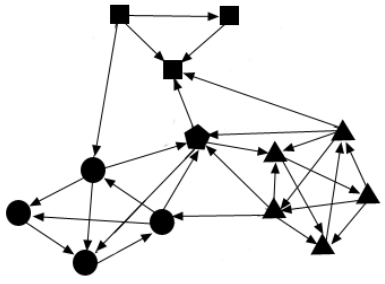

Figure 1. Example network showing community structure, the nodes of this network are divided into three groups; node pentagon is the common node of both the circle and triangle communities

2.2. The Shortest Path

Let u, v be two nodes in a network G. Then a sequence of nodes from u to v is a path from u to v. The geodesic distance, $d(u,v)$, from u to v is the length of the shortest path from u to v in G. If on such path exists, then we set $d(u,v)=\infty$ and the shortest path from u to u is 0 [34].

2.3. Modularity

In order to quantify the community structure of a network, Newman and Girvan [20] proposed the modularity Q as a measure of a network partition:

$$Q = \sum_{i=1}^{k} \left(e_{ii} - a_i^2 \right) \tag{1}$$

Where e_{ii} is the fraction of links belonging to community i in the total weights of all links and a_i is the fraction of links connecting community i with other communities.

2.4. Link Similarity

The link similarity is a measure of closeness between a pair of links. It is clear that in the same community of network the node-node connections are more densely and the shortest paths between pairs of nodes are shorter than in different communities. In accordance with the this principles, the similarity between links e_{il} and e_{jk} is:

$$S(e_{il}, e_{jk}) = \frac{\left| W(e_{il}, e_{jk}) \right|}{D^2(e_{il}, e_{jk})} \tag{2}$$

Where $S(e_{ik},e_{jk})$ means link similarity value, $D(e_{il},e_{jk})$ is the shortest paths between links e_{il} and e_{jk}. As shown in Figure 2, there are eight shortest paths among four nodes, represented by d_{lk}, d_{ik}, d_{lj}, d_{ij}, d_{kl}, d_{jl}, d_{ki}, d_{ji} respectively. $D(e_{il},e_{jk})$=min(d_{lk}, d_{ik}, d_{lj}, d_{ij}, d_{kl}, d_{jl}, d_{ki}, d_{ji})+1, if e_{il} and e_{jk} have a common node, $D(e_{il},e_{jk})$=1.

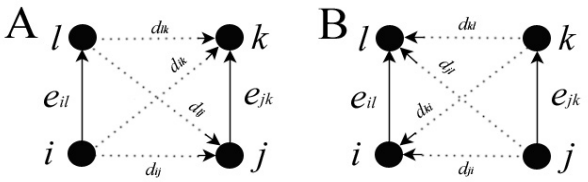

Figure 2. The Shortest Paths between Nodes of a Pair of Links

$W(e_{ik},e_{jk})$ is the compactness between links e_{il} and e_{jk}. After found the shortest path nodes, there are four similarity measures shown in Figure 3.

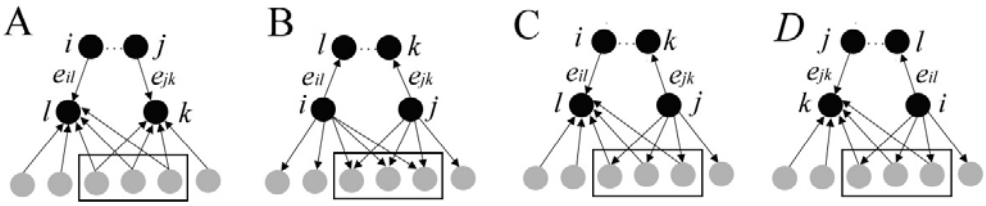

Figure 3. Four cases of the link compactness measure $W(e_{ik},e_{jk})$ between link e_{ik} and e_{jk}

The compactness between links in Figure 3 is:

$$\text{A: } S(e_{il}, e_{jk}) = \frac{\left| n_+(l) \cap n_+(k) \right|}{\left| n_+(l) \cup n_+(k) \right|} \tag{3}$$

$$\text{B: } S(e_{il}, e_{jk}) = \frac{\left| n_-(l) \cap n_-(k) \right|}{\left| n_-(l) \cup n_-(k) \right|} \tag{4}$$

$$\text{C: } S(e_{il}, e_{jk}) = \frac{\left|n_+(l) \cap n_-(k)\right|}{\left|n_+(l) \cup n_-(k)\right|} \quad (5)$$

$$\text{D: } S(e_{il}, e_{jk}) = \frac{\left|n_-(l) \cap n_+(k)\right|}{\left|n_-(l) \cup n_+(k)\right|} \quad (6)$$

Where $n_+(i)$ is the neighbors of a node *i* which directing it, $n_-(i)$ is the neighbors of a node *i* which it directing.

2.5. Algorithm

Our algorithm comprises the following phases:

a) Calculate the link similarities $S(e_{ik},e_{jk})$ for link eik and e_{jk} and transform original network into a new weighted undirected network, whose nodes are the links of the original network and the weight of links is links similarity of the original network.

b) Calculate the shortest path between each pair of nodes in the transformed network. According to the shortest path, hierarchical clustering algorithm [35] is used to find community structures.

c) Using modularity on the transformed network to find meaningful communities rather than just the hierarchical organization pattern of communities.

3. Results and Analysis

To evaluate the performance of the proposed method a simple artificial network and three real-networks containing Gene network, Email network and Amazon.com network are used to be the test networks.

3.1. Simple Artificial Network

To make our method clear to readers, we show a small-scale example directed network consisted of five nodes and six links, as shown in Figure 4A. There are two communities, sharing the No. 5 node, in this network. According to Equation (2), the similarities of the links in Figure 4A are computed and transform to a new weighted undirected network composed of six nodes and ten links as shown in Figure 4B. Using hierarchical clustering algorithm on the network shown in Figure 4B with the shortest path, the results presented in Figure 1C illustrates two overlapping communities, one contain No. 1, 2 and 5 nodes and the other contain No. 3, 4 and 5 nodes, detected by our method.

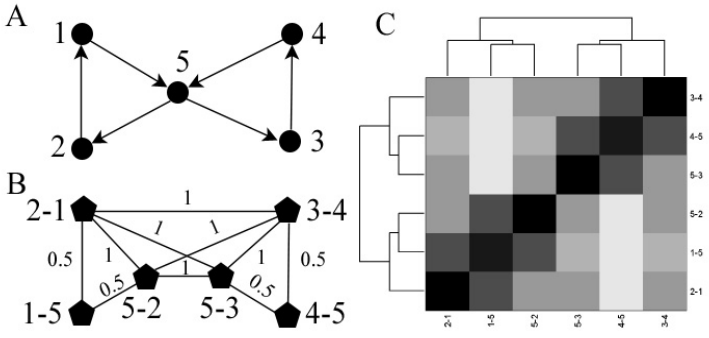

Figure 4. Detecting Community Structure in the Example Network using the Proposed Method (A) original network (B) transformed network (C) hierarchical clustering result.

3.1. Application to Real-network

We have run the clique percolation algorithm, link partition algorithm, modularity spectral optimization link algorithm and our algorithm on three real-world networks. 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. [32] 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:

$$Q_{OV} = \frac{1}{m}\sum_{c\in C}\sum_{i,j\in V}\left[r_{ijc}A_{ij} - s_{ijc}\frac{k_i^{out}k_j^{in}}{m}\right] \quad (7)$$

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 5 shows the modularity, Q_{ov}, of the networks listed in Table 1.

Table 1. Properties of Real-networks

Network name	Nodes	Links	Degree	Shortest path length	Clustering coefficient
Gene network	1860	4150	4.796	2.715	0.313
Email network	1133	5451	19.245	3.606	0.297
Amazon.com network	409687	2464630	12.03	3.865	0.171

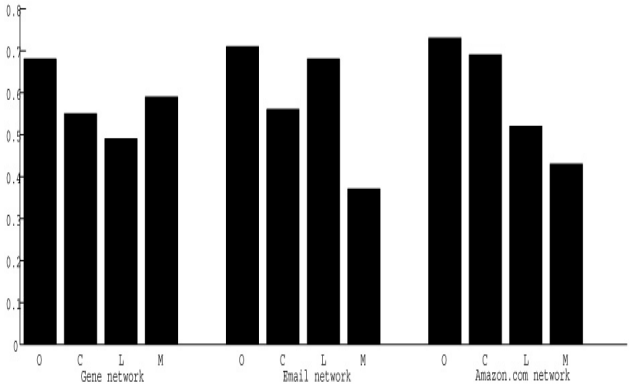

Figure 5. Q_{ov} of each Real-network Calculated by Four Community Detecting Algorithm
O:Our algorithm; C:Clique percolation algorithm; L:Link partition algorithm; M:Modularity spectral optimization algorithm.

The gene transcriptional regulatory network (TRN) of Escherichia coli is one of the most elaborate reconstructions currently available. In order to evaluate our method, we use the method presented in article [36] to build the TRN of E. coli that were organized in RegulonDB [37]. Removing duplicate interactions, the resulting TRN involving 1680 nodes and 4150 interactions, with 186 TFs (regulatory genes) controlling the expression of 1499 genes, some main parameters are listed in Table 1. A TRN model represents the molecular regulation process by which genes regulate transcription of other genes. A gene X directly regulates a gene Y, if protein that is encoded by X is a transcriptional factor for Y.

The email communication network [38] 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.

The Amazon.com network [39] is a purchasing network from the online vendor Amazon.com, collected in August 2003. Amazon sells a variety of products, particularly books and music, and as part of their web sales operation they list for each item A the ten other items most frequently purchased by buyers of A. This information can be represented as a directed network in which vertices represent items and there is a link from item A to another item B if B was frequently purchased by buyers of A.

4. Conclusion

In this paper, we presented a new algorithm for detecting overlapping communities in directed networks based on the link similarity, which incorporate communities overlap and link direction. A new measure of link similarity has been introduced. Using link similarity values, we have transformed an unweighted directed network into a new weighted undirected network and detected communities using hierarchical clustering method on the transformed 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 function community structure of directed networks. However its full potential remains unexplored. Our work has primarily focused on the highly overlapping communities structure of complex networks, but the hierarchy that organizes these overlapping communities holds great promise for further study.

Acknowledgements

This research work is supported partially by the Jilin Province Science and Technology Development projects 10100505, and partially by the Jilin University Graduate innovative research projects 20121101.

References

[1] Newman MEJ. Communities, modules and large-scale structure in networks. *Nature Physics*. 2012; 8(1): 25-31.
[2] Maslov S. Complex networks - Role model for modules. *Nature Physics*. 2007; 3(1): 18-19.
[3] Strogatz SH. Exploring complex networks. *Nature*. 2001; 410(6825): 268-276.
[4] Newman MEJ. *Modularity and community structure in networks.* Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(23): 8577-8582.
[5] Dorogovtsev SN, Goltsev AV, Mendes JFF. Critical phenomena in complex networks. *Reviews of Modern Physics.* 2008; 80(4): 1275-1335.
[6] Yang B, Jin D, Liu JM, et al. Hierarchical community detection with applications to real-world network analysis. *Data & Knowledge Engineering.* 2013; 83: 20-38.
[7] Wu ZH, Lin YF, Wan HY, et al. Efficient overlapping community detection in huge real-world networks. *Physica a-Statistical Mechanics and Its Applications.* 2012; 391(7): 2475-2490.
[8] Li K, Gong X, Guan S, et al. Efficient algorithm based on neighborhood overlap for community identification in complex networks. *Physica a-Statistical Mechanics and Its Applications.* 2012; 391(4): 1788-1796.
[9] Becker E, Robisson B, Chapple CE, et al. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics.* 2012; 28(1): 84-90.
[10] Hou L, Wang L, Qian MP, et al. Modular analysis of the probabilistic genetic interaction network. *Bioinformatics.* 2011; 27(6): 853-859.
[11] Steinhaeuser K, Chawla NV. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters.* 2010; 31(5): 413-421.
[12] Kaltenbach HM, Stelling J. Modular analysis of biological networks. *Adv Exp Med Biol.* 2012; 736: 3-17.
[13] Fortunato S. Community detection in graphs. *Physics Reports-Review Section of Physics Letters.* 2010; 486(3-5): 75-174.
[14] Wilkinson DM, Huberman BA. *A method for finding communities of related genes.* Proc Natl Acad Sci USA. 2004; 101 Suppl 1: 5241-5248.
[15] 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.
[16] Zanjani AAH, Darooneh A H. Finding communities in linear time by developing the seeds. *Phys Rev E.* 2011; 84(3).

[17] Rosvall M, Bergstrom CT. *An information-theoretic framework for resolving community structure in complex networks.* Proc Natl Acad Sci USA. 2007; 104(18): 7327-7331.
[18] Karrer B, Newman ME. Stochastic blockmodels and community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys.* 2011; 83(1 Pt 2): 016107.
[19] Li Z, Zhang S, Wang RS, et al. Quantitative function for community detection. *Phys Rev E.* 2008; 77(3).
[20] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E.* 2004; 69(2): 1-16.
[21] Girvan M, Newman MEJ. *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.
[22] Radicchi F, Castellano C, Cecconi F, et al. *Defining and identifying communities in networks*. Proc Natl Acad Sci USA. 2004; 101(9): 2658-2663.
[23] 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.
[24] Pons P, Latapy M. Computing communities in large networks using random walks. *Lect Notes Comput Sc.* 2005; 3733: 284-293.
[25] Yang B, Liu JM. Discovering Global Network Communities Based on Local Centralities. *Acm Transactions on the Web.* 2008; 2(1).
[26] Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature.* 2010; 466(7307): 761-U711.
[27] Evans TS, Lambiotte R. Line graphs, link partitions, and overlapping communities. *Phys Rev E.* 2009; 80(1).
[28] Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, et al. Modular analysis of the transcriptional regulatory network of E-coli. *Trends in Genetics.* 2005; 21(1): 16-20.
[29] Newman MEJ. Detecting community structure in networks. *European Physical Journal B.* 2004; 38(2): 321-330.
[30] 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.
[31] Leicht EA, Newman MEJ. Community structure in directed networks. *Physical Review Letters*. 2008; 100(11).
[32] 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.
[33] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics.* 2009; 11.
[34] Mason O, Verwoerd M. Graph theory and networks in Biology. *Iet Systems Biology.* 2007; 1(2): 89-119.
[35] Day WE, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification.* 1984; 1(1): 7-24.
[36] Salgado H, Martinez-Flores I, Lopez-Fuentes A, et al. Extracting regulatory networks of Escherichia coli from RegulonDB. *Methods Mol Biol.* 2012; 804: 179-195.
[37] 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.
[38] 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).
[39] Clauset A. Finding local community structure in networks. *Phys Rev E.* 2005; 72(2).