A Dynamical Method to Extract Communities Induced by Low or Middle-degree Nodes

Junhua Zhang*,§, Zhi-Ping Liu†,§, Xiang-Sun Zhang‡,§ and Luonan Chen†,§

*Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
Email: zjh@amt.ac.cn
†Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai 200032, China
Email: zpliu@sibs.ac.cn, lnchen@sibs.ac.cn
‡Key Laboratory of Management, Decision and Information Systems, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
Email: zxs@amt.ac.cn
§National Center for Mathematics and Interdisciplinary Sciences,
Chinese Academy of Sciences, Beijing 100190, China

***Abstract*—Many networks are proved to have community structure. Dense communities have been intensively investigated in recent years, oppositely seldom attention has been paid to sparse ones, which refer to those communities induced by low or middle-degree nodes rather than high-degree components. Recently, it has gradually been recognized that sparse community is also an important structure in biological networks because most disease genes and drug targets are within it. In this paper, we propose a dynamical method to extract sparse communities in complex networks by constructing local synchronization properties of phase oscillators. Compared to dense communities, sparse ones provide more general building and functional blocks in the networks without emphasis on the dominance of internal degrees over outside ones as well as the constraints of high degree connectors.**

***Index Terms*—Complex network; sparse community structure; dynamical method; local synchronization.**

I. Introduction

In complex network study, community structure refers to a group of nodes with denser linkages internally and sparser connections between groups. Often, the difference between the internal degrees and those of intragroup is used to define the network structure from topology only. Identifying communities within a network is a rather difficult task. The number of communities, if any, within the network is typically unavailable. And the communities are often of various density, size and practical meaning. A number of methods have been developed and achieved various level of success. Most community detection methods focus on directly partitioning the entire network into communities, without considering that some nodes may not fit in with any of them, and forcing every node into a community can distort the truly underlying results. Very recently, Zhao et al. [1] proposed an approach that extracts only dense communities, allowing for arbitrary structure in the remainder of the network. The dense subgraphs in a network are generally identified as the communities, while the other particular subnetwork structures also have particular implications.

Different from the existing works to detect dense community structures in complex networks, in this paper our attention focuses on the sparse communities. Compared to these dense communities, the communities considered here do not have the constraints of components with high degrees or of hubs in the network. They refer to the communities of functional implications without the preference of nodes of the highest degree. These communities correspond to more certain functionally important building blocks without constraints of dense topological structure.

Generally, the dense communities in a complex network are regarded as important blocks from topological perspective. The dense subgraph tends to contain nodes with high degrees and many of them are hubs and connectors. While in many cases, hub nodes should not highly relate to certain functions. The building blocks of a community in the complex network are not singly referred to the dense subnetworks. For instance, disease genes are usually not hubs in a protein interaction network because it would be mortal for individuals [2]. Also, from the controllability perspective, driver nodes in a network tend to avoid hubs and the fraction of driver nodes is significantly higher among low-degree nodes than among the hubs [3]. Those important disease and driver genes do not show the preference of hubs in the networks. Hence, to identify these important functional components, we need to focus on these communities with low or middle-degree nodes, i.e., sparse communities, which are functional blocks of the network in which those driver nodes performing critical functions are contained.

By constructing a local synchronization strategy of phase oscillators, in this paper we propose a dynamical method to extract communities induced by low or middle-degree nodes in complex networks. Without special requirements

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

and constraints of high degrees and linkage densities in the communities, we identify sparse communities in the network with particular implications and functional importance. Moreover, instead of decomposing nodes into certain communities, we extract sparse communities by synchronization strategy directly. Specifically, we identify these interesting sparse communities in two social networks as well as one protein-protein interaction network. The results show that these sparse communities play important roles with both practical and functional implications.

II. Methods

To extract sparse communities in the network, our attention will be focused on the area with low or middle-degree nodes. The methodology adopted here is based on synchronization properties of phase oscillators, for which a local synchronization strategy is constructed for the considered nodes.

A. The dynamical model

Synchronization is one of the most captivating cooperative phenomena in nature. It is widely observed in biological, chemical, physical, and social systems, and it has attracted the interest of scientists for centuries. The dynamics of complex networks has been largely studied in recent years [4], [5]. The emergence of synchronization patterns in these networks has been shown to be closely related to the underlying topology of interactions. For example, some researchers demonstrated that the dynamical process towards synchronization shows different patterns over time intrinsically connected with the hierarchical organization of communities in complex networks [6]. The ubiquity of synchronization phenomena in the real world makes this approach appealing from a physical and biological perspective.

One of the most successful attempts to understand synchronization phenomena was from Kuramoto [7], who analyzed a model of phase oscillators coupled through the sine of their phase differences. The model is rich enough to display a large variety of synchronization patterns and sufficiently flexible to be adapted to many different contexts [8]. The Kuramoto model consists of a population of n coupled phase oscillators where the phase of the ith unit, denoted by θ_i, evolves in time according to the following dynamics:

$$\frac{d\theta_i}{dt} = \omega_i + \sum_j K_{ij} sin(\theta_j - \theta_i) \quad i = 1, \ldots, n \qquad (1)$$

where ω_i stands for its natural frequency and K_{ij} describes the coupling between units. The original model studied by Kuramoto assumed mean-field interactions $K_{ij} = K, \forall i, j$. Some algorithms for community detection have been developed based on Eq. (1) [9].

In this paper our goal is to extract low or middle-degree modules, so we construct a local synchronization strategy of phase oscillators for this purpose. In detail, consider an undirected network $G(V, E)$ with the node set V, the edge set E depicted by the symmetric adjacency matrix $A = [a_{ij}]_{n\times n}$, where $a_{ij} = 1$ if nodes i and j are connected and otherwise $a_{ij} = 0$, and n is the size of the network. The modified Kuramoto model is used as:

$$\frac{d\theta_i}{dt} = \sigma \sum_j a_{ij} e_{ij} sin(\theta_j - \theta_i), \quad i = 1, \ldots, n \qquad (2)$$

where σ is the coupling strength, and $E = [e_{ij}]_{n\times n}$ is another 0-1 matrix whose elements can be determined based on the topological structure of the low or middle-degree nodes which will be introduced in the next subsection.

The coefficient a_{ij} means that only the two connected nodes could be coupled, and e_{ij} assigns real coupling such that only the low or middle-degree nodes satisfying certain conditions can be coupled. These two terms jointly ensure that the sparse communities can be extracted by this model.

B. The algorithm for extracting sparse communities

Based on Eq. (2), an iterative algorithm is proposed here. Because the goal is to extract sparse communities in the area of the low to middle-degree nodes in the network, two parameters d_1 and d_2 are introduced, which are the minimal and maximal degrees of the nodes considered, respectively. A local synchronization strategy of phase oscillators is realized by properly defining the e_{ij}, which ensures that the extracted sparse communities are relatively dense in the sparse area of the network. In detail, the algorithm for extracting sparse communities is as follows:

1) Choose d_1 and d_2, respectively, according to the degree distribution of the network.
2) Let S denote the set of nodes with degrees in the interval $[d_1, d_2]$. For $i, j \in S$, if i, j have common nearest neighbors in S, or the intersection of i's nearest neighbors in S with j's second nearest neighbors in S is nonempty and vice versa, $e_{ij} = 1$; otherwise, $e_{ij} = 0$.
3) Initially, θ_i is randomly and uniformly distributed in the intervals $[0, 2\pi]$. In this paper we choose the coupling strength σ as 10. Numerical results are obtained by integrating Eq. (2) using the Runge-Kutta method with step size 0.01.
4) The iteration will be stopped if the number of steps reaches some preassigned number, where 1000 is used in this paper.
5) If the difference of the phases of the nodes is less than 0.0001, they are in the same sparse communities.

III. Experiments

We test the performance of the proposed method by applying it to three real-world networks. At first, two small social networks are employed and the corresponding results can provide us an intuitive understanding about the sparse community. Then a large biological network, i.e., the human protein-protein interaction network, is introduced and the distributional properties of middle-degree drug targets and disease genes in the sparse communities are also investigated.

For the coupling strength σ in Eq. (2), if it is too low or high, the number and the size of the extracted sparse communities usually tend to small. For the experiments in this paper the

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

value between 8 and 30 is considered to be appropriate, here $\sigma = 10$ is used. On the other hand, because of the randomness of the initial phases of the oscillators, the results may be different for different runs. Here the results are obtained by picking out all the largest different subgraphs through three runs. Without special statement, only the sparse communities with more than three nodes are presented in the following experiments.

A. The karate club network

The famous karate club network analyzed by Zachary [10] is widely used as a test example for benchmark of detecting communities in complex networks [11]. The network consists of 34 members of a karate club as nodes and 78 edges representing friendship between members of the club which was observed over a period of two years (Fig. 1). Due to a disagreement between the club's administrator (node 34) and the club's instructor (node 1), the club split into two smaller ones, which are represented by circles and squares in Fig. 1, respectively.

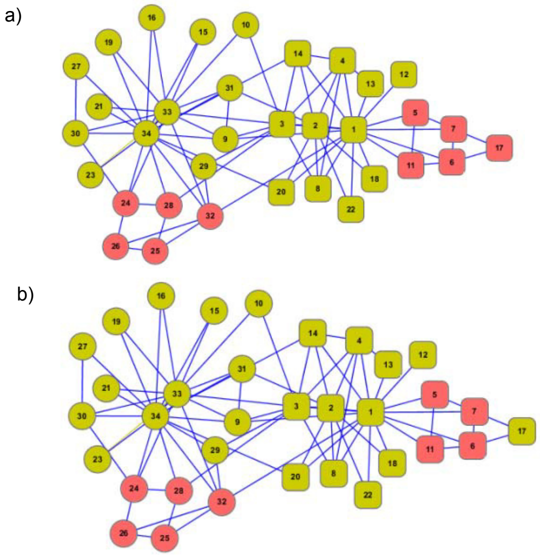

Fig. 1. The karate club network and the sparse communities (the groups of red nodes) extracted by Eq. (2) and the proposed algorithm. Two situations are investigated: a) $d_1 = 2$ and $d_2 = 6$, and b) $d_1 = 3$ and $d_2 = 6$.

A huge number of the existing methods intended to uncover the actual division of the original club, and therefore a proper partition of the network is a solution of these methods. Our goal is not to partition the network into communities, but just to extract the sparse ones induced by low or middle-degree nodes. From Fig. 2, we know that most nodes in the karate club network have degrees from 2 to 6 (about 82%). Although high-degree nodes (or hubs) (now with degrees not less than 9) are important in the network (for examples, nodes 1 and 34), we concentrate here on the low to middle-degree nodes only for our purpose of detecting interesting communities. Using Eq. (2) and the proposed algorithm, two sparse communities were extracted. The results are displayed in Fig. 1, where a) and b) correspond to $d_1 = 2$ and $d_2 = 6$, and $d_1 = 3$ and $d_2 = 6$, respectively.

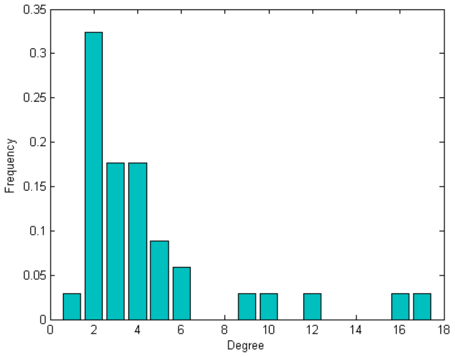

Fig. 2. The degree distribution of the karate club network.

For each of the sparse communities in Fig. 1, although any node's degree is not too high, close relationship is held among the members of the community. For practical implications, they may usually as a whole form a sparse community to carry out the commands of the administrator or the instructor in the karate club.

B. The scientific collaboration network

The scientific collaboration network collected by Girvan and Newman [12] is another widely used test example for methods of detecting communities in complex networks. This network consists of 118 nodes (scientists) and 200 edges (representing the collaboration among these scientists). It is rather a sparse network because the average degree of each node is only about 1.7. The degree distribution of the network is displayed in Fig. 3.

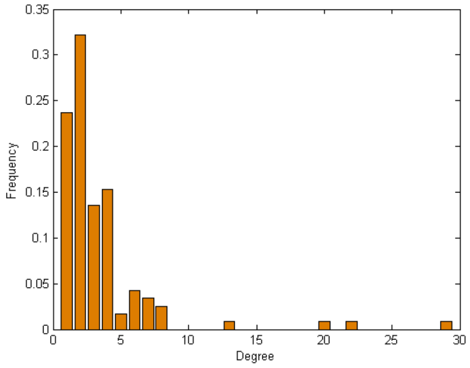

Fig. 3. The degree distribution of the scientific collaboration network.

Naturally the scientists belong to four groups according to their research interests, i.e., structure of RNA, statistical physics, mathematics ecology and agent-based models, which are represented by different colors in Fig. 4 and Fig. 5. Some collaboration is observed between the last two groups, but it is very little between other groups. Usually scientists cooperated with others in the same group, especially many members

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

worked with the hubs (very large-degree nodes) in the first three groups.

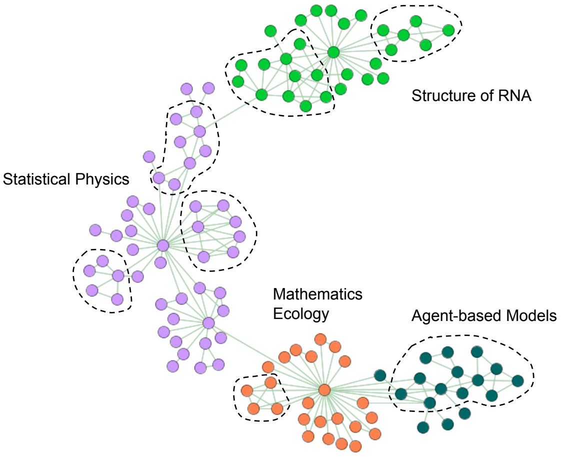

Fig. 4. The scientific collaboration network and the extracted sparse communities encompassed by dashed lines ($d_1 = 2$ and $d_2 = 8$).

Fig. 3 indicates that the network has many leaf nodes (one-degree nodes). Its degrees are gathered in the interval [1, 8]. Ignoring the leaf nodes, those with degrees 2 to 8 have a proportion of about 73%. Using Eq. (2) and the proposed algorithm, when we chose $d_1 = 2$ and $d_2 = 8$, seven sparse communities were extracted (Fig. 4); and when $d_1 = 3$ and $d_2 = 8$ were used, five sparse communities were obtained (Fig. 5).

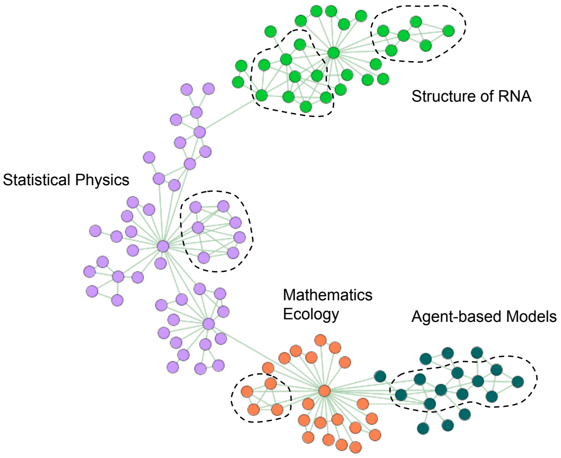

Fig. 5. The scientific collaboration network and the extracted sparse communities encompassed by dashed lines with $d_1 = 3$ and $d_2 = 8$.

Fig. 4 and Fig. 5 demonstrate that, in the area of low to middle-degree nodes in the network, we extracted the subgraphs although each member in them has not so much collaboration with others like the hubs, there are still much collaboration within the subgraphs relative to outside parts. For each interest group, because of the wide collaboration between the hub with others, the research interest may be influenced by the hub. But on the other hand, the hub's interest may also be influenced by the sparse communities connecting to it due to the close collaboration within the communities.

C. Simple applications to drug target and disease gene analysis in the human protein-protein interaction network

There is a growing awareness that networks of protein interactions and gene regulations are the keys to understand diseases and find accurate drug targets [13]. There have been several studies on the structure and statistical properties of protein interactions and how disease genes and drug targets are distributed over the protein-protein interaction (PPI) networks [2], [14], [19]. Here we further investigate the possible modules to which disease genes and drug targets may belong in the human PPI network, with the expectation that it should provide us a clue to find other disease genes and get possible drug targets in the future.

The human PPI data here used are from [2], [14], which is obtained by using two high-quality systematic yeast two-hybrid experiments [15], [16] and PPIs from literature by manual curation [15]. The integrated set of PPIs contains 22,052 non-self-interacting, nonredundant interactions between 7,496 genes, of which 1,203 are associated with diseases by the Online Mendelian Inheritance in Man (OMIM; [17]) and 263 are targets of FDA-approved drugs [18].

Hase et al. [19] studied the architectural properties of the PPI network structures, and revealed that there are extensive interconnections among middle-degree nodes that form the backbone of the networks. Further analysis on the degree distribution of human drug targets and disease genes indicated that there are advantageous drug targets and disease genes among nodes with low to middle-degree nodes.

Our analysis focuses on middle-degree nodes, i.e., the nodes with degrees from 6 to 30 in the human PPI network according to [19]. There are 89 middle-degree drug targets and 396 middle-degree disease genes respectively in the network. Using Eq. (2) and our algorithm, four sparse communities were extracted, in which there is a very large community with 1357 nodes. We found that 73 of 89 (i.e., about 82%) middle-degree drug targets are in this community. That is to say, these middle-degree target genes are mostly on the backbone of the network. Such network properties provide the rationale for combinatorial drugs that target less prominent nodes to increase synergetic efficacy and create fewer side effects.

Furthermore, if the nodes with degrees 6 to 10 are considered, 27 sparse communities with the size of not less than three are extracted. Most of them are with sizes 3 to 6, except one with size 10. Seven of the sparse communities are displayed in Fig. 6, where the red nodes represent disease genes.

IV. Conclusion

In this paper, a dynamical method was proposed to extract communities induced by low or middle-degree nodes in complex networks by constructing a local synchronization strategy of phase oscillators. The results on two small social networks validated the effectiveness of the proposed method. Then for

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

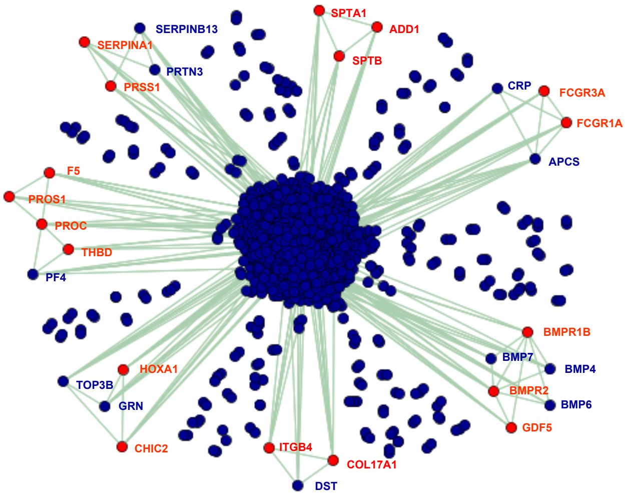

Fig. 6. The human PPI network with seven of the extracted sparse communities by using $d_1 = 6$ and $d_2 = 10$, where the red nodes represent disease genes.

the human PPI network, we investigated the possible modules to which disease genes and drug targets may belong, and some interesting sparse communities were extracted.

The main idea of the program proposed here is to extract the relatively dense subgraphs in the sparse area of the network, i.e., the area of nodes with low to middle degrees. Therefore in some cases the extracted communities may not be quite sparse in linkages. For the human PPI network, although we found some interesting sparse communities containing drug targets and disease genes, there are also some sparser communities not extracted. That is the problem worthy to be further studied. In any case, these findings in this paper may provide us a clue to find other disease genes and get possible drug targets in the near future.

Acknowledgments

This work was supported by a fund from National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, by Shanghai Natural Science Foundation under Grants No. 11ZR1443100, and by the Knowledge Innovation Program of SIBS of CAS with Grant No. 2011KIP203. Part of the authors were also supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61072149 and No. 91029301.

References

[1] Y. Zhao, E. Levinal and J. Zhu, Community extraction for social networks. *Proc. Natl. Acad. Sci. USA* 108(18), 7321-7326 (2011).

[2] K. I. Goh, M. E Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabási, The human disease network. *Proc. Natl. Acad. Sci. USA* 104(21), 8685-8690 (2007).

[3] Y. Y. Liu, J. J. Slotine and A. L. Barabasi. Controllability of complex networks. *Nature* 473(7346), 167-173 (2011).

[4] S. H. Strogatz, Exploring complex networks. *Nature* 410, 268-276 (2001).

[5] S. Boccalettia, V. Latorab, Y. Morenod, M. Chavezf and D.-U. Hwang, Complex networks: structure and dynamics. *Phys. rep.* 424(4-5), 175-308 (2006).

[6] A. Arenas, A. Diaz-Guı́lera and C. J. Pérez-Vicente, Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.* 96, 114102 (2006).

[7] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*, 2nd ed. Dover, Mineola, NY, 2003.

[8] J. A. Acebron, L. L. Bonilla, C. J. Perez Vicente, F. Ritort and R. Spigler, The Kuramoto model: a simple paradigm for synchronization phenomena. *Rev. Mod. Phys.* 77, 137-185 (2005).

[9] J. A. Almendral, I. Leyva, D. Li, I. Sendiña-Nadal, S. Havlin and S. Boccaletti, Dynamics of overlapping structures in modular networks. *Phys. Rev. E* 82, 016115 (2010).

[10] W. W. Zachary, An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* 33, 452-473 (1977).

[11] M. E. J. Newman, Detecting community structure in networks. *Eur. Phys. J. B* 38, 321-330 (2004).

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

[13] A. Henney and G. Superti-Furga, A network solution. *Nature* 455, 730-731 (2008).

[14] M. A Yildirim, K.-I. Goh, M. E Cusick, A.-L. Barabási and M. Vidal, Drug-target network. *Nat. Biotechnol.*, 25(10), 1119-1126 (2007).

[15] J.-F. Rual, et al. Toward a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173-1178 (2005).

[16] U. Stelzl, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957-968 (2005).

[17] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl. Acids Res.* 33, D514-D517 (2005).

[18] D. S. Wishart, et al. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668-D672 (2006).

[19] T. Hase, H. Tanaka, Y. Suzuki, S. Nakagawa, H. Kitano, Structure of protein interaction networks and their implications on drug design. *PLoS Comput. Biol.*, 5(10), e1000550 (2009).

978-1-4577-1666-9/11/$26.00 ©2011 IEEE