Analysis of stability of community structure across multiple hierarchical levels

Hui-Jia Li[1] and Xiang-Sun Zhang[(a)2,3]

[1] *School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100080, China.*
[2] *Academy of Mathematic and Systems Science, Chinese Academy of Science, Beijing 100190, China.*
[3] *National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China.*

PACS 89.75.Hc – First pacs description
PACS 89.75.Fb – Second pacs description

Abstract –The analysis of stability of community structure is an important problem for scientists from many fields. Here, we propose a new framework to reveal hidden properties of community structure by quantitatively analyzing the dynamics of Potts model. Specifically we model the Potts procedure of community structure detection by a Markov process, which has a clear mathematical explanation. Critical topological information regarding to multivariate spin configuration could also be inferred from the spectral significance of the Markov process. We test our framework on some example networks and find it doesn't have resolute limitation problem at all. Results have shown the model we proposed is able to uncover hierarchical structure in different scales effectively and efficiently.

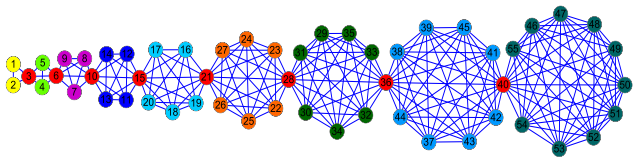
Appendix

Experiment on network with different modular sizes. – To illustrate the framework can uncover hierarchical community structures with different modular sizes, we apply the framework to a synthetic hierarchical network. The network contains 9 cliques of different sizes and we consider a line of cliques from size 3 to 11, joined only by a common node between each other. The clique network is shown in Fig.1(a) and the common nodes are presented in red color. One can consider a specific clique is an overlapping part between the neighbor ones. So 2-8 are also reasonable numbers of modules which reveal fuzzy levels of the hierarchical structure.

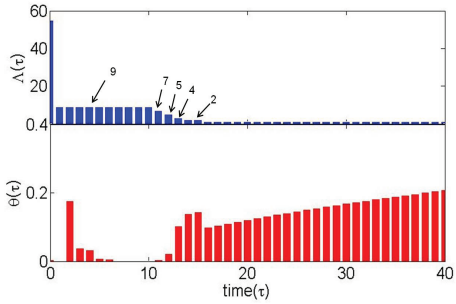
The significance of such levels can be quantified by their corresponding persistent time length. The longer the time persists, the more robust the configuration is. In the upper subgraph of Fig.1(b), one can observe that 9 modules and 2 modules are the most significant community structure. However, 3-8 are also reasonable although they don't have very long persistent timescales. This is in perfect consistence with the generation mechanisms if we consider the overlapping parts of the network. Furthermore, in the lower subgraph of Fig.1(b), we plot the curve of Θ. One can observe that the curve of Θ is a approximate parabolic shape for a specific Λ. It can be used to estimate the modularity property of complex networks, and larger Θ indicates stronger community structure. We explore the trend of stability Θ and find the largest value of Θ is corresponding to 9 communities with $\Gamma(9)$=0.187, much larger than 0.145 corresponding to 2 communities. The significance of community structure indicated by stability Θ favors finer but obvious modules. This is in keeping with the network formation and reasonable for many real networks.

Experiment on real networks. – We tested our framework on the largest connected component of a scientific collaboration network, collected by Girvan and Newman [1]. The network illustrates the research collaborations among physicists in terms of their coauthored papers posted on the Physics E-print Archive at arxiv.org which is shown in Fig.2(a). Totally, this network contains 379 nodes which are divided into 21 and 5 communities obtained by maximizing the modularity or markov endurance measure [2]. The partitions of two different s-

(a)Corresponding authors: zxs@amt.ac.cn



(a)



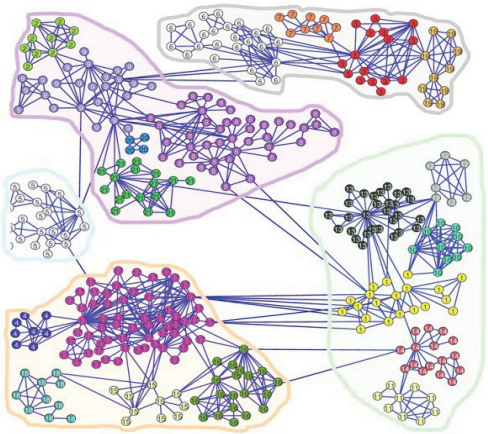
(b)

Fig. 1: (a) Structure of network contains a line of nine cliques with 3-11 nodes. The overlapping nodes are highlighted in red color. (b) The value of $\Lambda(\tau)$ and $\Theta(\tau)$ versus time τ.

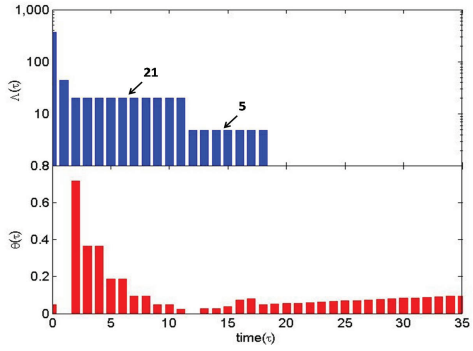
cales are corresponding to fine and coarse physical classifications. From Fig.2(b), we observe 21 and 5 communities are indeed the most significant partitions corresponding to the largest stability $\Theta(\tau)$. The result is reasonable and exactly the same as [1] and [2].

Then, we apply the framework to an important biological application,i.e. finding the communities of *S.cerevisiae* proteins based on their interactions [3]. In Fig.3(a), proteins in 10 communities are shown in different colors and annotated. These communities can be associated with either protein complexes or certain functions, which can be looked up by using the GO-Term Finder package [4] and the online tools of the Saccharomyces Genome Database (SGD) [5]. Fig.3(b) shows that 10 is indeed the optimal number of communities revealed by the corresponding largest stability $\Theta(\tau)$ in our framework. This result demonstrates that our framework can provide the real functional classifications of biological networks, which has broad applications in the future studies.

The comparison between modularity and our framework . – Finally, we emphasize the difference between the stability measure proposed and the modularity Q proposed by Newman [6]. Q is a well-known criterion for evaluating a specific partition scheme of a network. It is defined as "the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random regardless of the community structure". Different partition schemes will get different Q values for the same network, and larger ones mean better partitions. Λ and Γ try to directly characterize and evaluate the structure property which is based on network's spectra, rather than a specific network partition. Therefore, a network only has exactly self-deterministic Λ and Γ values regardless of how many partition schemes it would have, and the larger the Γ is, the stronger the network community structure is. In addition, Fortunato *et al* [7] pointed out the resolution limit problem of the modularity Q, that is, there exists an intrinsic scale beyond which small qualified communities cannot be detected by maximizing the modularity. As shown in Fig.4, when a clique ring contains cliques with different scales (i.e.,the heterogeneous community size), the intrinsic community structure can be exactly revealed by Λ. With Λ and Γ, we can quantitatively compare the modularity structure of different types of complex networks.



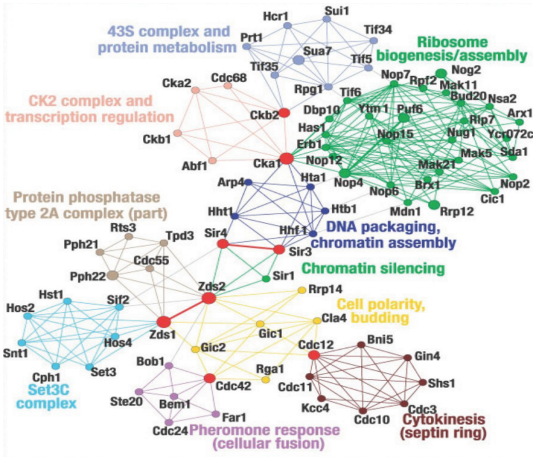
(a)



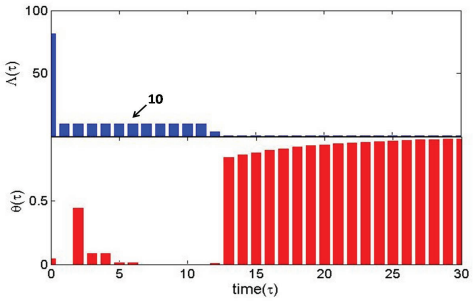
(b)

Fig. 2: (a) The largest connected component of scientific collaboration network. The nodes corresponding to 379 researchers which color coded into 21 and 5 communities obtained by maximizing the markov endurance measure [2] at $\tau = 1$ (or equivalently, modularity). The researcher are grouped mainly based on the physical classifications. (b) The value of $\Lambda(\tau)$ and $\Theta(\tau)$ versus time τ.

The relationships between our work and some famous concepts. – In [8], the authors proposed a heuristic fuzzy community detection method by minimizing an

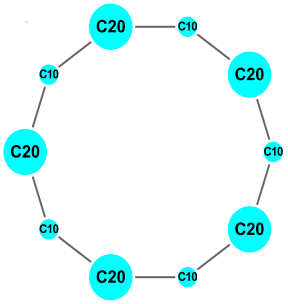


(a)

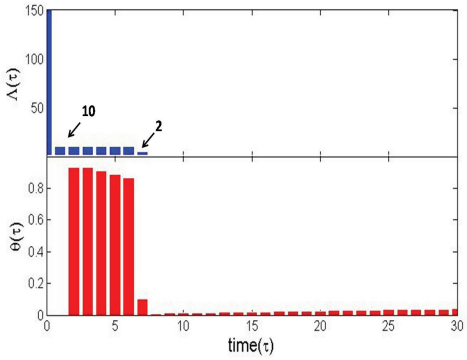


(b)

Fig. 3: (a) The protein−protein interactions network of *S.cerevisiae* containing 10 communities [3]. Different communities are described by different colors and the biological functions are annotated beside correspondingly. (b) The value of $\Lambda(\tau)$ and $\Theta(\tau)$ versus time τ.



(a)



(b)

Fig. 4: (a) Ring of clique network as a schematic example. Each circle corresponds to a clique, whose size is marked by its label C20 (contains 20 nodes) or C10 (contains 10 nodes). (b) The value of $\Lambda(\tau)$ and $\Theta(\tau)$ versus time τ.

efficient Hamiltonian function. Then, a simulated annealing algorithm is used to execute the optimization procedure. This work mainly focus on the extraction of community structure based on the objective function optimization and doesn't reveal the detail of dynamical process. Moreover, the optimal number of communities revealed by our framework can be directly used to [8] because it can't be find explicitly.

In [2] written by Delvenne et al, the authors have shown that random walk process enable one to introduce a general quality function, expressing the persistence of clusters in time. A cluster is persistent with respect to a random walk after t time steps if the probability that the walker escapes the cluster before t steps is low. Then, Delvenne et al defined the stability of the clustering aiming at, for a given time t, finding the partition with the largest value. This stability measure is a general quality function which is similar to Modularity Q [1] to some extent. It can be used to extract the optimal community partition. Compared with it, our framework tries to unveil the dynamical details of the community structure without using any particular partition. A lot of useful information, such as the optimal number of communities and dynamical changes of robustness, can be revealed directly. Thus, the objective and analytical methods between these two works are mainly different.

REFERENCES

[1] Newman.M.E.J and Girvan.M, *Phys. Rev. E*, **69** (2004) 026113.
[2] Delvenne.J.C, Yaliraki.S.N and Barahona.M, *Proc.Natl.Acad.Sci*, **107(29)** (2010) 12755-12760.
[3] Palla.G, Derényi.I, Farkas.I and Vicsek.T, *Nature*, **4635** (2005) 814-818.
[4] Boyle.E.I et al, *Bioinformatics*, **20** (2004) 3710-3715.
[5] Cherry.J.M et al, *Nature*, **433** (2005) 392-395.
[6] Newman.M.E.J, *Proc.Natl.Acad.Sci*, **103** (2006) 8577-8582.
[7] Fortunato.S and Barthelemy.M, *Proc.Natl.Acad.Sci*, **104** (2007) 36.
[8] Reichardt.J and Bornholdt.S, *Phys. Rev. Lett*, **93** (2004) 218701.