Symmetry consideration in identifying network structures

Jiao Wang\textsuperscript{1,3} and C.-H. Lai\textsuperscript{2,3}

\textsuperscript{1}Temasek Laboratories, \textsuperscript{2}Department of Physics and
\textsuperscript{3}Beijing-Hong Kong-Singapore Joint Center for Nonlinear and Complex Systems (Singapore)
National University of Singapore, 117542, Singapore

The topological information of a network can be retrieved equivalently from its complement consisting of the same nodes but complementary edges. Hence the partition of a network into certain substructures based on given criteria should be the same as that of its complement based on the equivalent criteria if the topological information is considered exclusively. This symmetry of partitioning between a network and its complement is due to the equivalence of their topological information and hence should be respected regardless of the detailed characteristics of the substructures considered. In this work we suggest this symmetry consideration as a general guideline and propose a symmetric community detection scheme to show its implications. Our method has no resolution limit and can be used to detect hierarchical community structures at different levels. Our study also suggests that the community structure is unlikely a result of random fluctuations in large networks.

PACS numbers: 89.75.Hc, 89.75.Fb, 05.45.-a

In the last decade complex networks have been extensively studied with the aim to reveal and understand their structures at various scales.\textsuperscript{[1]} Besides the general statistical properties such as small-world\textsuperscript{[2]} and scale-free\textsuperscript{[3]} properties, the significance of some common structural features at the mesoscopic level has also been realized. The mesoscopic structures having received intensive studies include communities\textsuperscript{[4]} and similar groups\textsuperscript{[5, 6, 7]}, i.e. node sets whose components have similar connection patterns. These mesoscopic structures are of scientific interest because they may have a close relation to certain behavioral or functional units of the system\textsuperscript{[8]}, and meanwhile they provide an ideal basis for reduction or coarse-graining of networks\textsuperscript{[9]}, which could be particularly useful in dealing with networks of huge size as often encountered nowadays. Furthermore, these substructures also have important implications for various dynamical processes over the networks\textsuperscript{[10, 11]}

However, in spite of the efforts and fast progress made in this field, the detection of these substructures still remains challenging. (Here we restrict ourselves to networks consisting of these mesoscopic structures exclusively, such that the problem of detection is equivalent to that of partitioning.) One conceptual difficulty is the ambiguity in the definition of these substructures\textsuperscript{[4]}. and the question of what characterizations are essential to them has not been thoroughly understood yet. In this Letter we suggest a symmetry that should be taken into account in the definition of network structures. It does not address the details of individual substructures and their characterizations, but is a property of networks. It provides a consistency criterion with which the network structures can be specified more precisely. The detection of network structures can then be improved as a result.

This symmetry originates from the dual nature of connection states in networks. Consider a network of $N$ nodes whose connection topology is encoded in the adjacency matrix $A$ with $A_{ij} = 1$ if node $i$ and $j$ are connected and $A_{ij} = 0$ otherwise. Obviously, the topological information contained in $A$ is completely equivalent to that in its complement $\overline{A}$ related to $A$ via the one-to-one map $\overline{A}_{ij} = 1 - A_{ij}$. (The network corresponding to $\overline{A}$ is referred to as the complement of the network corresponding to $A$.) Due to this equivalence, it is natural to expect that any structure recognized in network $A$ based on certain characterizations should be recognized in network $\overline{A}$ based on the same or equivalent characterizations. By way of analogy, this is similar to recognizing a face in a photo; given its features it can be done in the negative film equivalently. Assuming this equivalence, it provides an approach to check if the characterizations used for defining a structure are consistent. Only those characterizations (and their equivalent) based on which we can recognize the same structures in a network and its complement are regarded as consistent and acceptable. We suggest this symmetry principle should be adopted as a necessary condition in defining the network structures.

This symmetry consideration has not been adopted as a general guideline in most investigations. In a recent study\textsuperscript{[7]} a symmetric definition of similar groups is proposed. It has been found that indeed the symmetric definition can overcome some difficulties encountered with the asymmetric definition\textsuperscript{[6]}. Moreover, the symmetric definition can be extended to the connection information weighted networks, resulting in a new perspective to see the role of weights in the problem\textsuperscript{[7]}. It is interesting to notice that a symmetric definition of the community in a general spectral detection algorithm has also been found to outperform the asymmetric definitions\textsuperscript{[12]}

To demonstrate the power of this symmetry guideline, in the following we apply it to the community detection problem by constructing a symmetric quality function of partition. An asymmetric version, which has been so far the most popular\textsuperscript{[4]}, is that suggested by Newman and...
is the total degree of nodes in module

modularity based methods have a resolution limit

issues however remain to be addressed. One is that the

rithms have also been developed for this purpose. Some

been developed into a popular method and many algo-

networks [4]. The modularity maximization has itself

tion to the definition and detection of communities in

C

values, and the corresponding partition is regarded to

tain fairly large values when being applied to partitioning

modules

α

by

is defined as the maximum value of

A

represents how much the fraction of links inside a mod-

q

M

the maximum of

A

Girvan [13]. For a given partition \( \pi^C \) of network \( A \) that

contains \( C \) modules, it reads

\[
q_{NG}(A, \pi^C) = \sum_{\alpha=1}^{C} \left[ \frac{l_\alpha}{L} - \frac{d_\alpha}{2L} \right]^2,
\]

where \( L \) is the total number of links in the network, \( d_\alpha \)

is the total degree of nodes in module \( \alpha \) and \( l_\alpha \)

is the number of internal links of module \( \alpha \). The summand

represents how much the fraction of links inside a module

is more than what is being expected in the null model

of \( A \), i.e. random networks sharing the same nodes and

the same degree sequence. For convenience let us denote

by \( M^C_{NG} \) the maximum of \( q_{NG} \) over all the possible parti-

tions containing \( C \) modules; then the modularity, \( M_{NG} \),

is defined as the maximum value of \( M^C_{NG} \) over all allowed

C values, and the corresponding partition is regarded to be

the optimal community partition of network \( A \) [13].

The concept of modularity is an important contribu-
tion to the definition and detection of communities in

networks [4]. The modularity maximization has itself

been developed into a popular method and many algo-
rithms have also been developed for this purpose. Some

issues however remain to be addressed. One is that the

modularity based methods have a resolution limit \( \sim \sqrt{L} \)

preventing them from identifying communities smaller

than this limit [14]; another is that modularity may at-

tain fairly large values when being applied to partitioning

random networks [12], making its meaning elusive [4].

We consider here instead a symmetric quality function of

partition:

\[
q(A, \pi^C) = \frac{1}{N} \sum_{\alpha=1}^{C} \left[ \frac{d^\text{in}_\alpha}{N_\alpha} - \frac{d^\text{out}_\alpha}{N - N_\alpha} \right] \tag{2}
\]

Here \( N_\alpha \) is the number of nodes in module \( \alpha \), \( d^\text{in}_\alpha \) (\( d^\text{out}_\alpha \))

is the total degree of nodes in module \( \alpha \) correspond-

ing to their connections to themselves (other modules).

The summand reflects the difference between the average

edges a node in a module and a node outside can have

to connect to the nodes in that module. For the parti-

tion \( \pi^1 \) that all \( N \) nodes are assigned into a single module

(\( C = 1 \)), it can be naturally extended to \( q(A, \pi^1) \equiv d/N^2 \)

(\( d \) is the total degree of all nodes). Apparently, \( q \) thus

defined is symmetric; i.e. \( q(A, \pi^C) = -q(A, \pi^C) \).

Of all the possible partitions that have \( C \) modules the one, denoted by \( \tilde{\pi}^C \), that generates the maximum \( q \) value

is regarded to be the optimal community partition \textit{given}

\( C \). This is in agreement with our expectations for a good

community partition. As \( q(A, \tilde{\pi}^C) = -q(A, \pi^C) \), it sug-

gests that to find the optimal partition with \( C \) modules

in \( A \) by maximizing \( q \) can be equivalently done by mini-

mizing \( q \) in the \( \overline{A} \). For this reason \( q \) is more consistent.

We denote by \( Q^C \equiv q(A, \tilde{\pi}^C) \) for the sake of convenience.

As an example Fig. 1 shows the optimal partition \( \tilde{\pi}^C \) of the karate network [16] with

(\( C = 2, 3 \) and 4). Indeed

they are consistent with our intuition of communities. In

all other networks we have investigated this is always the

case. Numerically we employ an accurate and very effi-
cient fusion algorithm (AdClust) [17] with a slight modi-

fication. Initially each node consists of a module. At

each fusion step followed there are two operations. First,

for each node one finds the target module which mov-
ing the node into may generate a maximum positive in-
crease of \( q \) value. If this is successful then the node is

moved into that module. After all the nodes are consid-

ered (one by one and with a random order) this process

is repeated with a new random order until all nodes are

stable. Next, for all possible module pairs one finds the

one whose merger may lead to the maximum increase

(or minimum decrease) of the \( q \) value and then combine

them. These two operations are repeated until all the
modules evolve into one. During this process we can obtain a series of partitions of different numbers of modules, and they are regarded to be good approximations of the optimal partition $\hat{\pi}_C$. Careful studies have shown that different node orders taken in the first operations may lead to different partition results. For this reason $10^3 \sim 10^4$ ‘random realizations’ are performed in our calculations and the largest $q$ values and the corresponding partitions are chosen to be the final approximations of $Q_C^r$ and $\hat{\pi}_C$. We have also checked the results obtained in this way with the stimulated annealing algorithm [18] and found that they cannot be improved any further.

Next, let us find out, among all the partitions with different number of modules $\{\hat{\pi}_C, C = 1, 2, ...\}$, which one could be the most relevant. For this purpose we consider the null model, i.e. random networks that share the same degree sequence with the network considered, and define the symmetric modularity for a given $C$ as

$$M_C = \frac{Q_C - \langle Q_C^r \rangle}{\langle Q_C^r \rangle}.$$  \hspace{1cm} (3)

Here $Q_C^r$ is the maximum $q$ value for the optimal partition of a random network of null model, and $\langle Q_C^r \rangle$ is the corresponding average over all such networks. $M_C$ measures how much more modular the communities found in the original network are as compared with those found in the corresponding random networks. If the communities are seen as certain ordered structures, then $M_C$ also reflects how orderly the communities found are as compared with their counterparts arising out of pure random fluctuations. The overall modularity is thus defined as $M = \max\{M_C, C = 1, 2, ...\}$ and the corresponding partition is assumed to be the most relevant.

Fig. 2 shows the analysis of the karate network as an example. There we have considered $10^4$ random networks of the null model generated with the rewiring technique [19]. It can be seen in Fig. 2(c) and (d) that the distribution of $Q_C^r$ is perfect Gaussian, and hence can be well characterized by its average $\langle Q_C^r \rangle$ and deviation $\delta Q_C$. Meanwhile $\langle Q_C^r \rangle$ is a function of $C$ (Fig. 2(a)); this is the reason why it is introduced as the denominator in the definition of $M_C$. The results of $M_C$ (Fig. 2(b)) suggest that the partition of three communities ($C = 3$; see Fig. 1 for the partition) is the most relevant. We have also studied the dolphin network [20] and the most relevant partition ($C = 2, M = 0.6258$) is found to be exactly the same as the natural split observed [21]. For another popular testing network of the American college football teams [22] our method suggests the partition of 10 communities ($C = 10, M = 1.3345$).

The fact that the distribution of $Q_C^r$ is Gaussian allows us to define another useful quantity

$$F_C^r = \frac{\delta Q_C}{\langle Q_C^r \rangle}.$$  \hspace{1cm} (4)

which gives how ‘modular’ a random network (of the null model) can be as a result of fluctuations. Obviously only the partitions of the original network whose $M_C \gg F_C^r$ may suggest meaningful community structures (see Fig. 2(b) for a comparison of $M_C$ and $F_C^r$ in the karate network). This should be seen as a necessary condition for the communities defined with the symmetric modularity and it concludes our community detection scheme.

Now let us discuss two useful properties of the symmetric modularity. First, the community detection method based on it has no resolution limit. As an example we consider a network of $N$ cliques sited on a circle. Each clique contains 3 nodes – the smallest size for a meaningful module – and any two neighboring cliques are linked with one edge. Our scheme can identify all cliques (for $N \geq 2$) without any ambiguity (see Fig. 3 for $N = 30$ as an example). In this simulation (and also in those for Fig. 4 and Fig. 5) $\langle Q_C^r \rangle$ and $F_C^r$ are evaluated over $10^3$ random networks with the same degree sequence. As a comparison, the method with the asymmetric modularity $M_{NG}$ suggests instead the partition of 10 communities each containing 3 neighboring cliques ($M_{NG} = 49/60$) due to its inherent resolution limit. ($M_{NG} = 97/120$ and 43/60 for $C = 15$ and $C = 30$ in this case.)

This high resolution even makes our method applicable to the hierarchical community networks – a challenge for the quality function method due to the multiple scales involved. In Fig. 4 we present the partition results for the model hierarchical network suggested in [11]: 256 nodes are divided into 16 compartments of equal size at the first level and every 4 of them make a bigger compartment at the second level. The internal degree of nodes at the first (second) level $z_{in1}$ ($z_{in2}$) and the degree for the links between the second level communities keep an average of $z_{in1} + z_{in2} + z_{out} = 18$ (hence the hierarchical
creasing roughly in a power law

\[ N \sim r \]

be seen that and the corresponding deviation (see Fig. 5(a)). It can sequences for each \( F \) or this reason we have considered eight different degree

levels can be indicated by \( (z_{in1} - z_{in2}) \). We find that the hierarchical structures are well characterized by the local maxima (also sharp turning points) on the \( M^C \) curve. The change of \( M^C \) values at the maxima from (a) to (c) reflects the competition of the two scales.

Second, the symmetric modularity does not take large value for a random network. Careful studies of Erdös-Rényi (ER) and Barabasi-Albertscale (BA) scale-free networks [3] are summarized in Fig. 5. For an ER network with \( N \) nodes and connection probability \( p \) studied there, we have verified that \( M^C \) is around zero and \( |M^C| \sim F^C \) as implied by definition. Meanwhile, the data suggest that \( F^C \) may depend on the degree sequence, but always takes the maximum value at \( C = 2 \).

For this reason we have considered eight different degree sequences for each \( N \), \( p \) pair and calculated their average and the corresponding deviation (see Fig. 5(a)). It can be seen that \( \langle F^C \rangle \) is small and does not depend on \( p \) significantly; more important as \( N \) is increased it keeps decreasing roughly in a power law \( \sim N^{-0.75 \pm 0.05} \). This suggests that the community structure cannot be a general property in ER networks. In addition, the dependence of \( F^C \) on the degree sequence is very weak \( (\delta F^C < 0.05) \), suggesting the chance for finding meaningful community structure in certain realizations of ER networks of particular degree sequences is also very slim. The study of the scale-free networks leads to the same results except that the power law dependence of \( \langle F^C \rangle \) on the network size is roughly \( \sim N^{-0.46 \pm 0.04} \) instead.

In summary, we suggest the equivalence between the topological information of a network and its complement should be considered generally in the definition and detection of network structures. As an important application we have focused on the community partition problem and proposed a symmetric quality function. The resulted community detecting scheme has a high resolution and can be used to identify hierarchical community structures. In addition, we have found that the effects of fluctuations on the community structure are weak and decrease as the size of network increases. This implies that the community structure is unlikely a result of fluctuations when the size of the network is large enough. The question of whether there are other relevant symmetries and how they may provide insights into network structures is interesting and deserves further efforts.

This work is supported by Defense Science and Technology Agency (DSTA) of Singapore under agreement POD0613356.

---

[1] R. Albert and A.L. Barabasi, Rev. Mod. Phys. 74, 47 (2002); M.E.J. Newman, SIAM Rev. 45, 167 (2003).
[2] D.J. Watts and S.H. Strogatz, Nature 393, 440 (1998).
[3] A.L. Barabasi and R. Albert, Science 286, 509 (1999).
[4] S. Fortunato and C. Castellano, arXiv: physics.soc-ph/0712.2716 (2007).
[5] M. Rosvall and C.T. Bergstrom, PNAS 104, 7327 (2007).
[6] M.E.J. Newman and E.A. Leicht, PNAS 104, 9564 (2007).
[7] J. Wang and C.-H. Lai, New J. Phys. 10, 123023 (2008).
[8] S. Shen-Orr, R. Milo, S. Mangan and U. Alon, Nature Genetics 31, 64 (2002); M.A. Porter, P.J. Mucha, M.E.J. Newman and C.M. Warmbrond, PNAS 102, 7057 (2005).
[9] D. Gfeller and P. De Los Rios P, Phys. Rev. Lett. 99, 038701 (2007); A. Arenas, J. Duch, A. Fernández and S. Gómez, New J. Phys. 9, 176 (2007).
[10] H. Zhou, Phys. Rev. E 67, 041905 (2003); K.A. Eriksen, I. Simonsen, S. Maslov and K. Sneppen, Phys. Rev. Lett.
[90] 148701 (2003); L. Huang, K. Park, Y.C. Lai, L. Yang and K.Q. Yang, Phys. Rev. Lett. 97, 164101 (2006); F. Sorrentino and E. Ott, Phys. Rev. E 76, 056114 (2007).
[11] A. Arenas, A. Díaz-Guilera and J. Pérez-Vicente, Phys. Rev. Lett. 96 114102 (2006).
[12] M. Zarei and K.A. Samani, Physica A 388, 1721 (2009).
[13] M.E.J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).
[14] S. Fortunato and M. Barthélemy, PNAS 104, 36 (2007).
[15] R. Guimerà, M. Sales-Pardo and L.A.N. Amaral, Phys. Rev. E 70, 025101(R) (2004).
[16] W.W. Zachary, J. Anthropol. Res. 33 452 (1977).
[17] Z. Ye, S. Hu and J. Yu, Phys. Rev. E 78, 046115 (2008).
[18] R. Guimerà, and L.A.N. Amaral, Nature 433, 895 (2005); S. Kirkpatrick, C.D. Gelatt, Jr., M.P. Vecchi, Science 220, 671 (1983).
[19] S. Maslov, K. Sneppen and A. Zaliznyak, Physica A 333, 529 (2004); R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman and U. Alon, arXiv: cond-mat/0312028 (2003).
[20] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten and S.M. Dawson, Behav. Ecol. Sociobiol 54, 396 (2003).
[21] M.E.J. Newman, Phys. Rev. E 74 036104 (2006).
[22] M. Girvan and M.E.J. Newman, PNAS 99 7821 (2003).