

Community Identification of Complex Network

Xiang-Sun Zhang

<http://zhangroup.aporc.org>

Chinese Academy of Sciences

2008.10.31, OSB2008

Outline

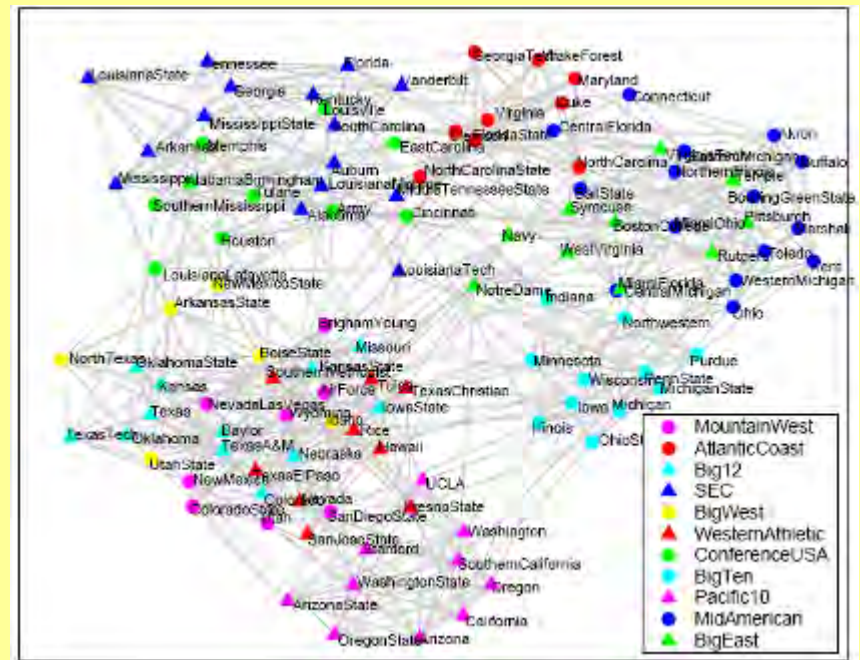
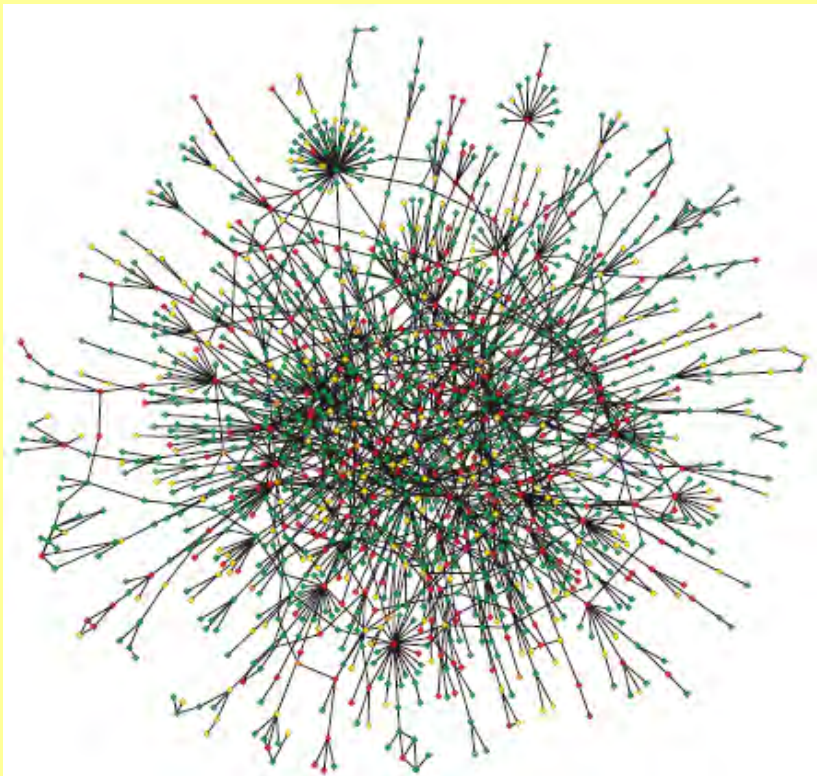
- Background
- Community identification definition
- Community identification methods
- Modularity measures for network community
- Conclusion

Complex networks

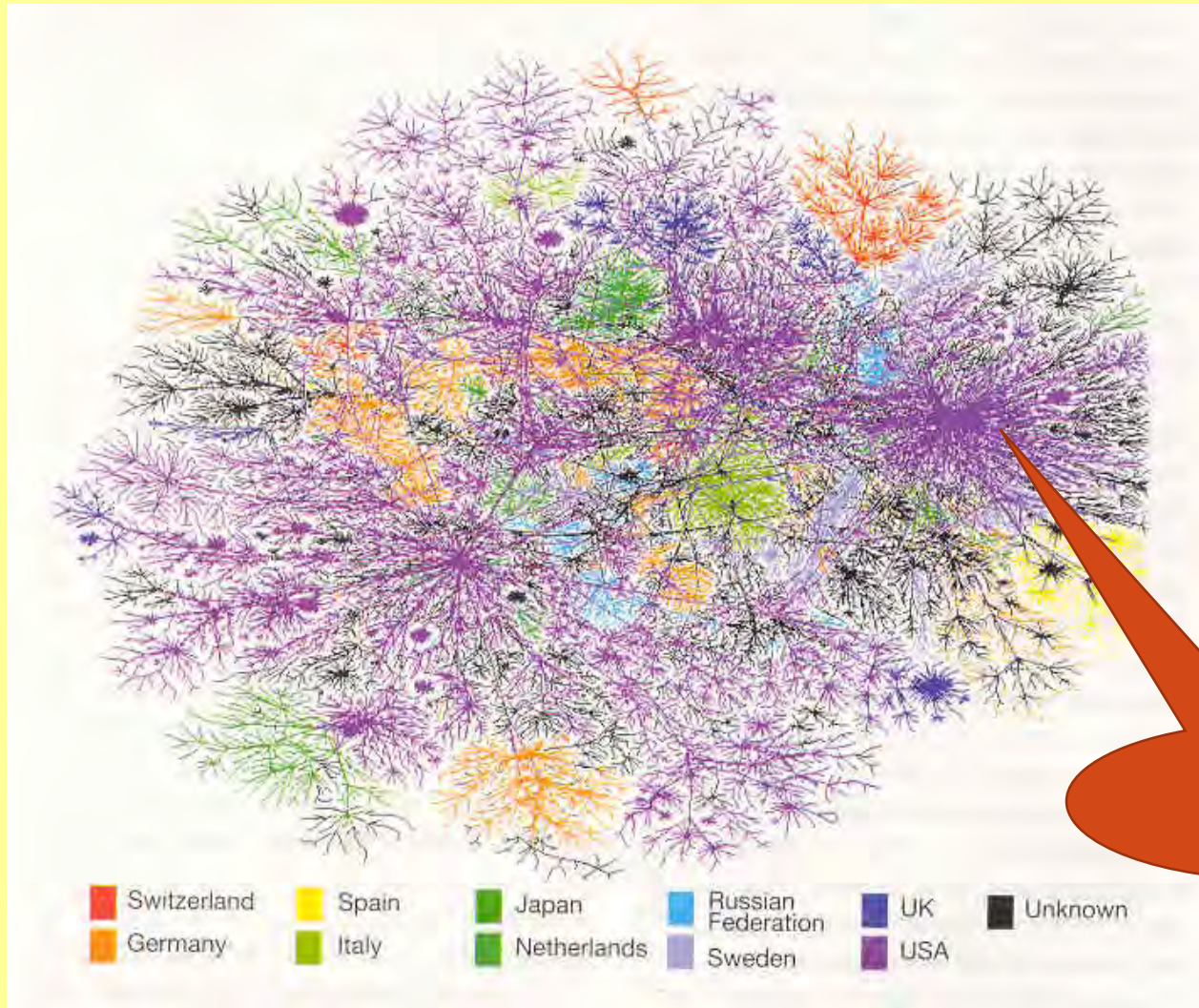
- Many systems can be expressed by a **network**, in which nodes represent the objects and edges denotes the relations between them.
- **Social networks** such as scientific collaboration network, food network, transport network, etc.
- **Technological networks** such as web network, software dependency network, IP address network, etc.
- **Biological networks** such as protein interaction networks, metabolic networks, gene regulatory networks, etc.
- ...

Examples

- Yeast protein interaction network (A.-L. Barabási, NATURE REVIEWS GENETICS, 2004)
- Football team network (S. White, P. Smyth, SIAM conference, 2004)



Computer IP address network



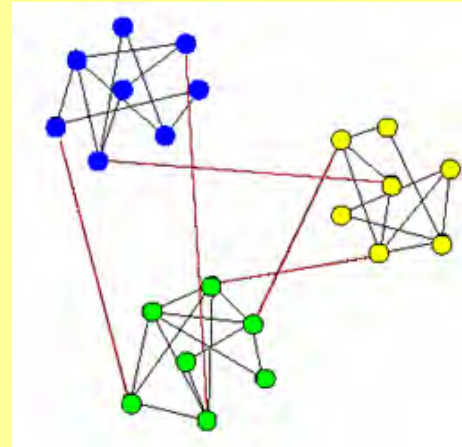
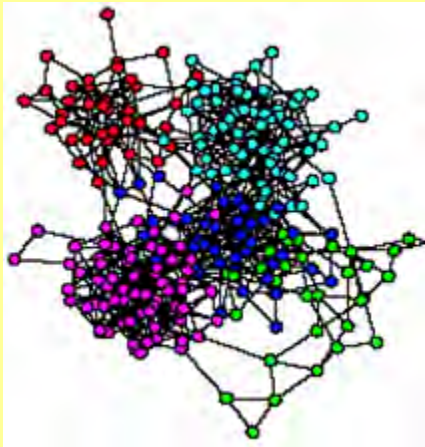
USA

Common topological properties

- **small-world property**: most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of steps
- **scale-free property**: degree distribution follows a power law, at least asymptotically. That is, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the fraction of nodes in the network having k connections to other and γ is a constant.
- ...

Modularity/Community structure

- **Modularity/Community structure** : common to many complex networks. It means that complex networks consist of groups of nodes **within** which the connection is **dense** but **between** which the connection is relatively **sparse**.



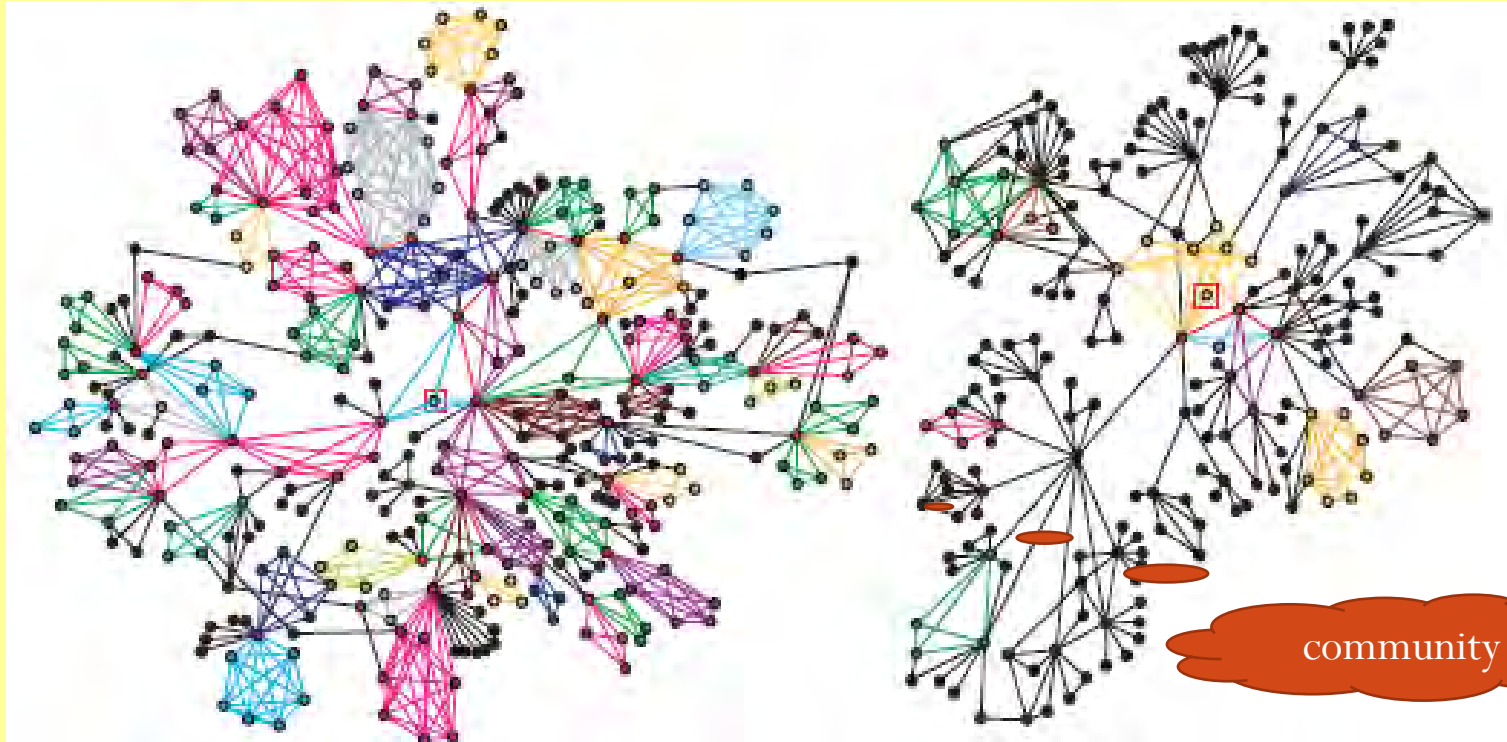
Community structure

- Nodes in a same tight-knit community tend to have common properties or attributes
- **Modules/communities** in biological networks or other types of networks usually have functional meaning

Community identification

- Identifying community structure of a complex network is fundamental for uncovering the relationships between sub-structure and function of the network.
- In biological network research, it is widely believed that the modular structures are formed from the long evolutionary process and corresponds to biological functions.

Community of complex networks

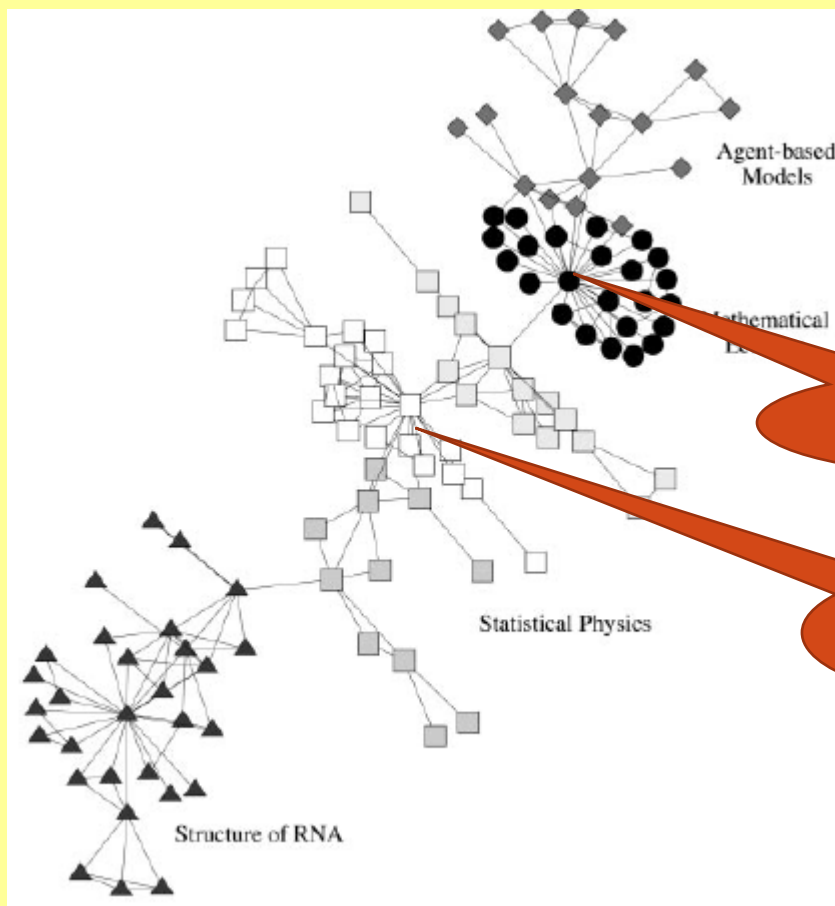


Paper-cooperation network

Phone network

Significance of community structure

- Common functions of many complex networks
- Global network structure and function decomposition



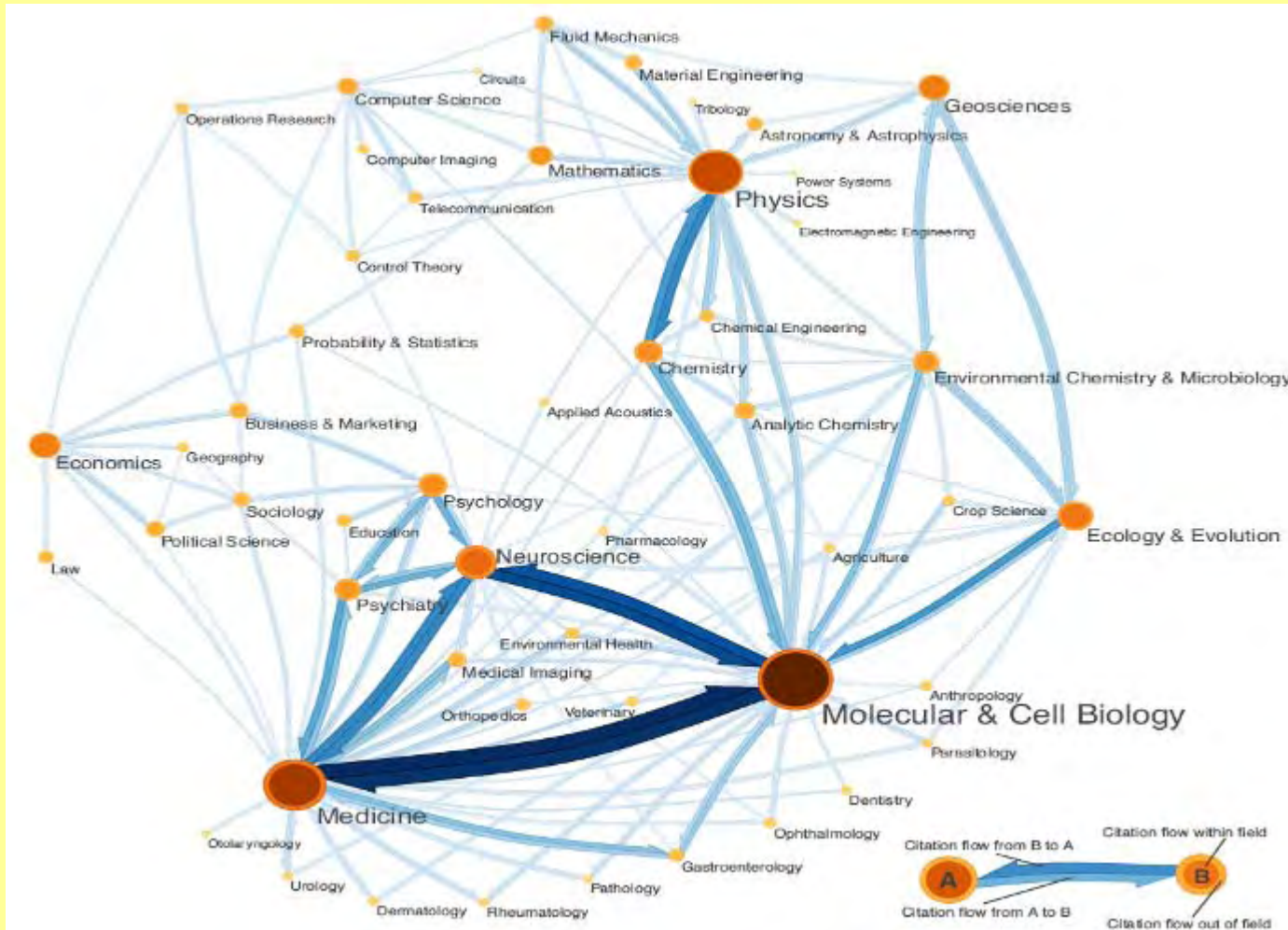
The scientific collaboration network in The Santa Fe Institute: the module denotes the groups of scientists in similar research field.

Mathematical ecology

Statistical physics

Martin Rosvall, Carl T. Bergstrom,
PNAS, vol. 105, no.4. 1118-1123,
2007

A network of science based on citation
patterns: **6,128 journals connected by**
6,434,916 citations.



The network is partitioned into 88 modules and 3,024 directed and weighted links, which represent a trace of the scientific activity.

Community identification definition

- Given a network/graph $N = (V, E)$, partition N into several subnetworks which satisfy community conditions
- In complex network research, a popular qualitative community definition is

The nodes in a community are densely linked but nodes in different communities are sparsely linked

Filippo Radicchi et. al. *Proc. Natl. Acad. Sci. USA (PNAS)*, Vol.101, No.9, 2658-2663, 2004

Community detection methods

- Some methods are based on topological properties of nodes or edges such as betweenness-based methods (Girvan, Newman, PNAS, 2002)
- Some of them are clustering-based, e.g. various spectral clustering algorithms (S. White, P. Smyth, SIAM conference, 2004)

Community detection methods

- In Newman and Girvan, *PRE*, 2004, a modularity function Q was proposed as following

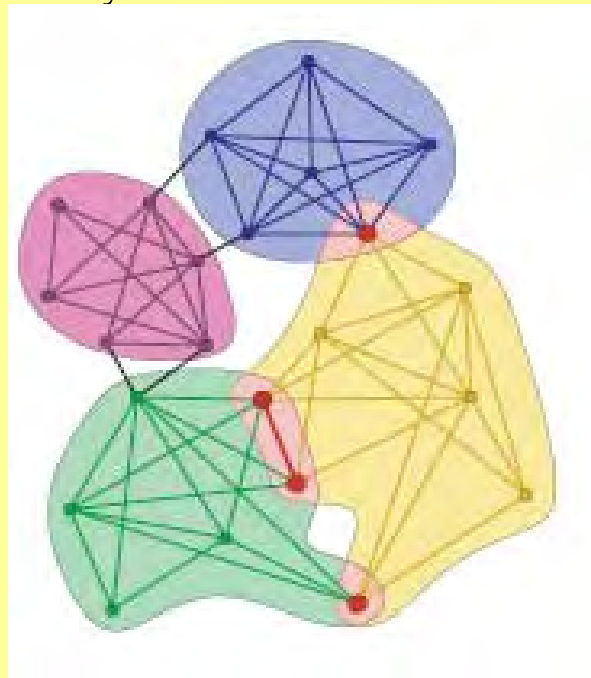
$$Q(P_k) = \sum_{c=1}^k \left[\frac{L(V_c, V_c)}{L(V, V)} - \left(\frac{L(V_c, V)}{L(V, V)} \right)^2 \right]$$

to measure the community structure of a network.

- A class of methods maximizing modularity Q appear. Heuristic algorithms such as Simulated Annealing, Genetic Algorithms, Local Search, etc. [Newman, *PNAS*, 2006; Guimera, *Nature*, 2005].

Overlapping/fuzzy communities

- In Palla et al., *Nature*, 2005, a clique-percolation method was proposed for community detection



- In Reichardt, Bornholdt, *PRL*, 2004, a Potts model was used for detecting fuzzy structure

Our work (I will not focus on)

- **Shihua Zhang, Rui-Sheng Wang, and Xiang-Sun Zhang.** Identification of Overlapping Community Structure in Complex Networks Using Fuzzy c-means Clustering. *Physica A*, 2007, 374, 483–490.
- **Shihua Zhang, Rui-Sheng Wang and Xiang-Sun Zhang.** Uncovering fuzzy community structure in complex networks. *Physical Review E*, 76, 046103, 2007
- **Rui-Sheng Wang, Shihua Zhang, Yong Wang, Xiang-Sun Zhang, Luonan Chen.** Clustering complex networks and biological networks by Nonnegative Matrix Factorization with various similarity measures. *Neurocomputing*, DOI: 10.1016/j.neucom.2007.12.043
- ...

Mathematical community definition

- Mathematically, let

$$d_i = d_i^{in} + d_i^{out}$$

then the condition for a subnetwork $N_k = (V_k, E_k)$ being a community is

$$\sum_{i \in V_k} d_i^{in} - \sum_{i \in V_k} d_i^{out} > 0$$

i.e.

$$2|E_k| - |\overline{E}_k| > 0$$

where \overline{E}_k is all edges linking V_k and $V \setminus V_k$

**Filippo Radicchi et. al. *Proc. Natl. Acad. Sci. USA (PNAS)*,
Vol.101, No.9, 2658-2663, 2004**

- A popular method to partition a network into community structure is to optimize a quantity called **modularity**, or some alternatives, which is a measure for a given partition.
- **Modularity definition** and **modularity optimization** are still in the state-in-art process.

Modularity function Q

- Newman and Girvan (*Physical Review E*, 2004) gives a quantitative measure Q

$$Q(N_1, \dots, N_k) = \sum_{i=1}^k \left[\frac{|E_i|}{|E|} - \left(\frac{d_i}{2|E|} \right)^2 \right]$$

where N_1, \dots, N_k is a partition of N . We can prove

$$2|E_i| - |\bar{E}_i| > 0 \Rightarrow Q(N_1, \dots, N_k) > 0$$

- But it is not necessary that

$$Q(N_1, \dots, N_k) > 0 \Rightarrow 2|E_i| - |\bar{E}_i| > 0$$

- It suggests that partition N into N_1, \dots, N_k such that $Q(N_1, \dots, N_k)$ is as large as possible to make sure that

$$2|E_i| - |\bar{E}_i| > 0$$

which leads to an optimization process below

- Step 1: Fix k ($k = 1, \dots, n$), $N_1 \cup \dots \cup N_k = N$,
compute

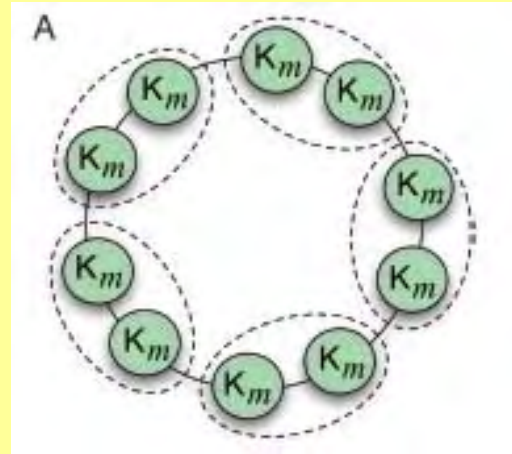
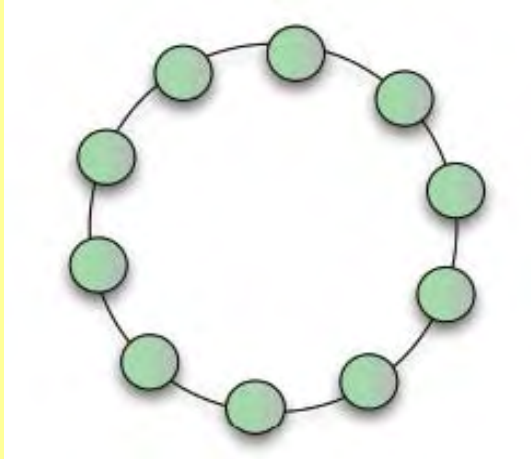
$$\max_{N_1, \dots, N_k} Q(N_1, \dots, N_k)$$

- Step 2: Compute

$$\max_{k \in \{1, \dots, n\}} \max_{N_1, \dots, N_k} Q(N_1, \dots, N_k)$$

This is an **enumeration** algorithm, then heuristic algorithms including simulation annealing, genetic algorithm are generally used (Newman, **PNAS**, 2006; Guimera, **Nature**, 2005).

Modularity Q fails to identify correct community structure in some cases

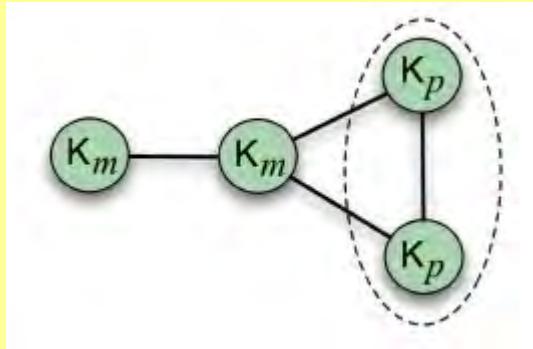


Left: a graph consists of **a ring of cliques** connected by single links, each clique is a qualified community.

Right: when the number of cliques is larger than about \sqrt{N} , the modularity optimization gives a partition where two cliques are combined into one community! This phenomena is called **resolution limit**.

Fortunato & Barthelemy, *Proc. Natl. Acad. Sci.* (2007)

Modularity Q fails to identify correct community structure in some cases



a graph consists of **four cliques** with different size, each clique is a qualified community.

when the clique size are quite heterogeneous, i.e. $p \ll m$, the modularity optimization gives a partition where two small cliques are combined into one community!

We suggested a new quantitative measure

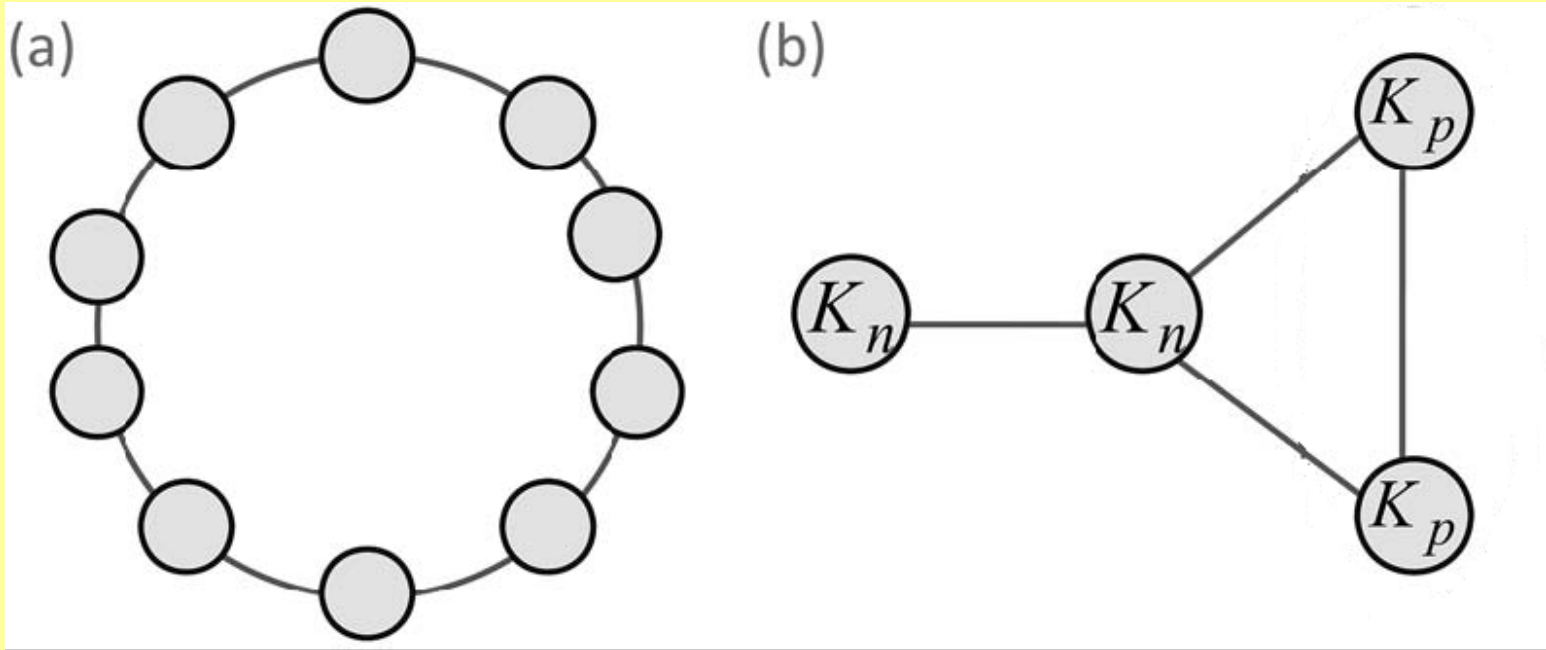
- Modularity Density D :

$$D(N_1, \dots, N_k) = \sum_{i=1}^k \left[\frac{2|E_i|}{|V_i|} - \frac{|\bar{E}_i|}{|V_i|} \right]$$

which obviously has property:

$$2|E_i| - |\bar{E}_i| > 0 \Rightarrow D(N_1, \dots, N_k) > 0$$

Zhenping Li, Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang,
Luonan Chen, Quantitative function for community detection.
***Physical Review E*, 77, 036109, 2008**



Modularity density D overcomes “resolution limit” problem in the cases of the **ring of L cliques and the **network with heterogeneous clique size****

Experiment Result

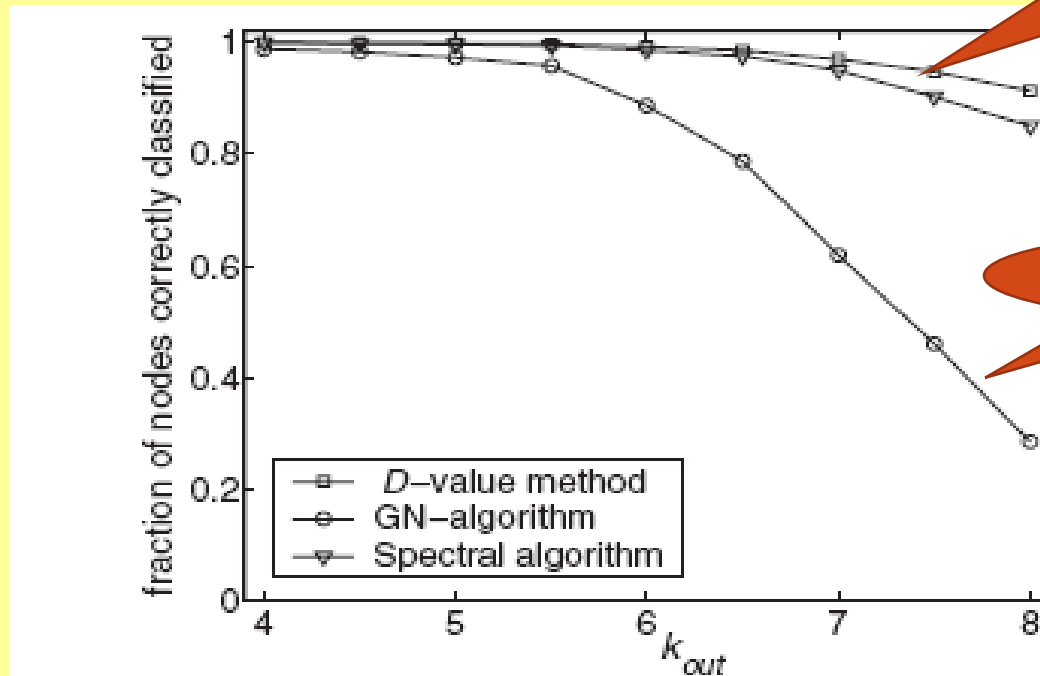


FIG. 2. Test of various methods on computer-generated networks with known community structures. It is a plot of the fraction of nodes correctly classified with respect to k_{out} . Each point is an average over 100 realizations of the networks.

Problem remained

- Fortunato & Barthelemy, *PNAS* (2007), analyzed the “resolution limit” numerically based on some special network structures.
- Zhenping Li etc, *Physical Review E* (2008), suggested a new measure D and compare the modularity density D and modularity Q based on special network structures and numerical examples.
- A theoretical/mathematical framework to evaluate the different measures and display community structure properties is needed.

A closed optimization model based on the modularity Q

- Given a network $N = (V, E)$, $V = (v_1, \dots, v_n)$, let (e_{ij}) be the adjacency matrix. Suppose that N is partitioned into k parts N_1, \dots, N_k . Use binary integer variable x_{ij} :

$$x_{ij} = \begin{cases} 1 & \text{if node } i \text{ is in community } j \\ 0 & \text{otherwise} \end{cases}$$

The community definition then can be expressed as

$$\sum_{s,t \in V} e_{st} x_{sj} x_{tj} \geq \sum_{s,t \in V} e_{st} x_{sj} (1 - x_{tj})$$

For $j=1,2,\dots, k$

Optimization model based on Q

- A nonlinear integer programming based on Q

$$\begin{aligned} \max \quad & \sum_{j=1}^k \left[\frac{\sum_{s,t \in V} e_{st} x_{sj} x_{tj}}{\sum_{(s,t) \in E} e_{st}} - \left(\frac{\sum_{s,t \in V} e_{st} x_{sj}}{\sum_{(s,t) \in E} e_{st}} \right)^2 \right] \\ \text{s.t.} \quad & \sum_{j=1}^k x_{ij} = 1, \quad i = 1, \dots, n \\ & x_{ij} \in \{0, 1\}, \quad i = 1, \dots, n, \quad j = 1, \dots, k \end{aligned}$$

Xiang-Sun Zhang and Rui-Sheng Wang, Optimization analysis of modularity measures for network community detection, **OSB 2008**.

Optimization model based on D

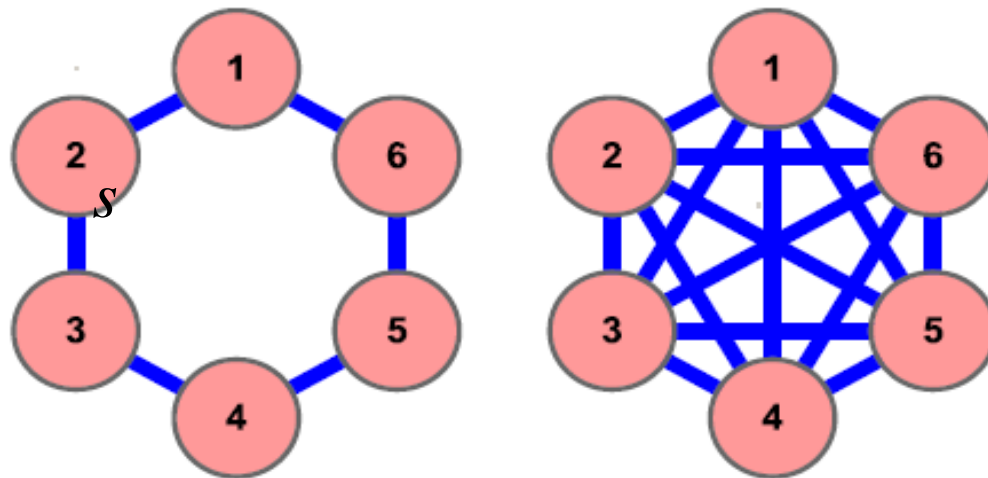
- A nonlinear integer programming based on D

$$\begin{aligned} \max \quad & \sum_{j=1}^k \left[\frac{\sum_{s,t \in V} e_{st} x_{sj} x_{tj}}{\sum_{t \in V} x_{tj}} - \frac{\sum_{s,t \in V} e_{st} x_{sj} (1-x_{tj})}{\sum_{t \in V} x_{tj}} \right] \\ \text{s.t.} \quad & \sum_{j=1}^k x_{ij} = 1, \quad i = 1, \dots, n, \\ & x_{ij} \in \{0, 1\}, \quad i = 1, \dots, n, \quad j = 1, \dots, k \end{aligned}$$

Xiang-Sun Zhang and Rui-Sheng Wang, Optimization analysis of modularity measures for network community detection, **OSB 2008**.

Convex analysis based some special structures

The following two exemplar networks are used in almost all *PNAS* papers that discuss the community identification



(a) The ring of lumps

(b) The ad hoc network

Figure 1: Diagrams of two exemplary networks.

A ring of dense lumps whose adjacency matrix is:

$$A^L = \begin{pmatrix} A & M & 0 & \cdot & \cdot & 0 & 0 & M \\ M & A & M & \cdot & \cdot & 0 & 0 & 0 \\ 0 & M & A & \cdot & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & A & M & 0 \\ 0 & 0 & 0 & \cdot & \cdot & M & A & M \\ M & 0 & 0 & \cdot & \cdot & 0 & M & A \end{pmatrix}$$

where $L > 4$, A is an $m \times m$ adjacency matrix to represent a connected subnetwork called as lump, then A^L is an $Lm \times Lm$ matrix, M stands for a random matrix with s non-zero elements. Note that these random matrices don't have to be identical, provided that they have the same number of non-zero elements .

The second exemplary network is a special version of the *ad hoc network* (a computer-generated network). Its adjacency matrix takes the form:

$$A^L = \begin{pmatrix} A & M & M & \cdot & \cdot & M & M & M \\ M & A & M & \cdot & \cdot & M & M & M \\ M & M & A & \cdot & \cdot & M & M & M \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ M & M & M & \cdot & \cdot & A & M & M \\ M & M & M & \cdot & \cdot & M & A & M \\ M & M & M & \cdot & \cdot & M & M & A \end{pmatrix}.$$

Denote a partition as $P = \{V_1, \dots, V_k\}$, the optimization process can be written as a two-stage optimization problem:

$$Q_p : \max_k \bar{Q}(k) = \max_k \max_{\sum_{i=1}^k |V_i|=n} Q(V_1, V_2, \dots, V_k);$$

$$D_p : \max_k \bar{D}(k) = \max_k \max_{\sum_{i=1}^k |V_i|=n} D(V_1, V_2, \dots, V_k);$$

We denote $\bar{Q}(k)$ and $\bar{D}(k)$ as the solutions from the first-step optimization problems: with a fixed k , partition the whole network into k subnetworks $N_1 = (V_1, E_1), \dots, N_k = (V_k, E_k)$ to maximize the quantitative functions Q and D . And $\max_k \bar{Q}(k)$ and $\max_k \bar{D}(k)$ are the second-step optimization problems.

Convex Analysis

- A function (or a programming) whose variables take discrete values (or, say, the sample values) is called as **discrete convex (concave) function (or programming)** if they can be embedded into a continuous convex (concave) function (or programming).
- **Result 1 :** For the **ring of A**,

$\max_{\sum_{i=1}^k |V_i|=n} Q(V_1, V_2, \dots, V_k)$ is a discrete concave programming

$\max_{\sum_{i=1}^k |V_i|=n} D(V_1, V_2, \dots, V_k)$ is a discrete concave programming

$\bar{Q}(k)$ is a discrete convex function

$\bar{D}(k)$ is a discrete convex function

Convex Analysis (continued)

- **Result 2** : For the *ad hoc* network,

$\max_{\sum_{i=1}^k |V_i|=n} Q(V_1, V_2, \dots, V_k)$ is a discrete concave programming

$\max_{\sum_{i=1}^k |V_i|=n} D(V_1, V_2, \dots, V_k)$ is a linear programming

$\bar{Q}(k)$ is a discrete convex function

$\bar{D}(k)$ is a linear function

Above analysis makes it possible that we solve the two exemplar networks analytically, then compare Q and D analytically.

Convex Analysis (continued)

Result 3 :

- for the ring of A where each A is the smallest community (known community), the modularity density model D can identify the known communities. But the modularity model Q fails if

$$s > \frac{|E|}{L-1}$$

which extends the result in Fortunato & Barthelemy, *Proc. Natl. Acad. Sci.* (2007) where s takes 1.

Further research in community identification

- The closed formulation of the Q and D optimization allows to design more efficient algorithm to solve the community identification problem
- Based on the comparison of Q and D , present new measures that exactly reflect the community definition
- Consider modularity measures in directed networks

Thanks

Welcome to visit us at

<http://zhangroup.aporc.org>