RESOLUTION PROBLEMS IN COMMUNITY DETECTION

János Kertész
Institute of Physics, Budapest University of Technology and Economics
and HUT

with

J Kumpula, J. Saramäki, K. Kaski (HUT) and A. Lancichinetti, S. Fortunato (ISI)

OUTINE

- Community detection: Global vs local
- Resolution limit in Newman Girvan modularity scheme (Fortunato and Barthelemy)
- Resolution limit in the Potts model approach, different null models
- How to avoid resolution problems? Multiresolution methods
- Relation to the problem of hierarchical community organization
- Multiresolution in local approaches
- Summary

COMMUNITY DETECTION: LOCAL VS GLOBAL

Communities: Vague definitions, „more inside links than outside ones”
-- definitions by algorithms

Natural approach: local algorithms (many)
Clauset (local modularity)
Luo, Wang, Promislow (weak community)
Bagrow (outwardness)
…
hierarchical clustering
…

Palla, Derényi, Farkas, Vicsek (clique percolation)
Lancichinetti, Fortunato, JK (node fitness)
…

Module identification often ambiguous
mostly considered as a partition problem } optimization

Global optimization

• Newman-Girvan modularity (quality function → optimization tool):

$$Q = \frac{1}{2L} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2L} \right) \delta(C_i, C_j)$$

L: total number of links

Null model: Here configurational model

• Generalization by Reichardt and Bornholdt: Find ground state of a Potts model (p_{ij} = prob. of link in the null model)

$$\mathcal{H} = -\sum_{i \neq j} \left(A_{ij} - \gamma p_{ij} \right) \delta(\sigma_i, \sigma_j)$$

RESOLUTION LIMIT IN THE N-G METHOD (Fortunato-Barthelemy)

Small, plausible communities cannot be identified if the network is large

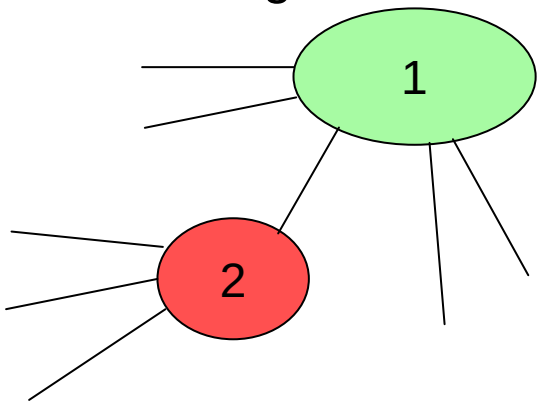
Rewrite *Q* as $$Q = \sum_{s=1}^{m} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right]$$ (same null model)

where l_s is # links inside module *s*
d_s is the total degree in module *s*
m is # modules

When is it worth considering two connected communities as a single one?

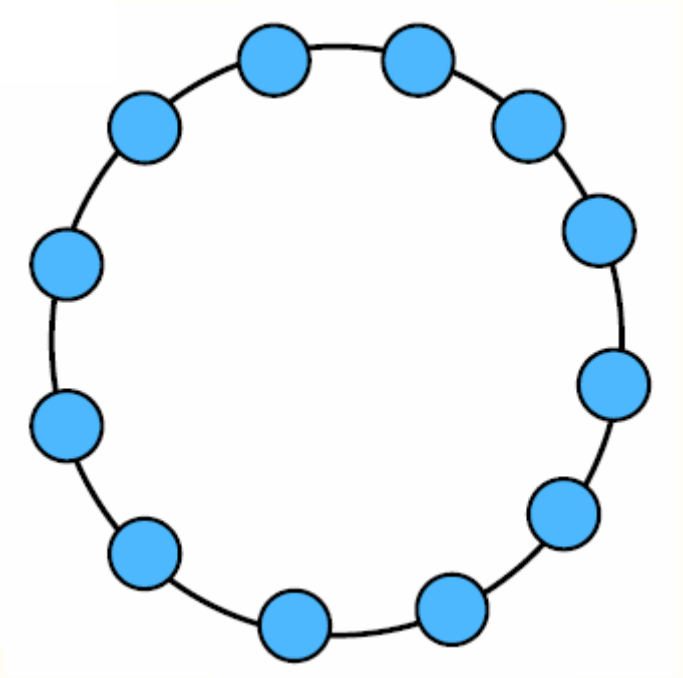
$$\Delta Q = \frac{l_1 + l_2 + l_{12}}{L} - \frac{(d_1 + d_2)^2}{(2L)^2} > \frac{l_1}{L} + \frac{l_2}{L} - \frac{(d_1)^2 + (d_2)^2}{(2L)^2}$$

$$\Delta Q = \frac{l_{12}}{L} - \frac{2 d_1 d_2}{(2L)^2} > 0 \qquad L > d_1 d_2$$



Even if the small communities are cliques and a single link connects them….

d_s is characteristic of the module size. Assuming equal size moduls we conclude that $d_s < \sqrt{L}$ size modules cannot be seen by the N-G method. This is a resolution limit.



Circles symbolize a m-cliques. Just by changing the length of the chain (changing *L*) it may be worth considering pairs of cliques as single communities. This is unphysical and shows the limitation of the global optimization method.

RESOLUTION LIMIT IN THE POTTS MODEL APPROACH

(Kumpula et al. I)

$$\mathcal{H} = -\sum_{i \neq j} (A_{ij} - \gamma p_{ij})\, \delta(\sigma_i, \sigma_j)$$

This is equivalent to the modularity with

$$Q = -\mathcal{H} / L, \qquad \gamma = 1, \qquad p_{ij} = k_i k_j / 2L$$

Thus this method enables to study the role of the null model and of the coupling constant γ.

$\gamma \rightarrow 0$: plain Potts ground state, i.e., all nodes in a single community

γ **>>** 1: communities break into small pieces because penalty for missing links is very high

γ **>** **1/**min(p_{ij}): each node is a separate community

What is the optimum number of communities for a chain of cliques, if *N* nodes *L* links and γ are given?

The Hamiltonian can be rewritten as:

$$\mathcal{H} = -\sum_{s=1}^{n} \left(l^s - \gamma [l]^s_{p_{ij}} \right)$$

links in module *s*

expected # links in module *s* in the null model

For $p_{ij} = \frac{1}{2L} k_i k_j$ (config. Model) $[l]^s_{p_{ij}} = \frac{1}{4L} d_s^2$

In the chain model each module has $L/n - 1$ links, leading to

$$\mathcal{H}_{min}(n, \gamma, L) = -(L - n - \gamma \frac{L}{n})$$

Looking at the optimum number of modules $d\mathcal{H}_{min}(n, \gamma, L)/dn$ = 0 has to be taken, leading to $n^* = \sqrt{\gamma L}$ With the ER null model $n^* = \sqrt{\gamma L \frac{N}{N-1}}$

The case of a general null model:

In the same spirit as before, we calculate when it is worth combining two communities. The criterion is:

$$\Delta E = E_2 - E_1 = -l^{s\leftrightarrow r} + \gamma \left([l]_{p_{ij}}^{s+r} - [l]_{p_{ij}}^{r} - [l]_{p_{ij}}^{s} \right) < 0$$

Since $[l]_{p_{ij}}^{s+r} - [l]_{p_{ij}}^{r} - [l]_{p_{ij}}^{s} \equiv [l]_{p_{ij}}^{s\leftrightarrow r}$

The criterion reduces to

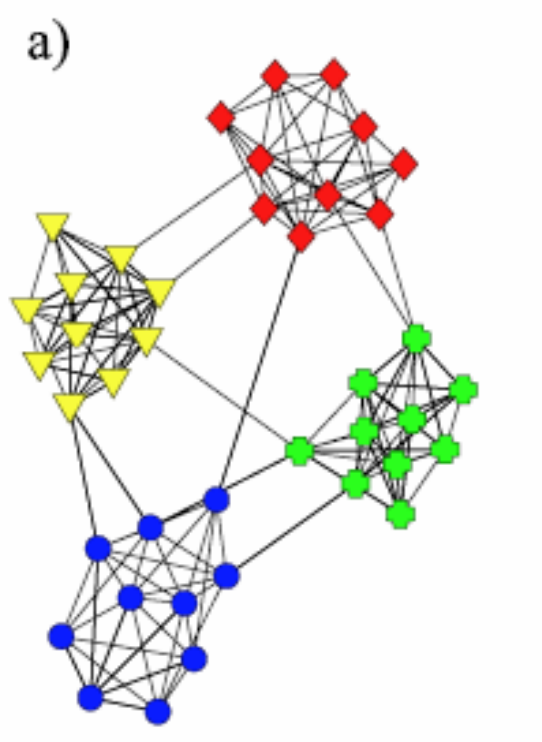
$$\gamma [l]_{p_{ij}}^{s\leftrightarrow r} < l^{s\leftrightarrow r}$$

In a large network $[l]_{p_{ij}}^{s\leftrightarrow r}$ is well approximated by $n_s n_r / N$

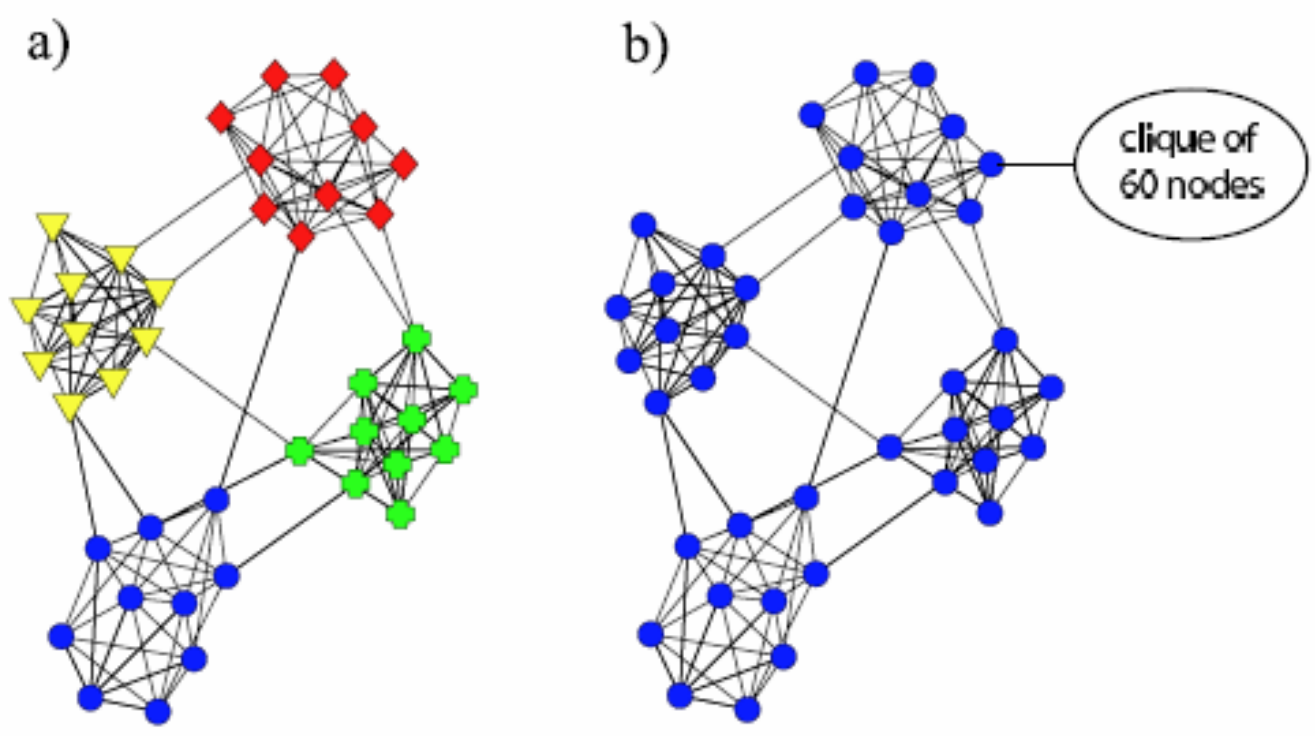
Assuming roughly equal size modules we have $n_s \approx n_r \lesssim \sqrt{N l^{s\leftrightarrow r}/\gamma}$

Even in the best case scenario $l^{s\leftrightarrow r} = 1$ modules smaller than $\sqrt{N/\gamma}$ cannot be resolved

The effect of network size on the resolution

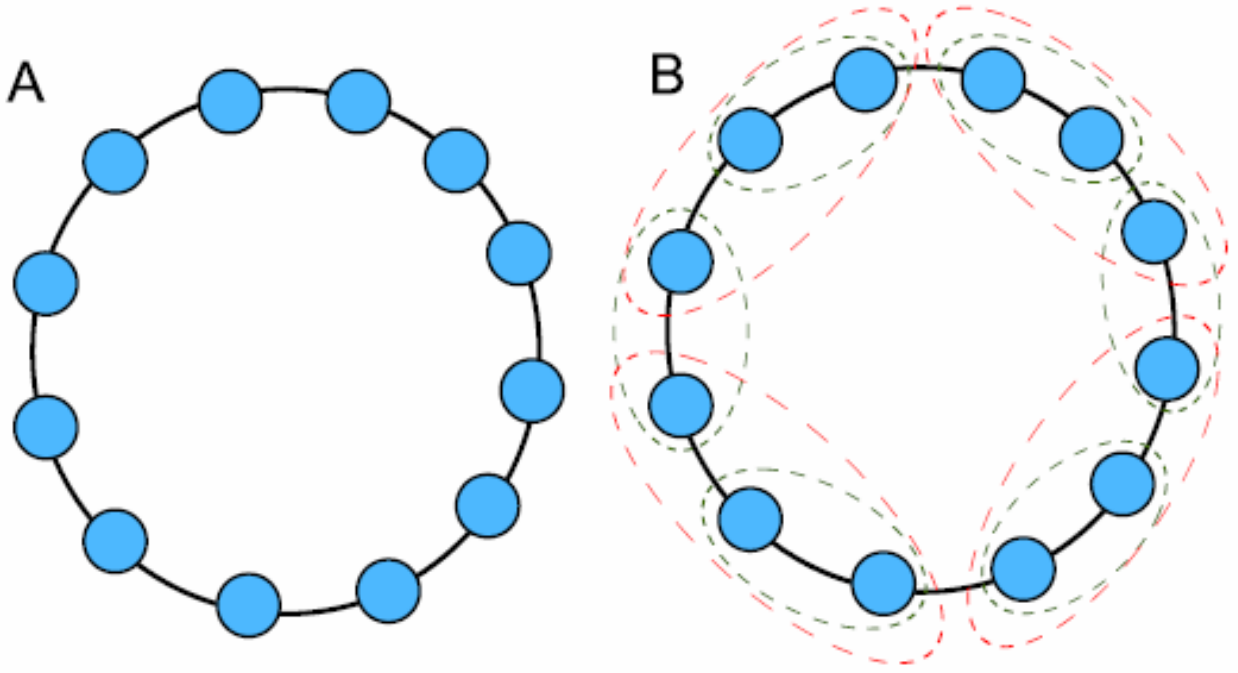


The effect of network size on the resolution



Action at distance!

Changing γ will change the resolution limit leading to the following:

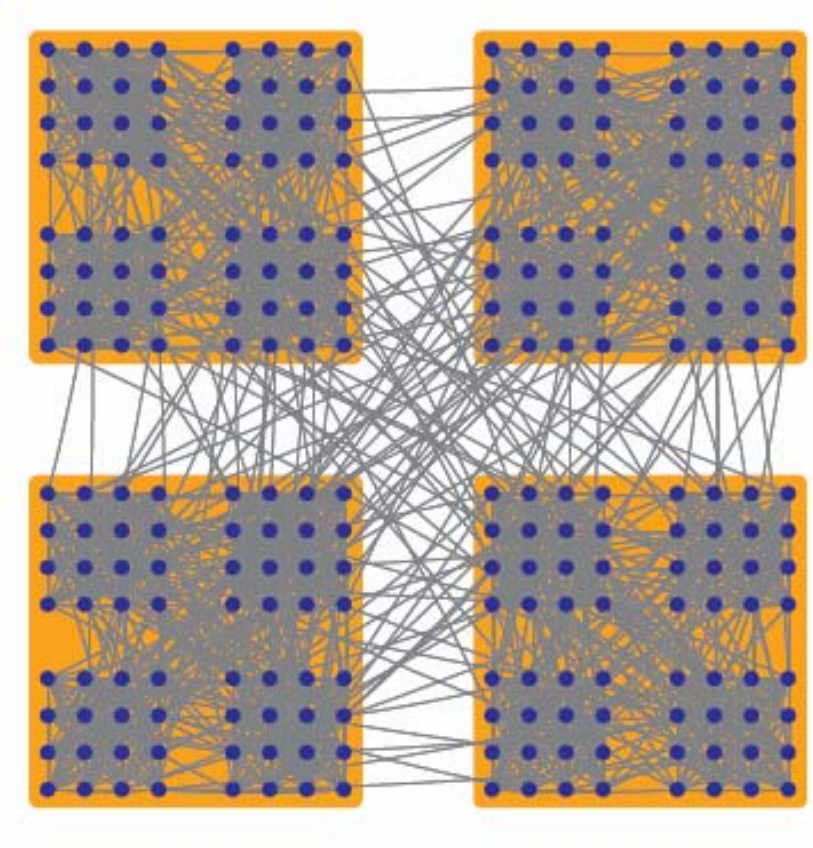


No physical meaning in merging two (or more) cliques connected by single links.

There is no *a priori* knowledge how γ should be chosen. Trial and error.

MULTIRESOLUTION METHODS

Simple global optimization methods are unable to identify modules on all scales in large networks. This is a major problem if the network modules are hierarchically nested.



Way out: Use many γ values! (Reichardt and Bornholdt, Kumpula et al. II)

Sweeping through the γ values corresponds to changing continuously the resolution limit in the community-microscope (multiresolution method). At different γ different size communities become visible.

Arenas, Fernandez and Gomes introduced a similar tool to the N-G modularity concept.

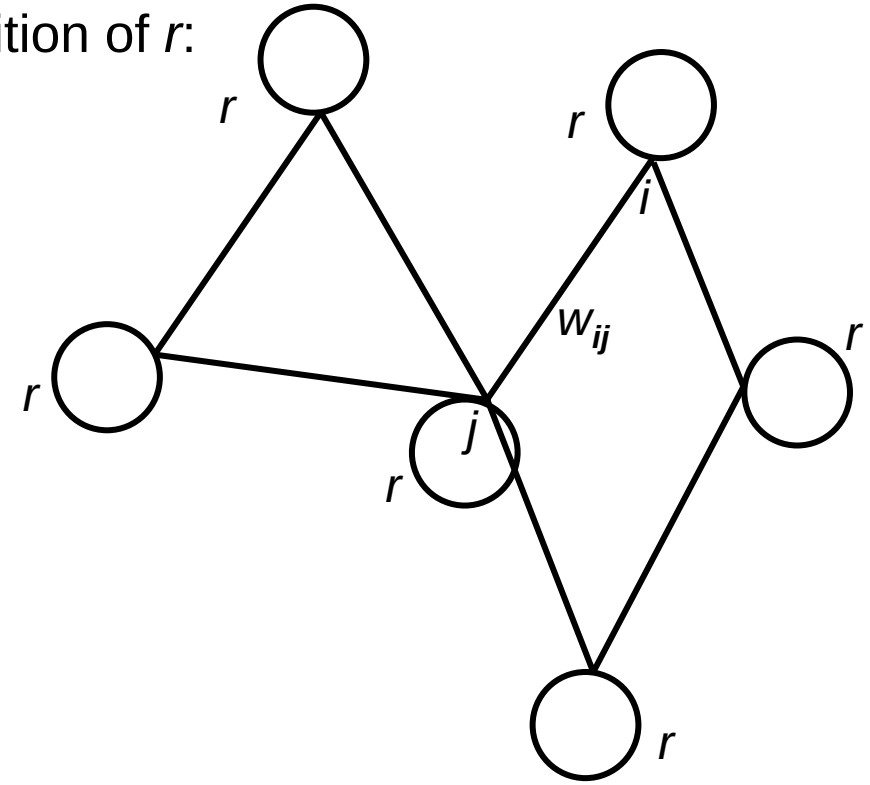
Consider the network as weighted. The modularity will become

$$Q_w(r) = \frac{1}{W(r)} \sum_{s=1}^{m} (w_{ss}(r) - [w_{ss}(r)])$$

where $[w_{ss}(r)]$ is the expectation of $w_{ss}(r)$, the total link weight in community *s* in the null model

W(*r*) is the total link weight in the network, *r* is the parameter, which tunes the resolution

Definition of *r*:



Self-link to each node with weight *r*

r changes the total weight and thus the resolution

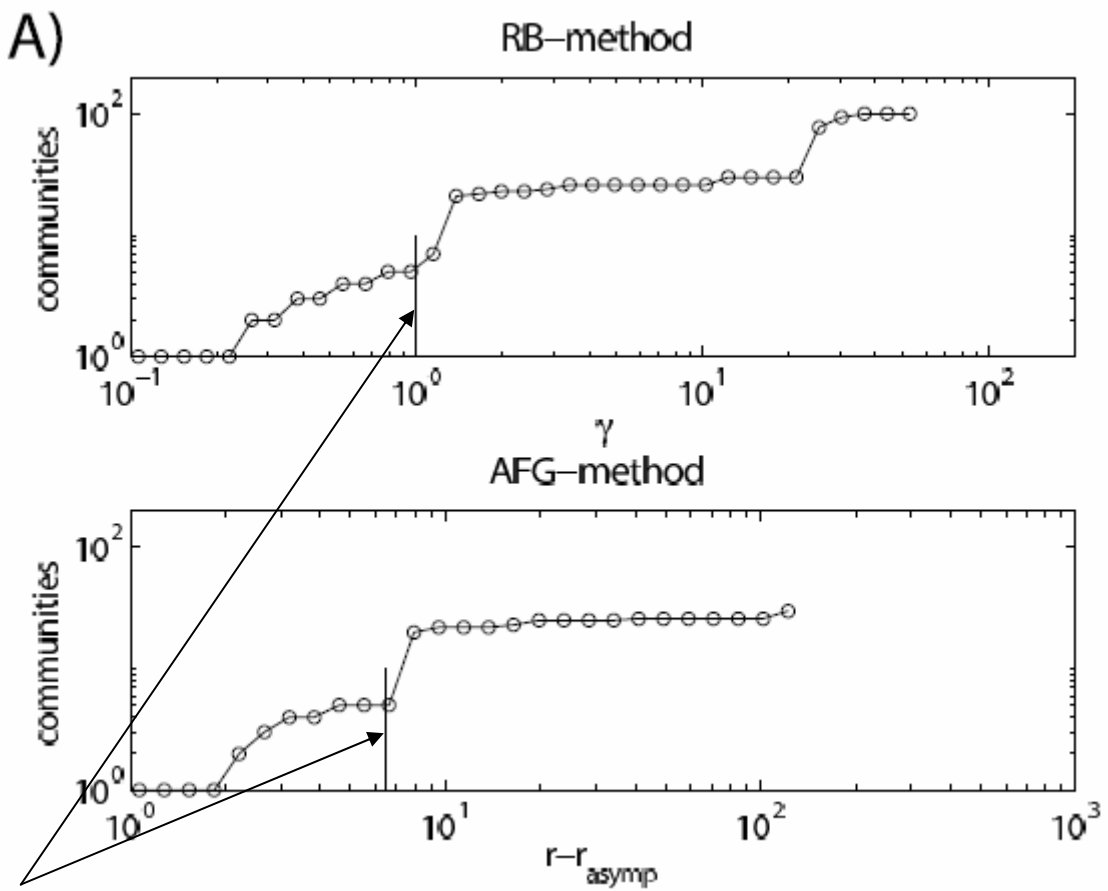
γ changes the effective number of links to L/γ

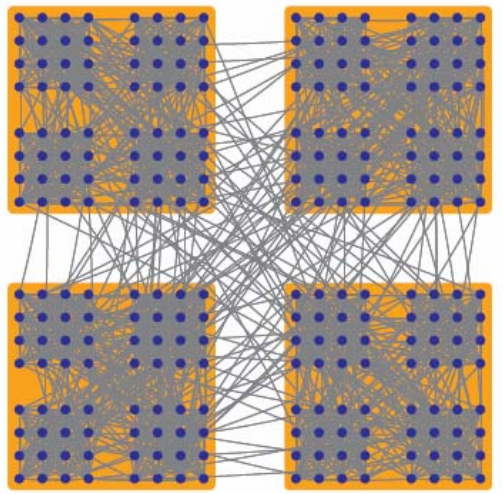
Difference: *r* changes the total weight in the module

Problem: Given a resolution limit (by γ or *r*) what distinguishes between spurious merger of smaller communities and communities of a higher hierarchical level?

Search for stability! If there is a long plateau in the # communities vs resolution parameter the community structure can be taken seriously.

Example A. 125 node artifitial hierarchical network





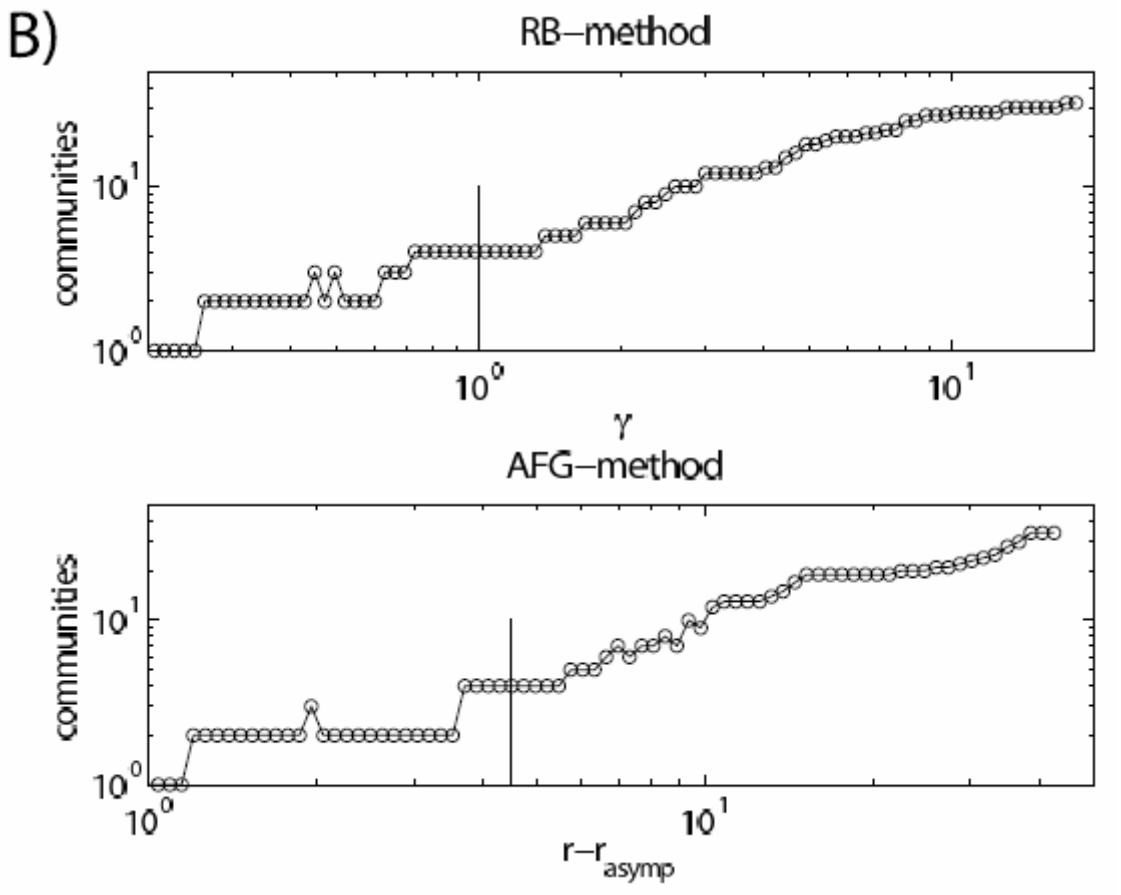
$$r_{\text{asymp}} = -\frac{W}{N}$$

where the total strength is 0, i.e., no meaningful scales for $r < r_{\text{asymp}}$

„Plain" methods: $\gamma = 1$, $r = 0$

Example B. : Zachary's karate club (34 nodes)

Modularity optimization leads to 4 communities, whereas sociologically 2 communities are identified.



For the 2-community solution the splitting happens along the „physical" line

MULTIRESOLUTION IN A LOCAL METHOD

(Lancichinetti, Fortunato, JK)

Use node fitness to decide if a node should be included into the module:

$$f_{\mathcal{G}} = \frac{k_{in}^{\mathcal{G}}}{(k_{in}^{\mathcal{G}} + k_{out}^{\mathcal{G}})^{\alpha}}$$

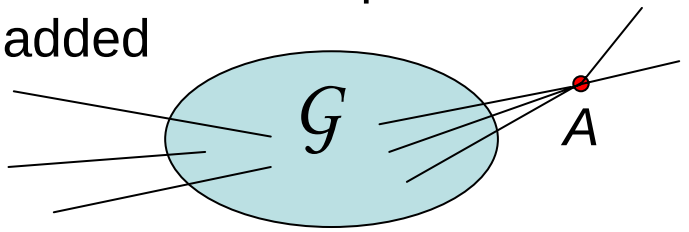
Subgraph (module) fitness

$$f_{\mathcal{G}}^{A} = f_{\mathcal{G}+\{A\}} - f_{\mathcal{G}-\{A\}}$$

Fitness of node *A*

Key step: Include *A* if $f_{\mathcal{G}}^{A} > 0$. (+Check the already member nodes eliminate those with negative fitness redo it until no node with positive fitness can be added

$$f_{\mathcal{G}}^{A} = \frac{k_{in}^{\mathcal{G}} + 2k_{in}^{A}}{(k^{\mathcal{G}} + k^{A})^{\alpha}} - \frac{k_{in}^{\mathcal{G}}}{k^{\mathcal{G}\alpha}} > 0$$



$$f_{\mathcal{G}}^{A} = \frac{k_{in}^{\mathcal{G}} + 2k_{in}^{A}}{(k^{\mathcal{G}} + k^{A})^{\alpha}} - \frac{k_{in}^{\mathcal{G}}}{k^{\mathcal{G}\alpha}} > 0$$

Assume $\frac{k^{A}}{k^{\mathcal{G}}} << 1$

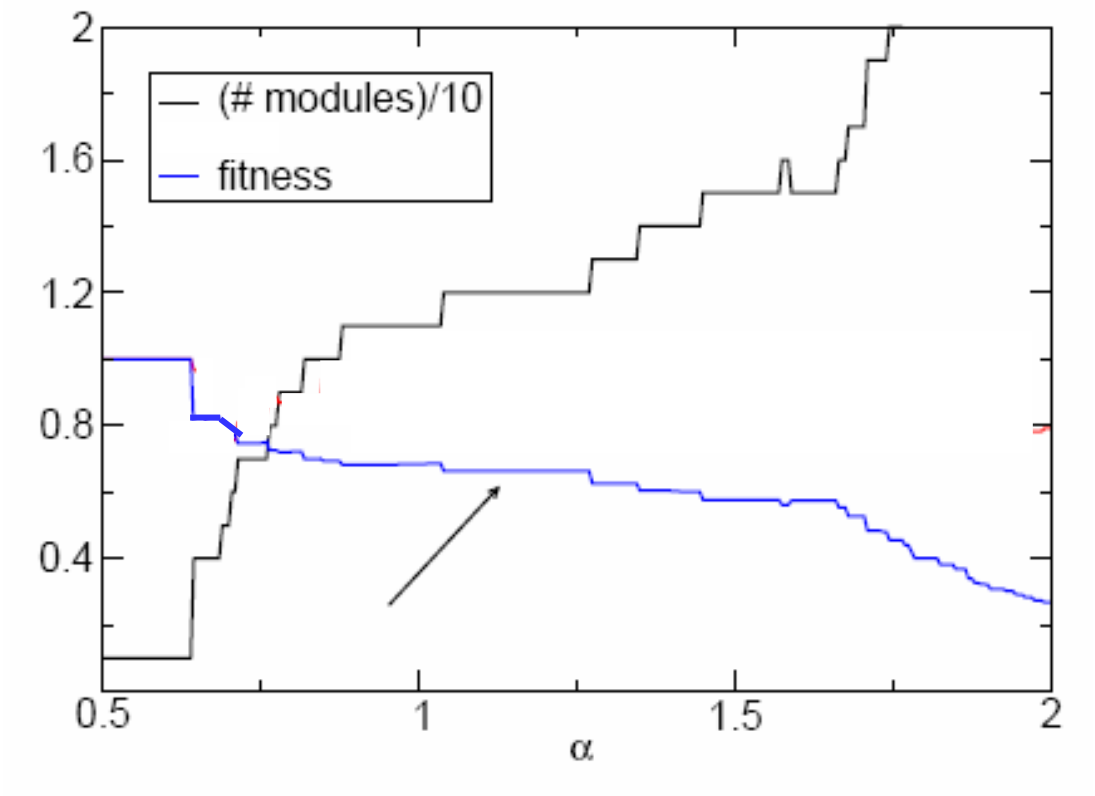
$$k_{in}^{\mathcal{G}} k^{\mathcal{G}\alpha} + 2k_{in}^{A} k^{\mathcal{G}\alpha} - k_{in}^{\mathcal{G}} k^{\mathcal{G}\alpha}(1 + \alpha \frac{k^{A}}{k^{\mathcal{G}}}) > 0$$

$$2k_{in}^{A} k^{\mathcal{G}\alpha} > \alpha k_{in}^{\mathcal{G}} k^{\mathcal{G}\alpha} \frac{k^{A}}{k^{\mathcal{G}}}$$

$$\frac{2k_{in}^{A}}{\alpha k^{A}} > \frac{k_{in}^{\mathcal{G}}}{k^{\mathcal{G}}}$$

i.e., α is a resolution parameter: If α is large, it is difficult to include new nodes → small modules, # modules large

Again: search for stability



The length of the plateaus measures stability.

American college football teams: 12 conferences within which more games played than outside. Tuning the resolution finds the most stable module configuration, corresponding to the 12 modules.

The „natural” value of the resolution parameter is often not the best!

Summary

- Global methods are appealing: They remind us to other problems (looking for ground state of a many body system, tasks in discrete math)
- These are NP complete problems (like finding the ground state of a spin glass); approximate methods are to be introduced
- Approximate method may be (ultra)fast, however, they may lead to spurious solutions
- Resolution limit is intrinsic to global optimization (it is NOT a result of approximate solutions). The reason is that in a large random network the probability of connecting two modules becomes very small
- Multiresolution resolves the problem by a suitable tuning parameter (time consuming…)
- Stability of solutions have to be checked
- Hierarchical structure becomes explorable
- Multiresolution is applicable to some local methods too

Thank you!