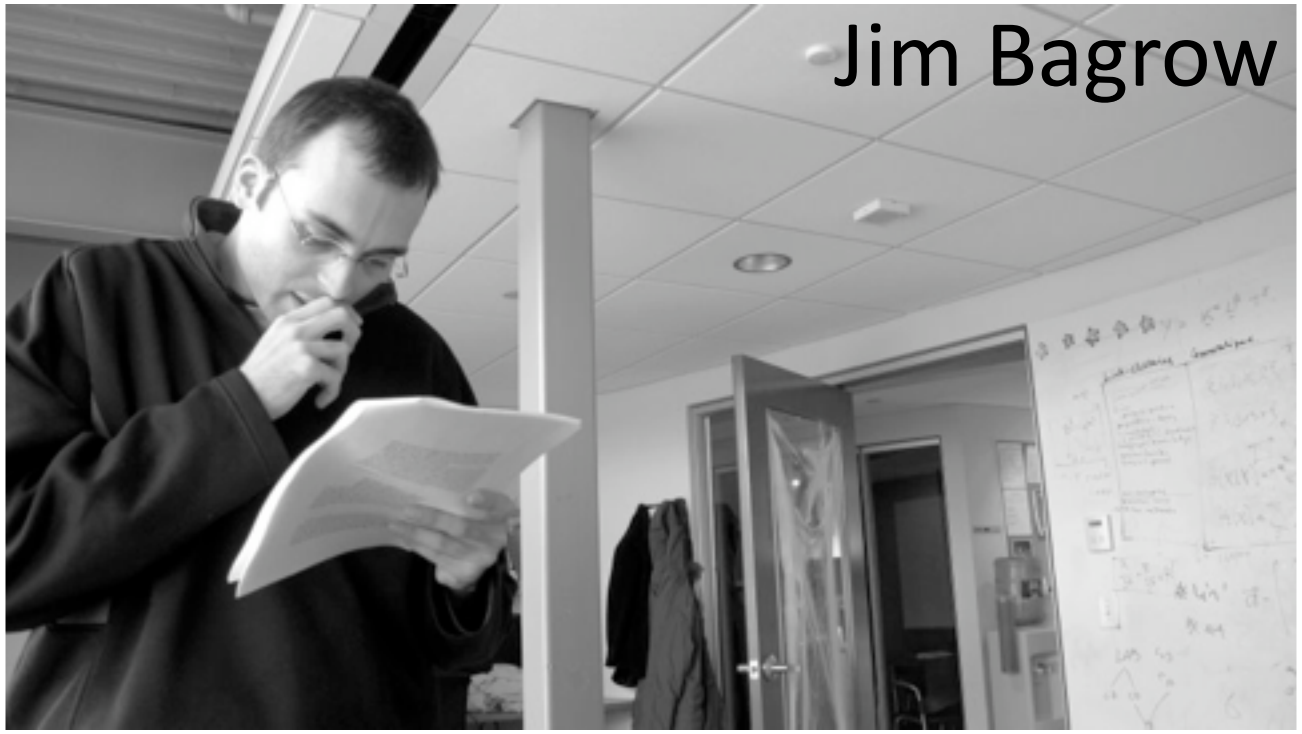
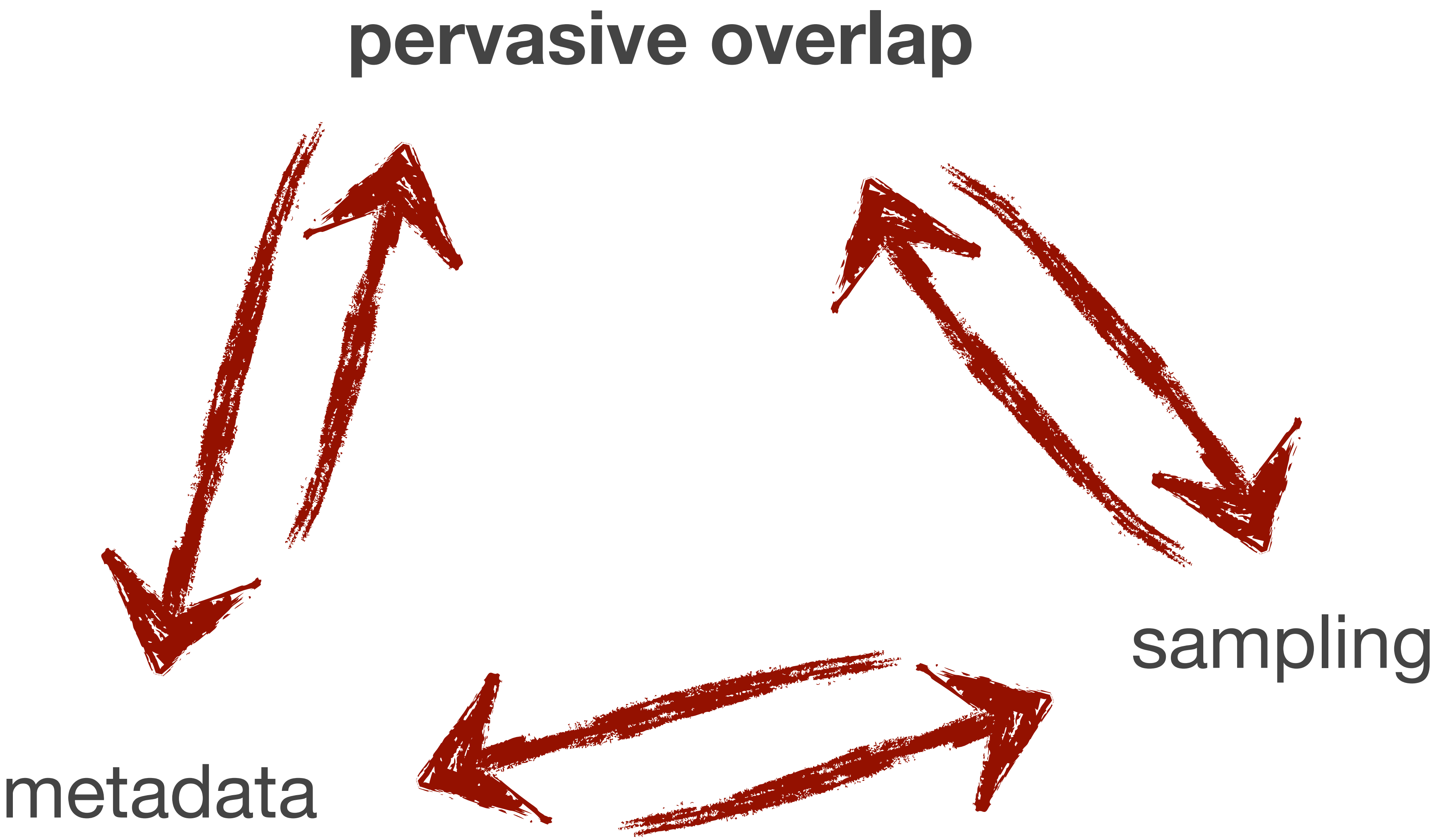
sampling & community structure

in densely connected networks

Sune Lehmann, YY Ahn and JP Bagrow
Technical University of Denmark


Yong-Yeol Ahn


Jim Bagrow


pervasive overlap
sampling
metadata

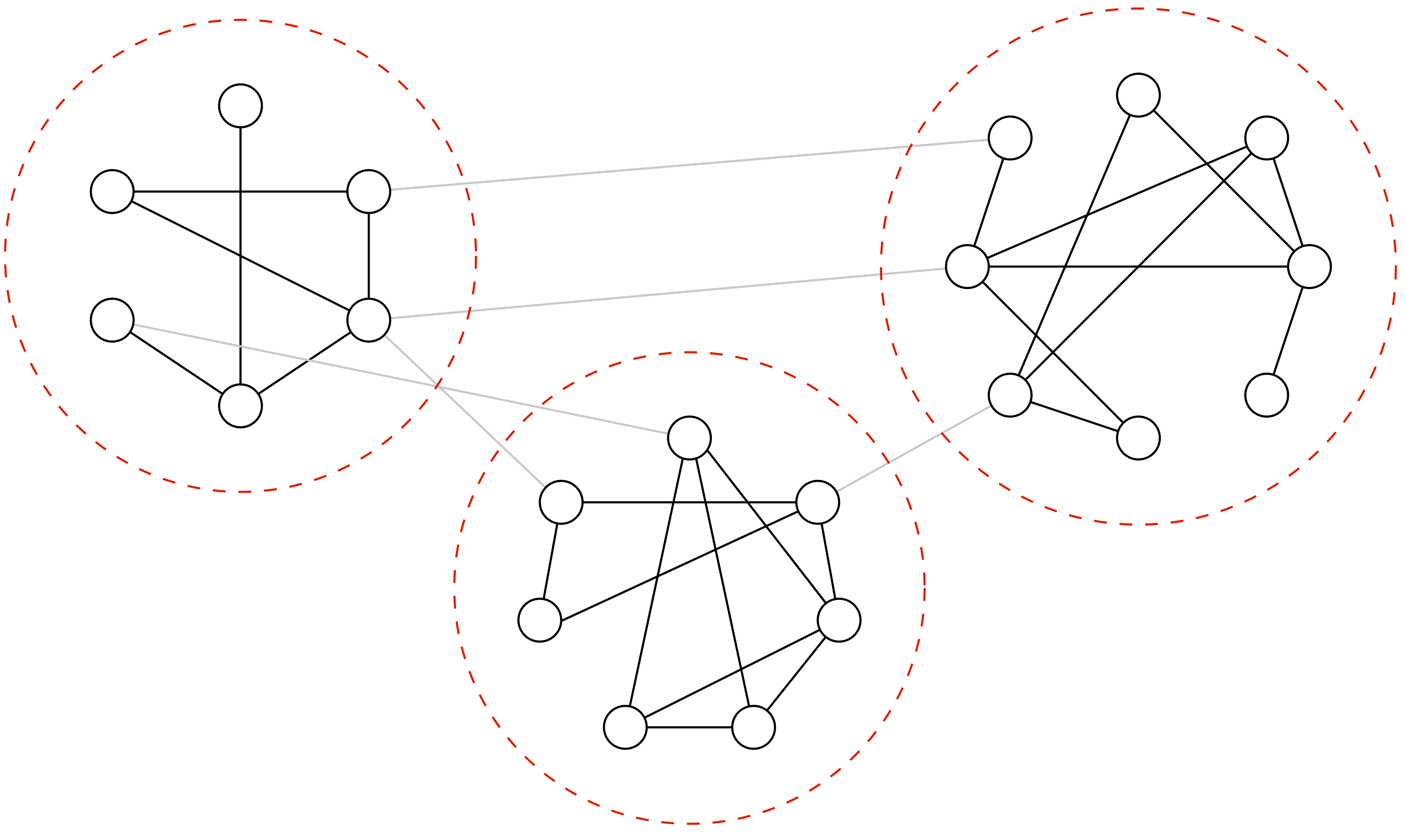
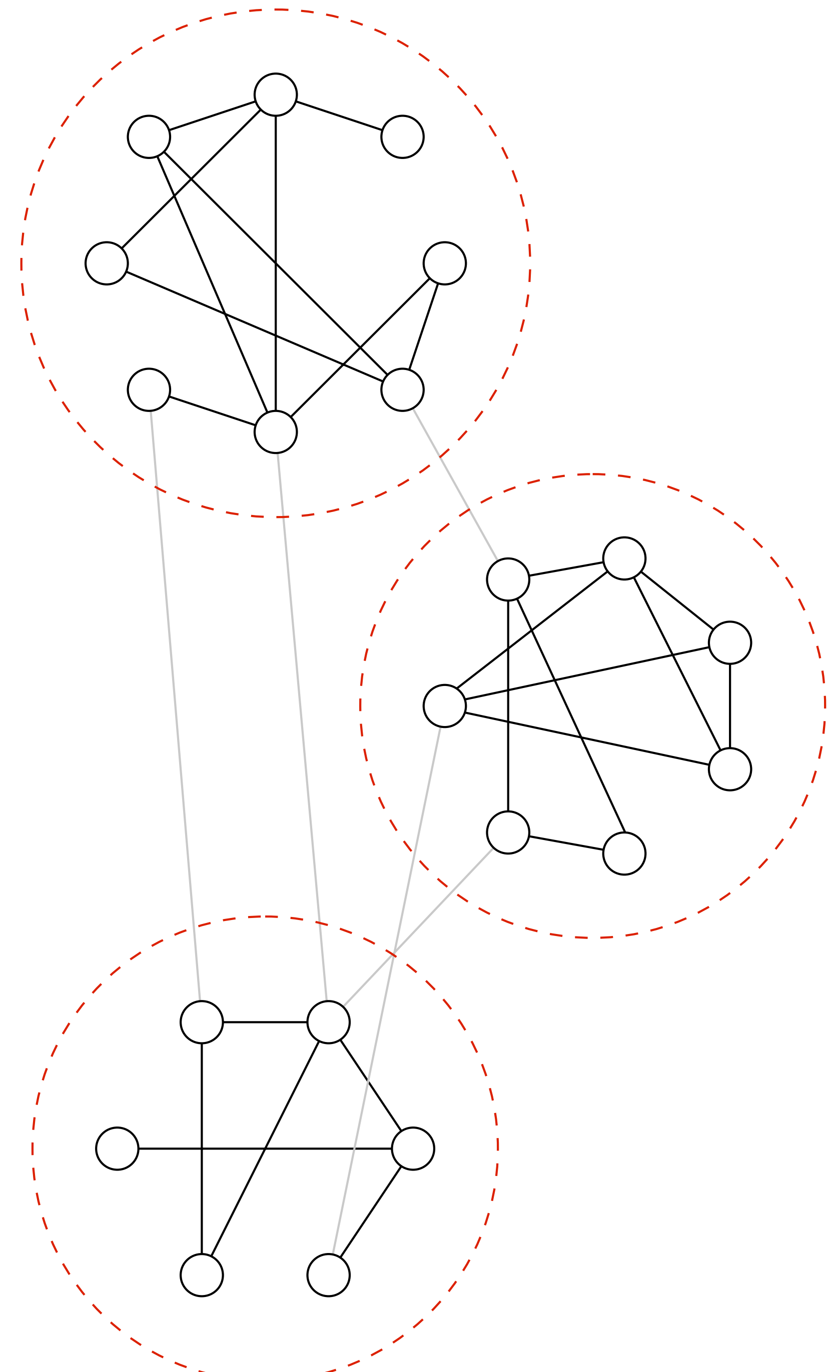
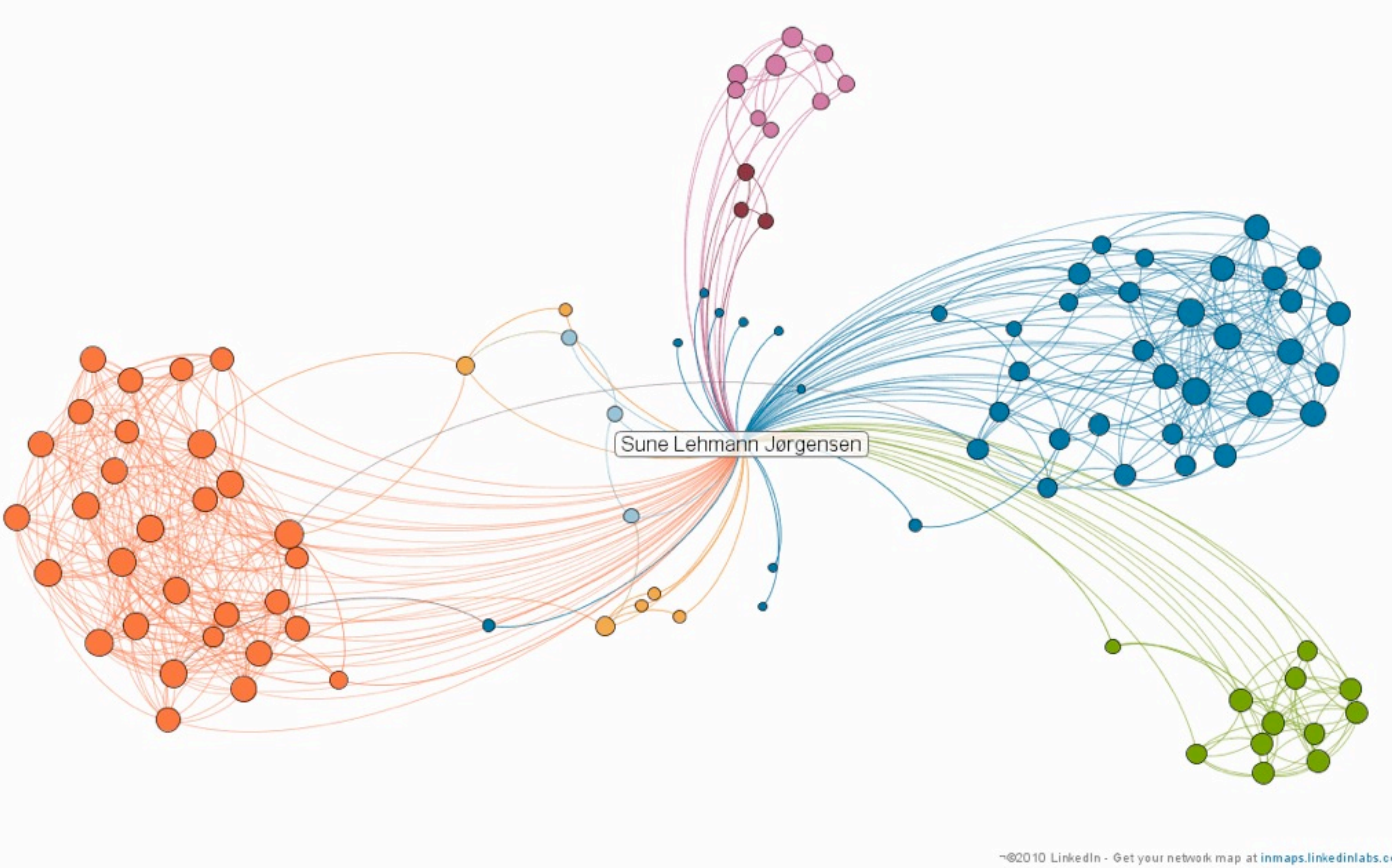
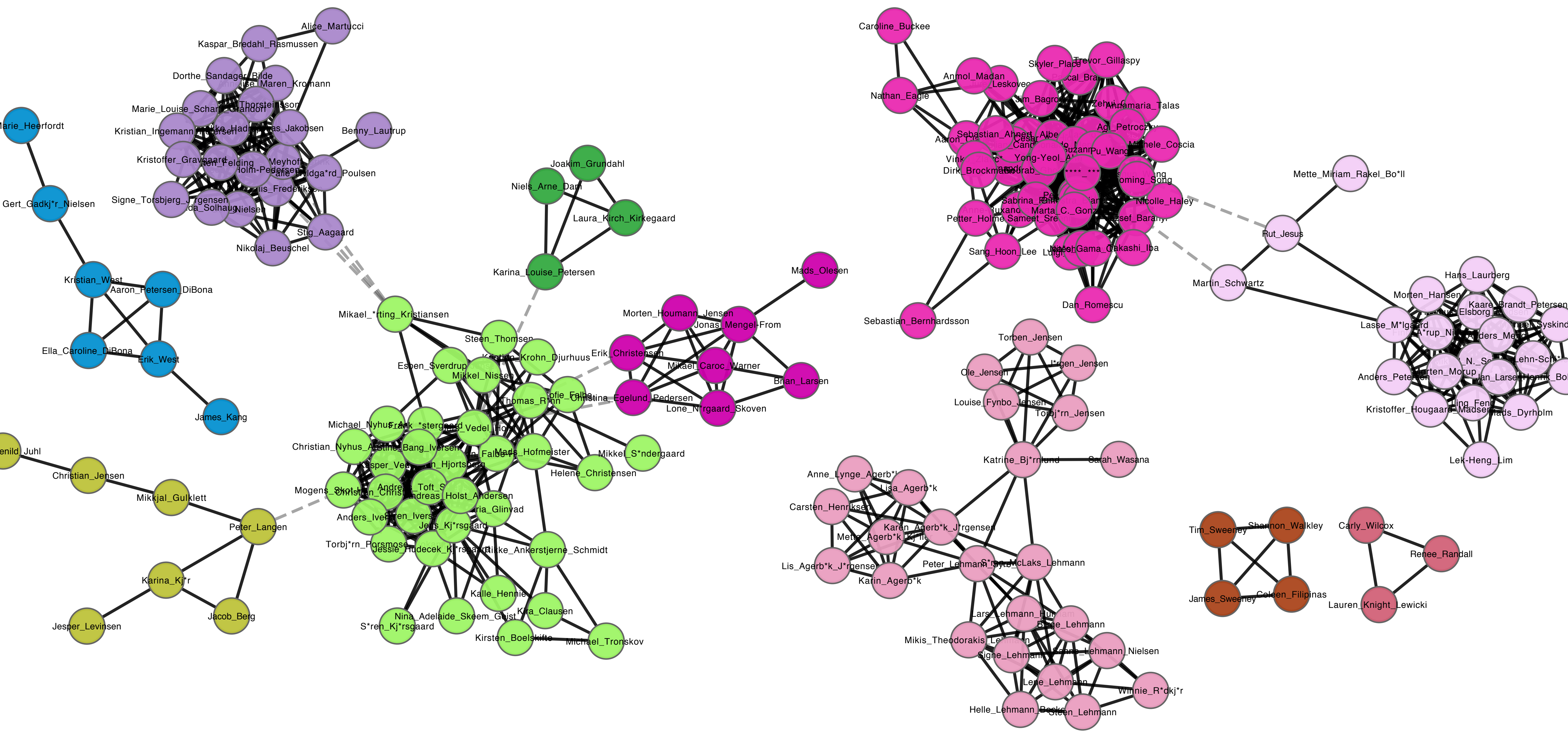
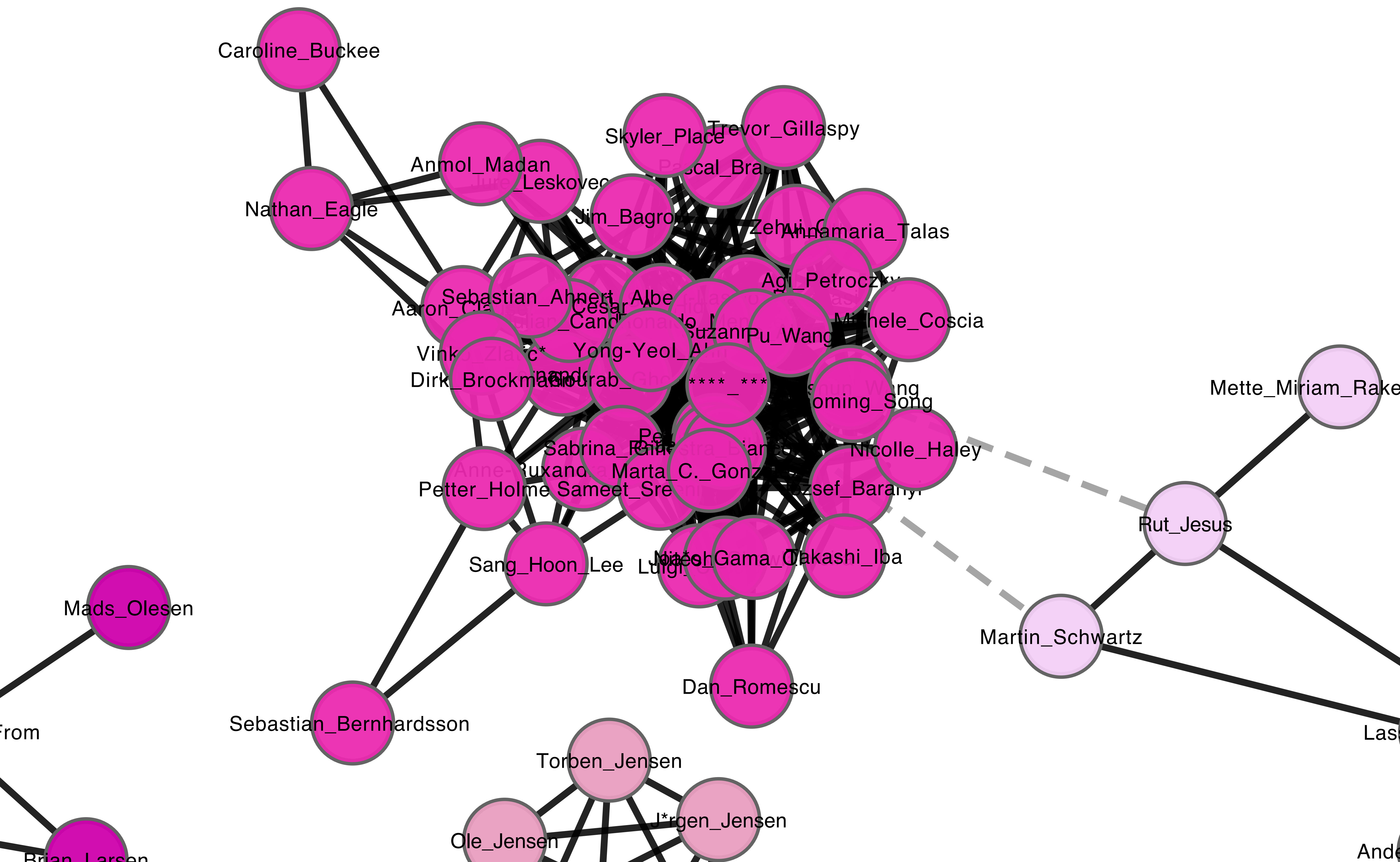


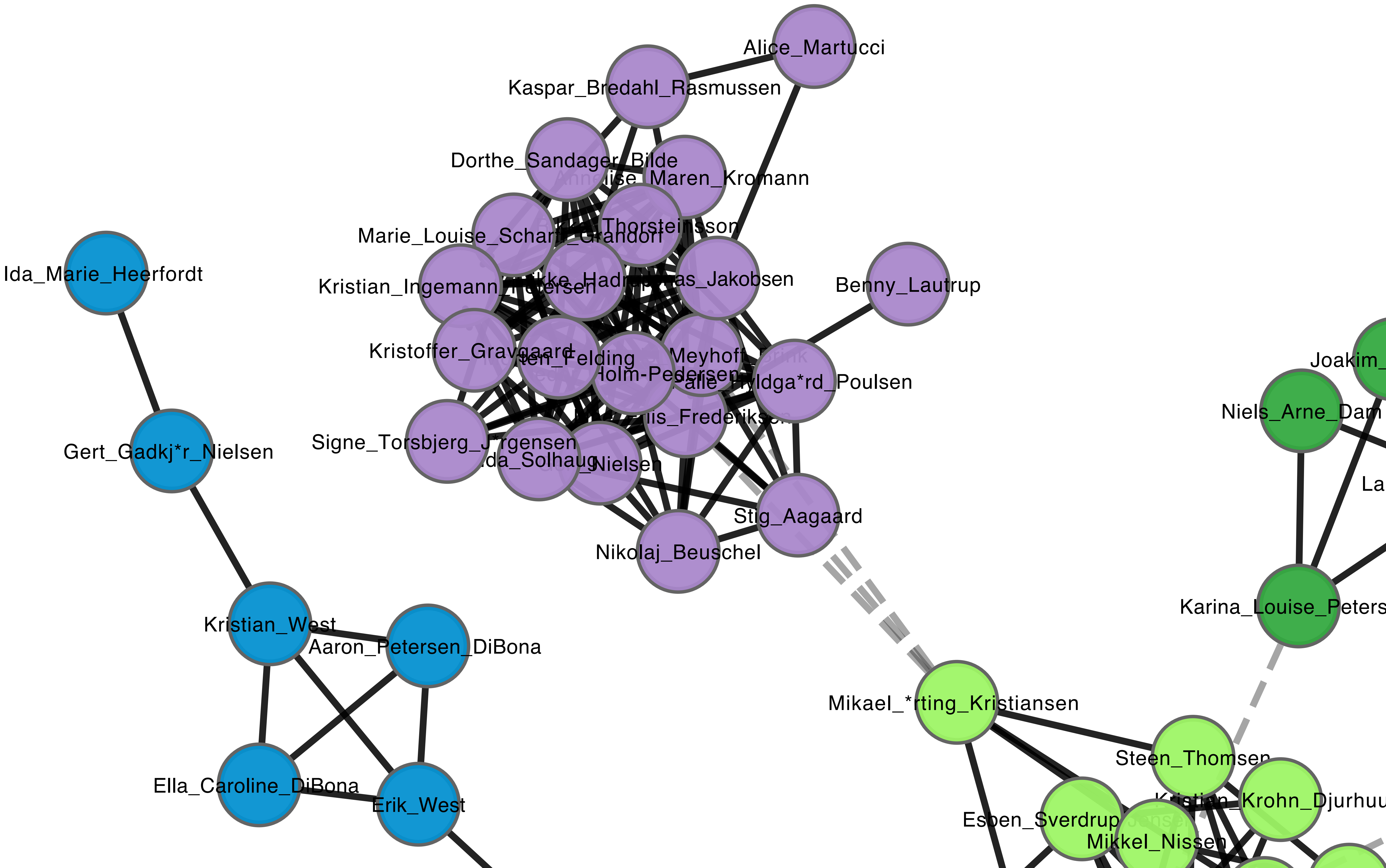
Illustration: Newman & Girvan. PRE 69, 026113 (2004)



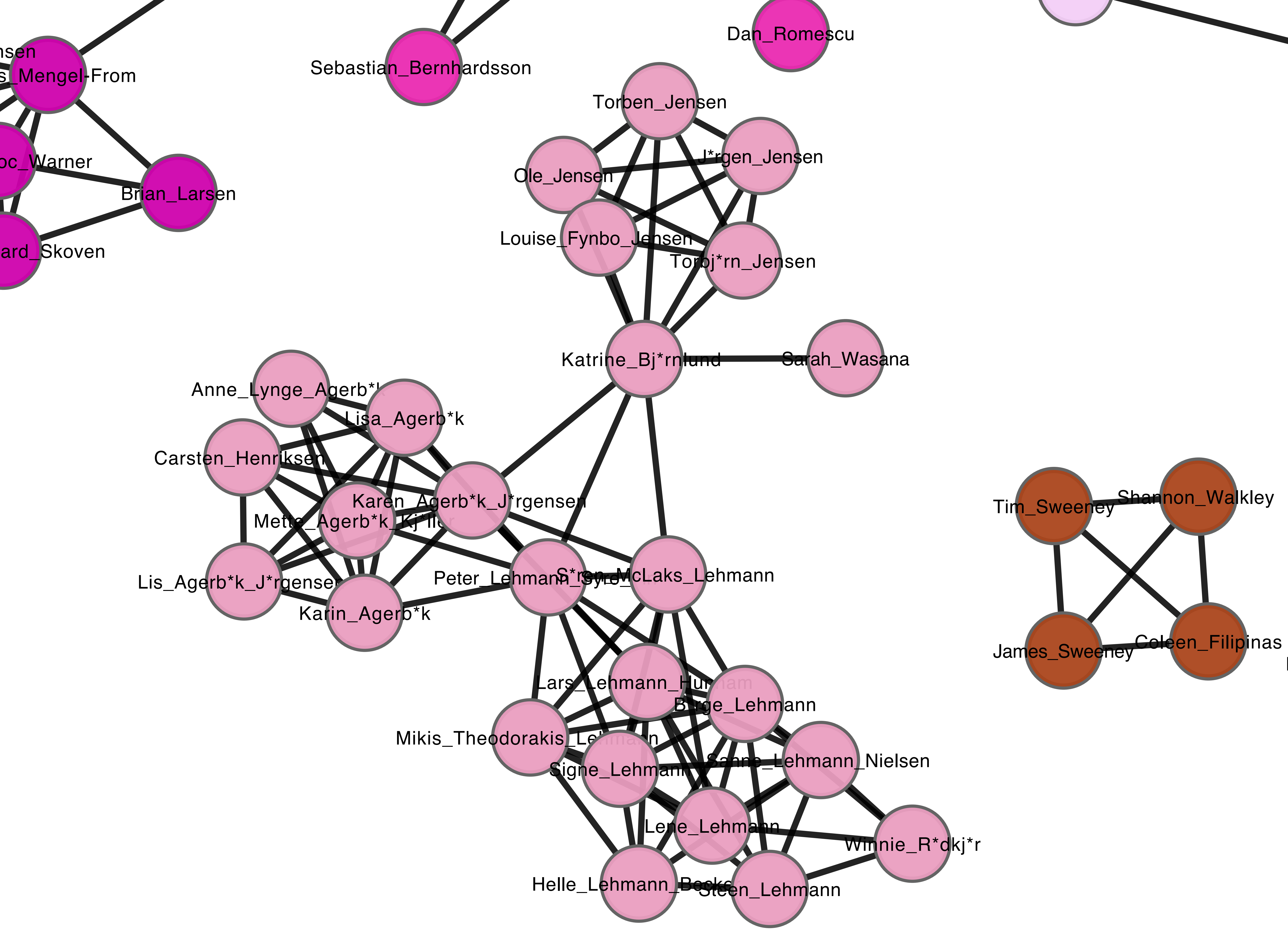

Sune Lehmann Jørgensen
©2010 LinkedIn - Get your network map at inmaps.linkedinlabs.c


Alice_Martucci
Kaspar_Bredahl_Rasmussen
Dorthe_Sandager_Bilde
Marie_Louise_Schantz_Grandorf
Kristian_Ingemann
Benny_Lautrup
Kristoffer_Gravgaard
Signe_Torsbjerg_J*rgensen
Stig_Aagaard
Nikolaj_Beuschel
Marie_Heerfordt
Gert_Gadkj*r_Nielsen
Kristian_West
Aaron_Petersen_DiBona
Ella_Caroline_DiBona
Erik_West
James_Kang
Christian_Jensen
Mikkjal_Gulklett
Peter_Langen
Karina_Kj*r
Jesper_Levinsen
Jacob_Berg
Joakim_Grundahl
Niels_Arne_Dam
Laura_Kirch_Kirkegaard
Karina_Louise_Petersen
Mikael_*rting_Kristiansen
Steen_Thomsen
Espen_Sverdrup
Mikkel_Nissen
Thomas_R*nn
Mads_Hofmeister
Mikkel_S*ndergaard
Helene_Christensen
Michael_Nyhus
Christian_Nyhus
Mogens
Anders_Ivers
Torbj*rn_Porsmose
Jesper_Veje
Holst_Andersen
Jens_Kj*rsgaard
Birke_Ankerstjerne_Schmidt
Kalle_Hennie
Kira_Clausen
Nina_Adelaide_Skeem_Geist
S*ren_Kj*rsgaard
Kirsten_Boelskifte
Michael_Tronskov
Mads_Olesen
Morten_Houmann_Jensen
Jonas_Mengel-From
Erik_Christensen
Mikael_Caroc_Warner
Christina_Egelund_Pedersen
Lone_N*rgaard_Skoven
Brian_Larsen
Caroline_Buckee
Nathan_Eagle
Anmol_Madan
Skyler_Place
Trevor_Gillaspy
Annamaria_Talas
Sebastian_Ahnert
Michele_Coscia
Yong-Yeol_Ahn
Dirk_Brockmann
Pu_Wang
Chaoming_Song
Nicolle_Haley
Marta_C._Gonzalez
Sabrina_Rahi
Petter_Holme
Sameet_Sreenivasan
Sang_Hoon_Lee
Takashi_Iba
Dan_Romescu
Sebastian_Bernhardsson
Mette_Miriam_Rakel_Bo*ll
Rut_Jesus
Martin_Schwartz
Hans_Laurberg
Morten_Hansen
Kaare_Brandt_Petersen
Lasse_M*lgaard
Anders_Petersen
Kristoffer_Hougaard_Madsen
Mads_Dyrholm
Lek-Heng_Lim
Torben_Jensen
Ole_Jensen
J*rgen_Jensen
Louise_Fynbo_Jensen
Torbj*rn_Jensen
Katrine_Bj*rnlund
Sarah_Wasana
Anne_Lynge_Agerb*k
Lisa_Agerb*k
Carsten_Henriksen
Karen_Agerb*k_J*rgensen
Mette_Agerb*k_Kjær
Lis_Agerb*k_J*rgensen
Karin_Agerb*k
Peter_Lehmann
S*ren_McLaks_Lehmann
Lars_Lehmann
Birge_Lehmann
Mikis_Theodorakis
Signe_Lehmann
Lene_Lehmann
Winnie_R*dkj*r
Helle_Lehmann
Steen_Lehmann
Tim_Sweeney
Shannon_Walkley
James_Sweeney
Coleen_Filipinas
Carly_Wilcox
Renee_Randall
Lauren_Knight_Lewicki

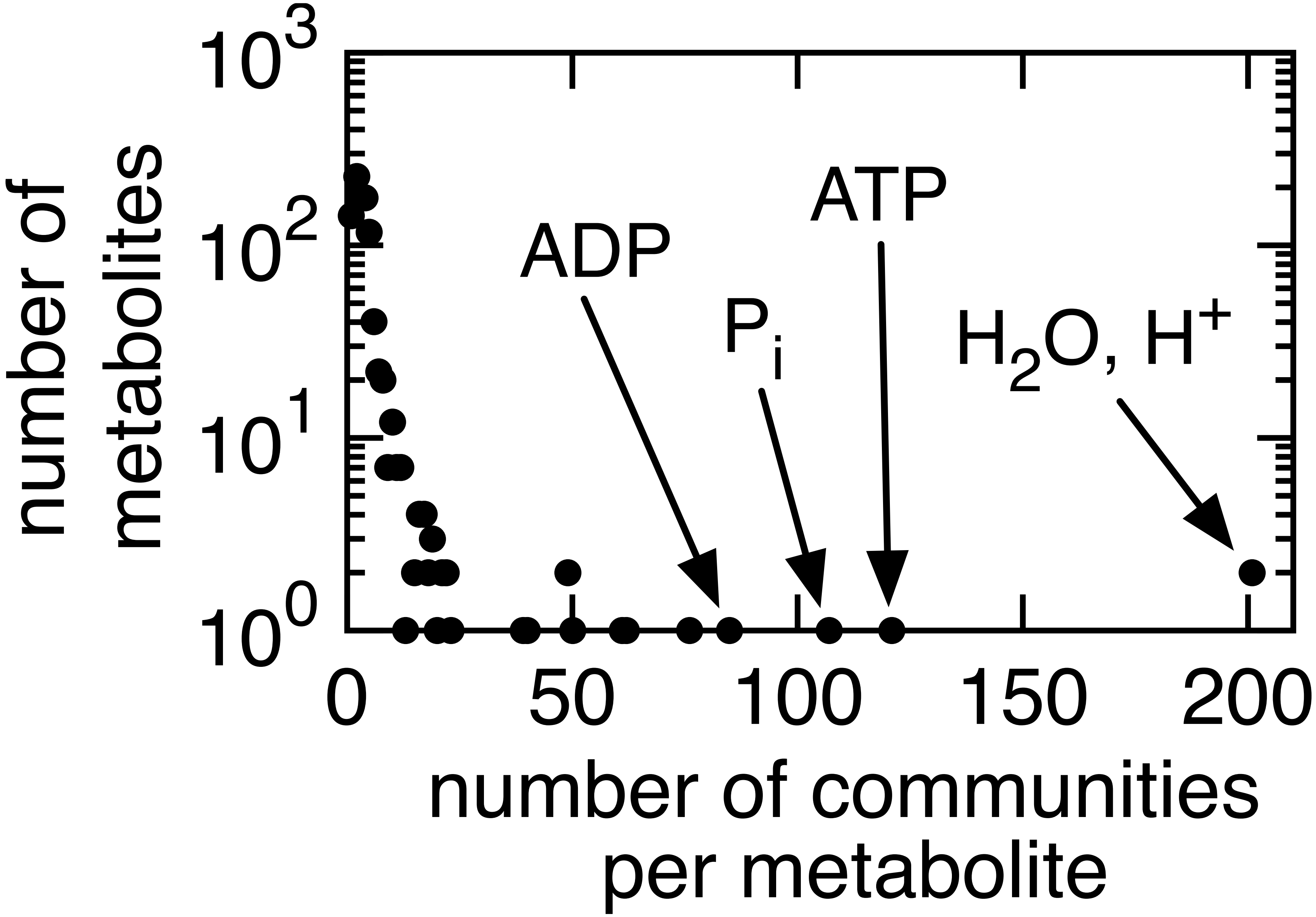

Caroline_Buckee
Nathan_Eagle
Anmol_Madan
Jure_Leskovec
Skyler_Place
Trevor_Gillaspy
Annamaria_Talas
Michele_Coscia
Pu_Wang
Sebastian_Ahnert
Nicolle_Haley
Petter_Holme
Sang_Hoon_Lee
Takashi_Iba
Dan_Romescu
Rut_Jesus
Martin_Schwartz
Mads_Olesen
Sebastian_Bernhardsson
Torben_Jensen
Ole_Jensen
J*rgen_Jensen



Alice_Martucci
Kaspar_Bredahl_Rasmussen
Dorthe_Sandager_Bilde
Benny_Lautrup
Kristoffer_Gravgaard
Stig_Aagaard
Nikolaj_Beuschel
Ida_Marie_Heerfordt
Gert_Gadkj*r_Nielsen
Kristian_West
Aaron_Petersen_DiBona
Ella_Caroline_DiBona
Erik_West
Mikael_*rting_Kristiansen
Steen_Thomsen
Mikkel_Nissen

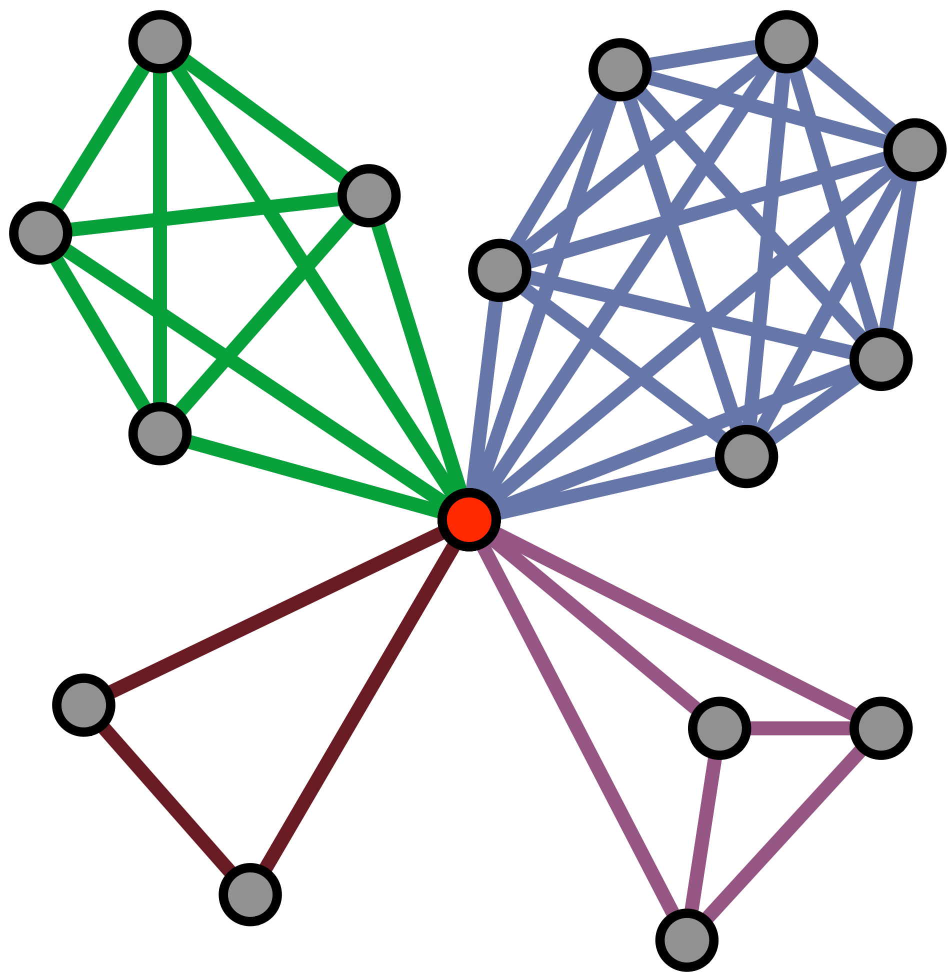


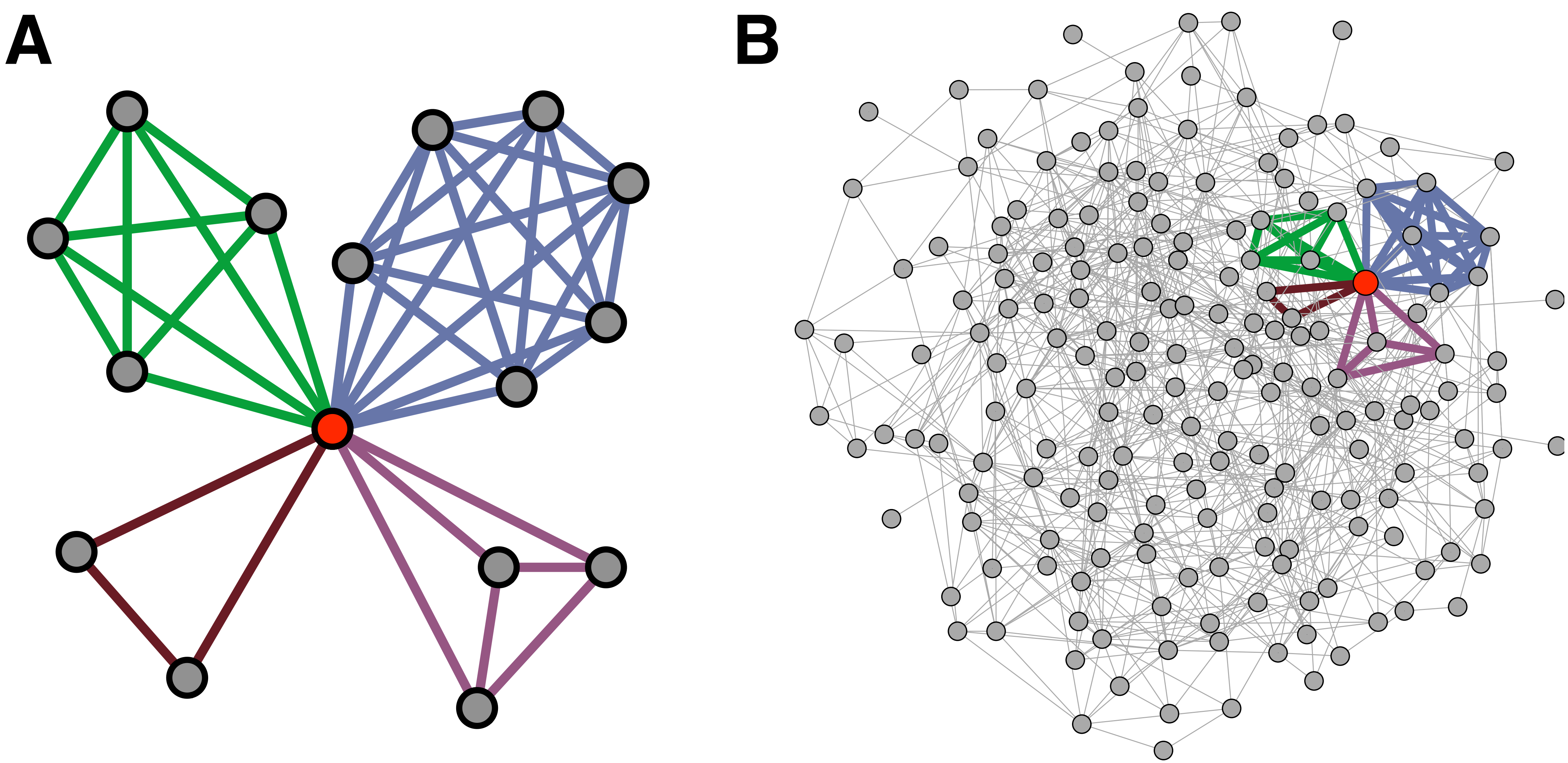
Sebastian_Bernhardsson
Dan_Romescu
Torben_Jensen
J*rgen_Jensen
Ole_Jensen
Louise_Fynbo_Jensen
Torbj*rn_Jensen
Katrine_Bj*rnlund
Sarah_Wasana
Brian_Larsen
Anne_Lynge_Agerb*k
Lisa_Agerb*k
Carsten_Henriksen
Karen_Agerb*k_J*rgensen
Lis_Agerb*k_J*rgensen
Karin_Agerb*k
Peter_Lehmann
McLaks_Lehmann
Lars_Lehmann_Hultham
B*rge_Lehmann
Mikis_Theodorakis_Lehmann
Signe_Lehmann
Sanne_Lehmann_Nielsen
Lene_Lehmann
Winnie_R*dkj*r
Helle_Lehmann_Becks
Steen_Lehmann
Tim_Sweeney
Shannon_Walkley
James_Sweeney
Coleen_Filipinas



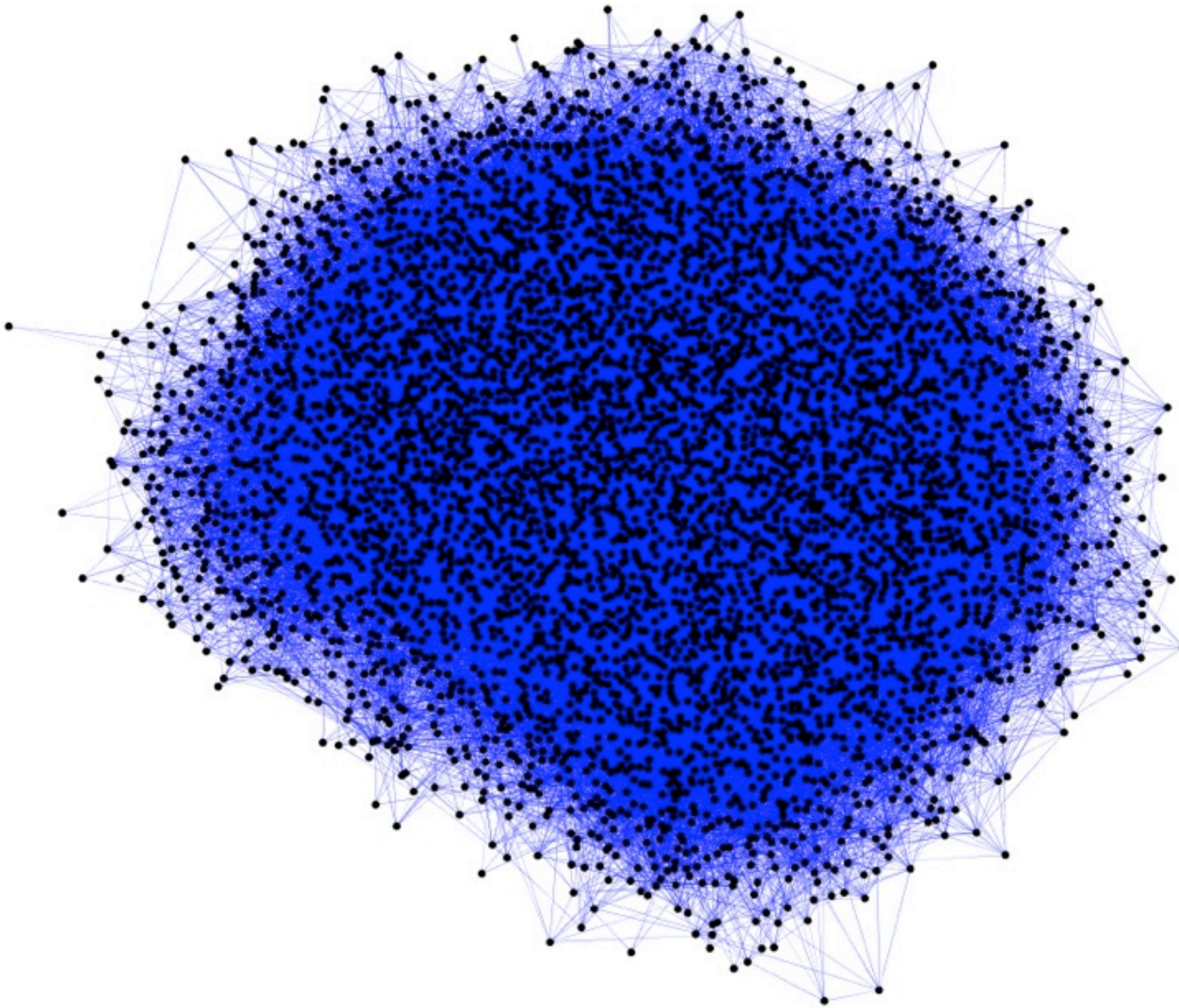
number of metabolites
number of communities per metabolite
10^0
10^1
10^2
10^3
0
50
100
150
200
ADP
P_i
ATP
H_2O, H^+

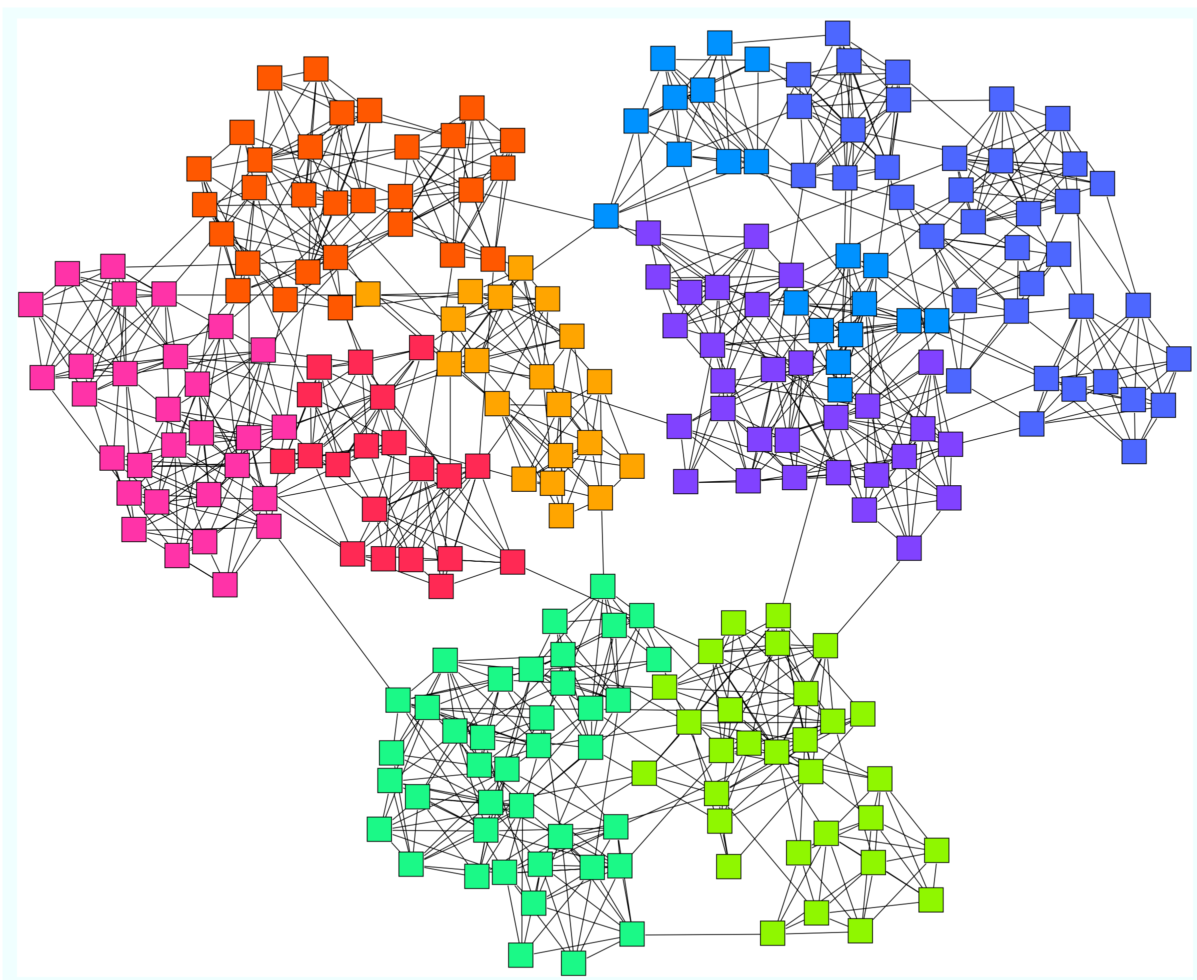
A



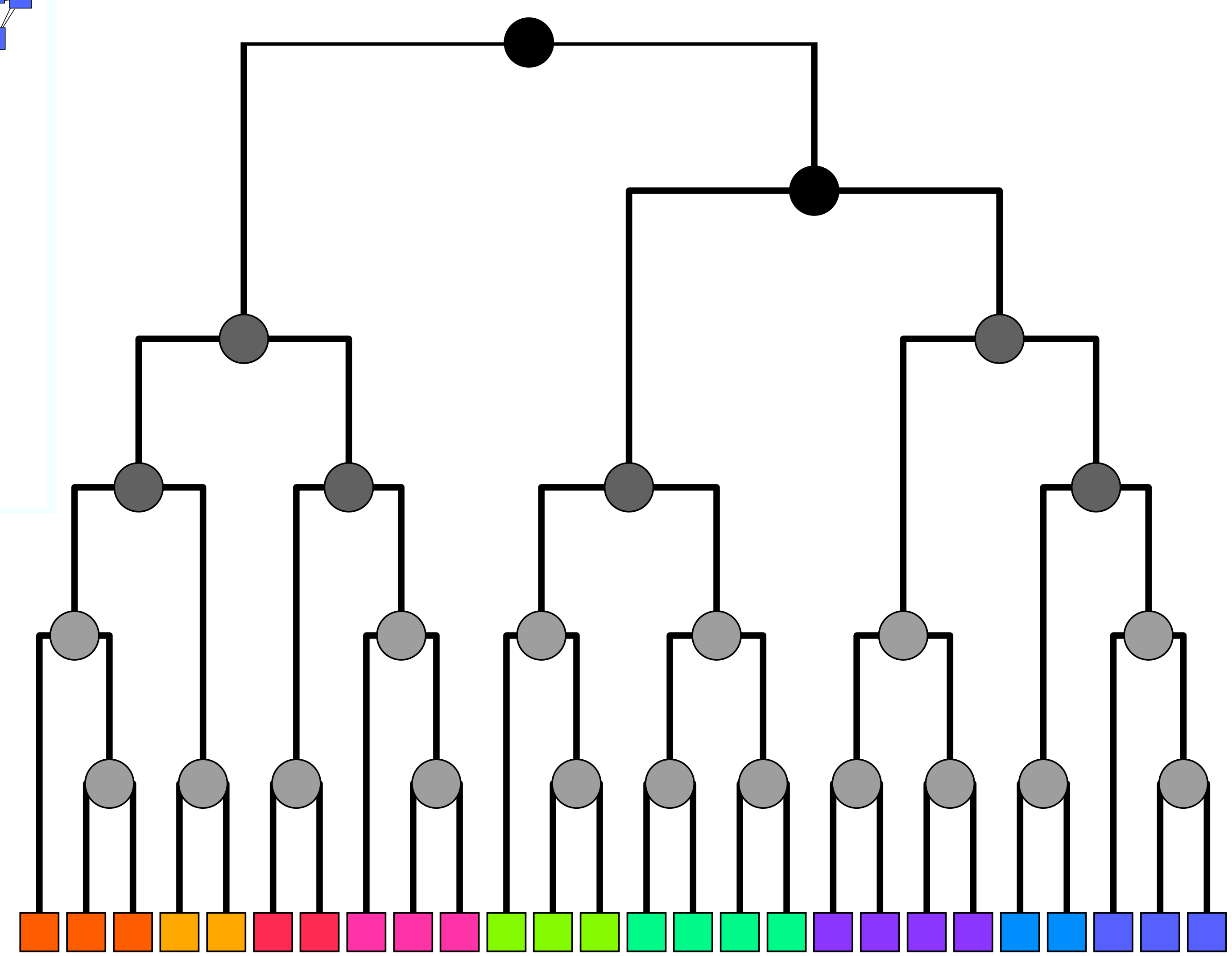


A
B



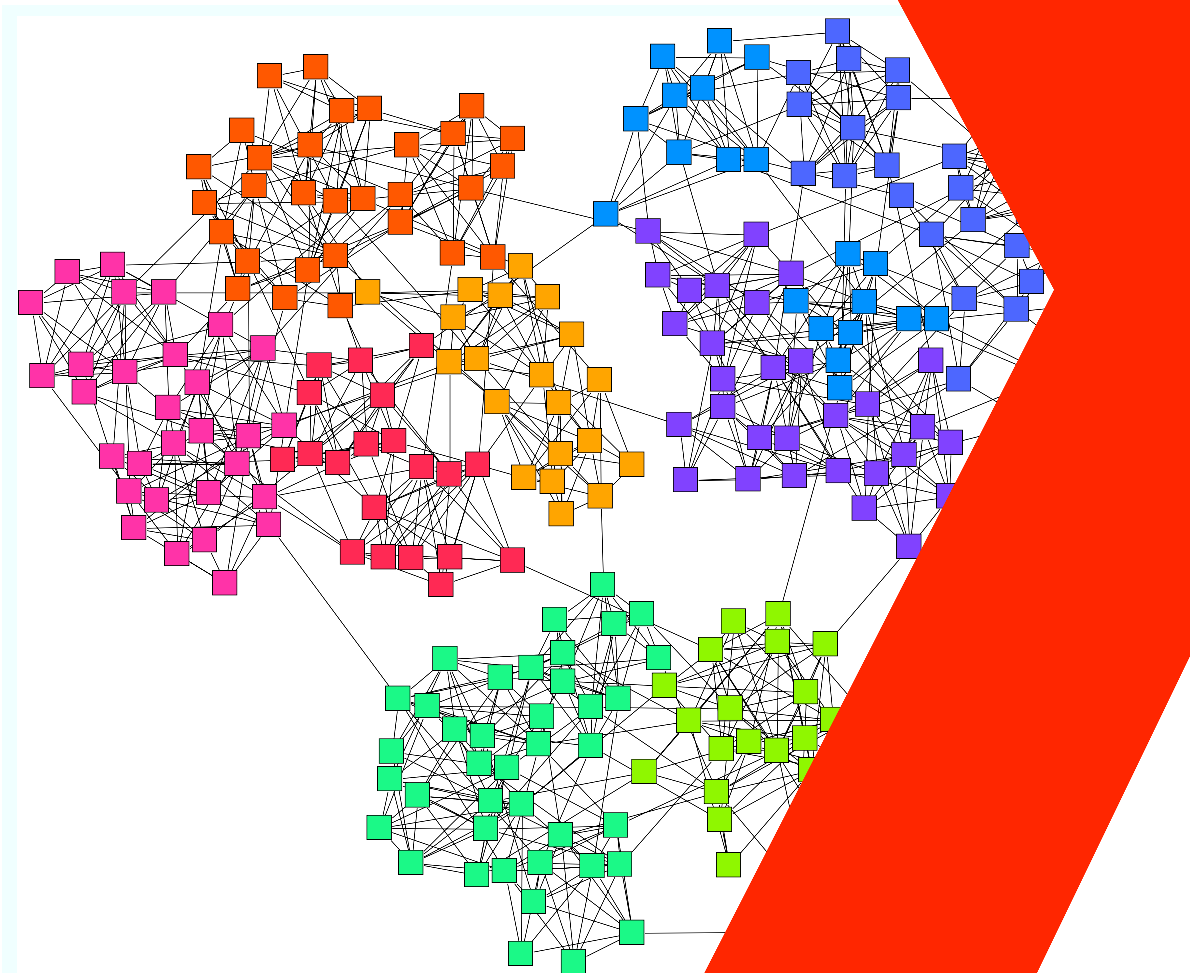


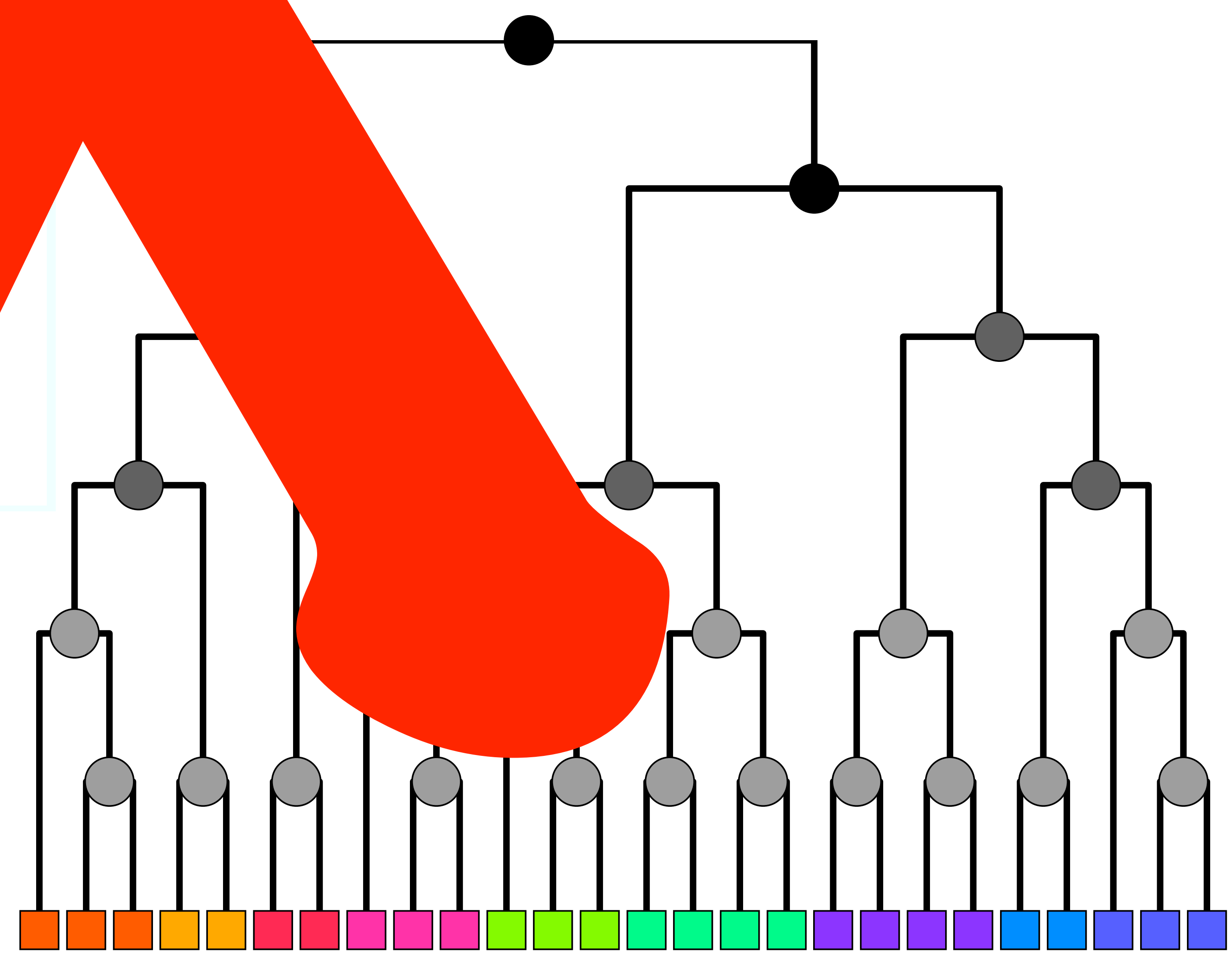
Hierarchy and community structure were thought to be two sides of the same story



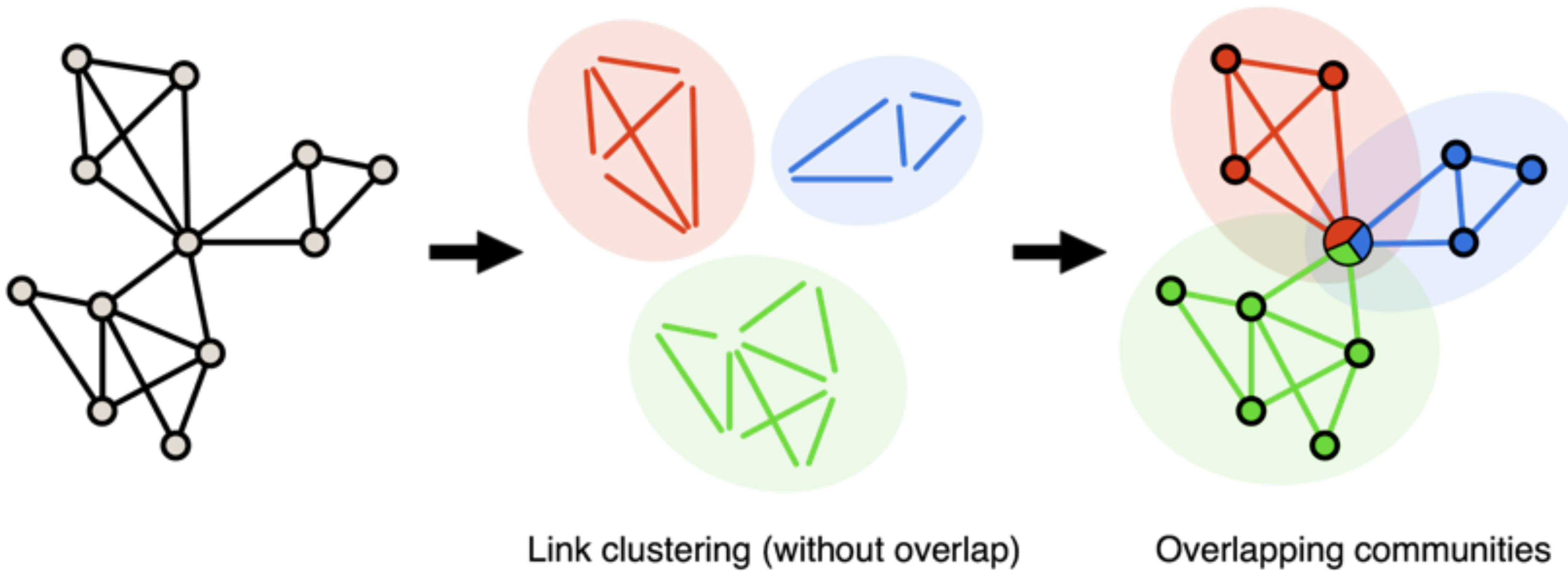
Clauset, Moore, Newman. Nature 453, 98 (2008)

rchy and community
cture were thought to be
o sides of the same story





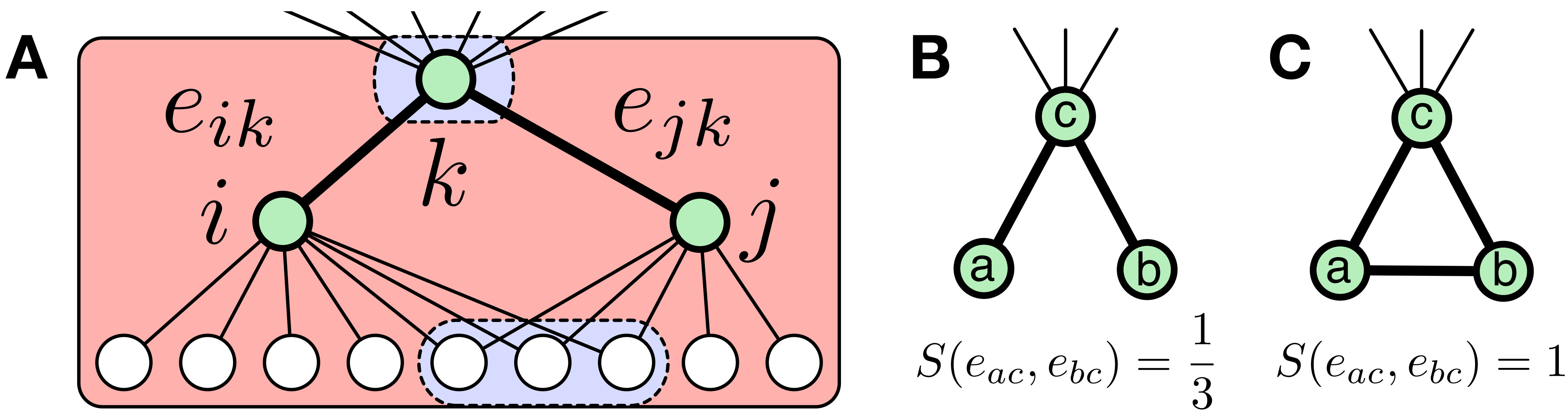
Clauset, Moore, Newman. Nature 453, 98 (2008)

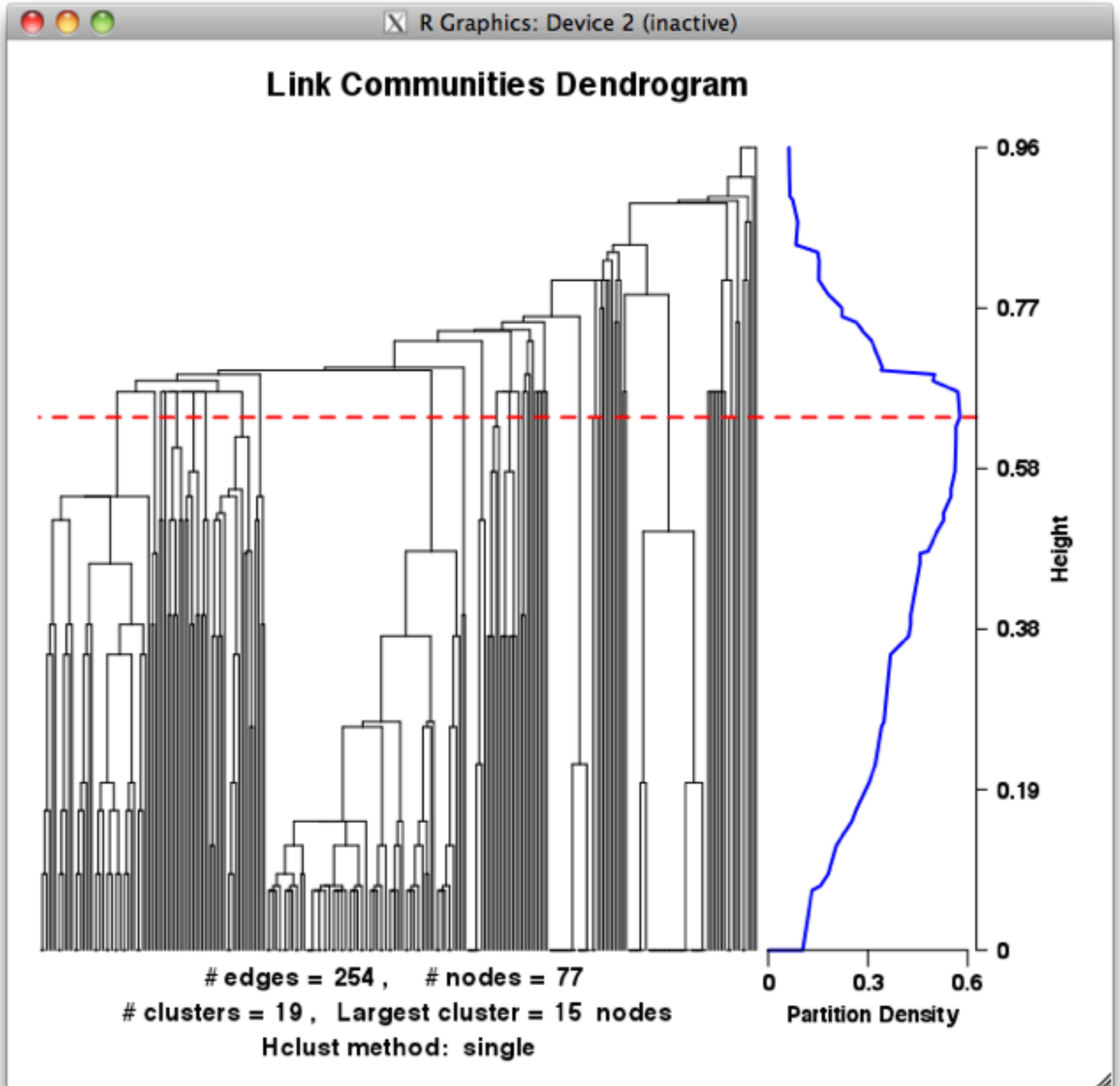


Even in the case when nodes belong to multiple communities, their ***links*** can be well categorized.

$$n_+(i) \equiv \{x \mid d(i,x) \leq 1\}$$

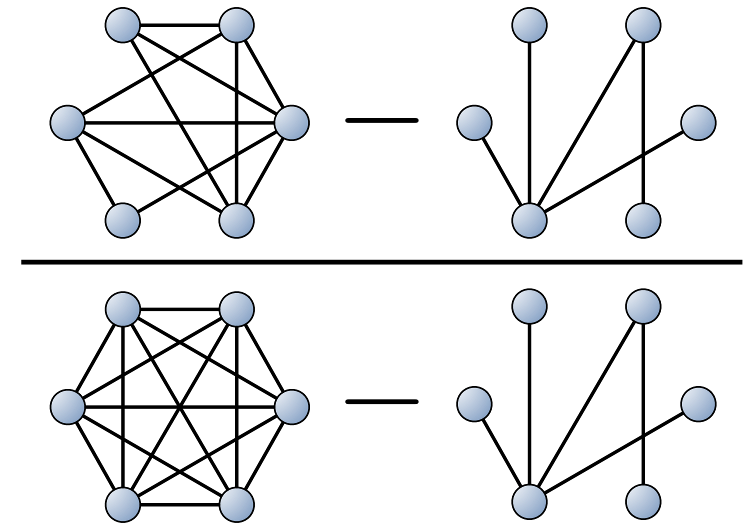
$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$$



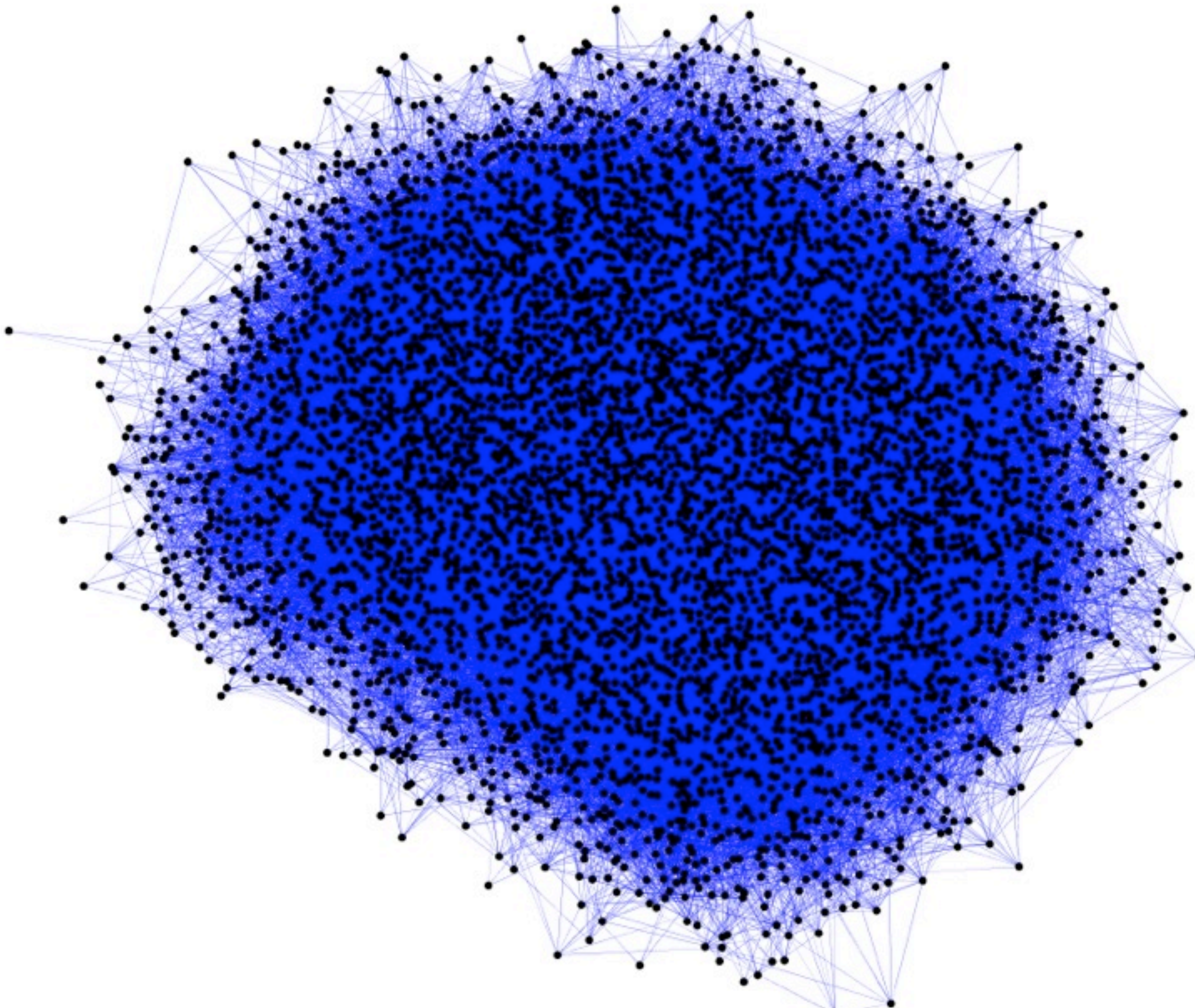


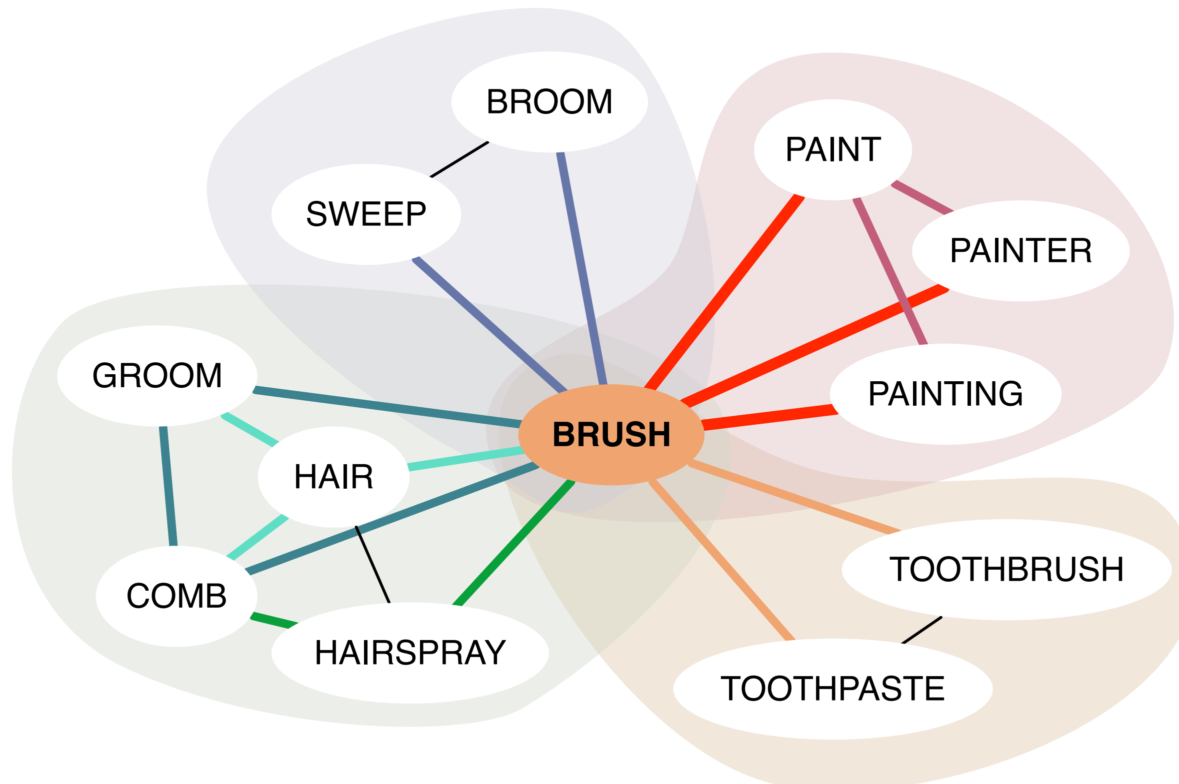
LinkComm R Package by Alex T. Kalinka

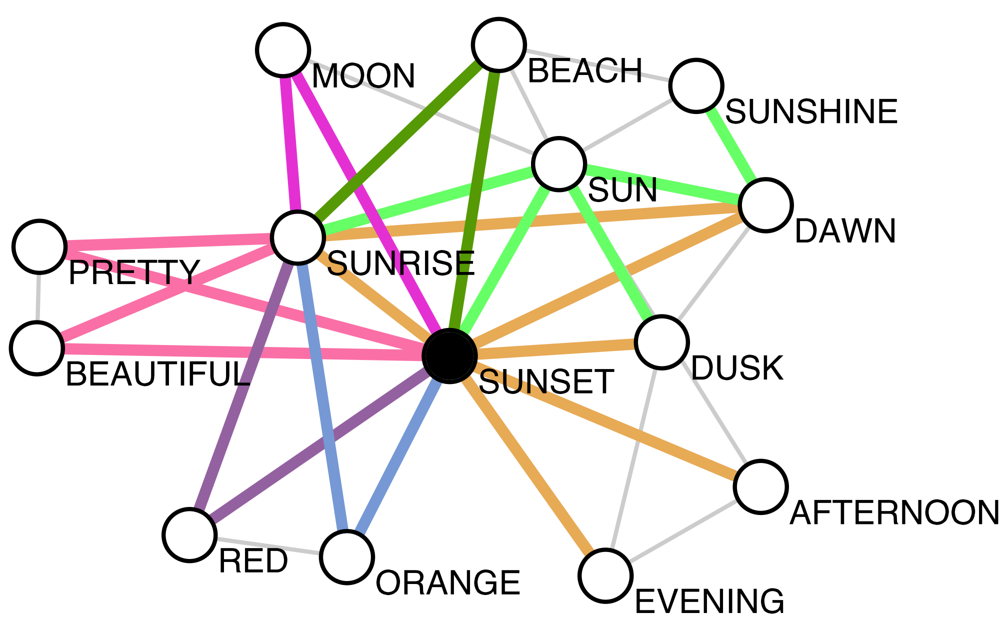
$$D_c = \frac{m_c - (n_c - 1)}{\frac{n_c(n_c-1)}{2} - (n_c - 1)}$$



$$\begin{aligned} D &= \sum_c \frac{m_c}{M} D_c \\ &= \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 1)(n_c - 2)} \end{aligned}$$

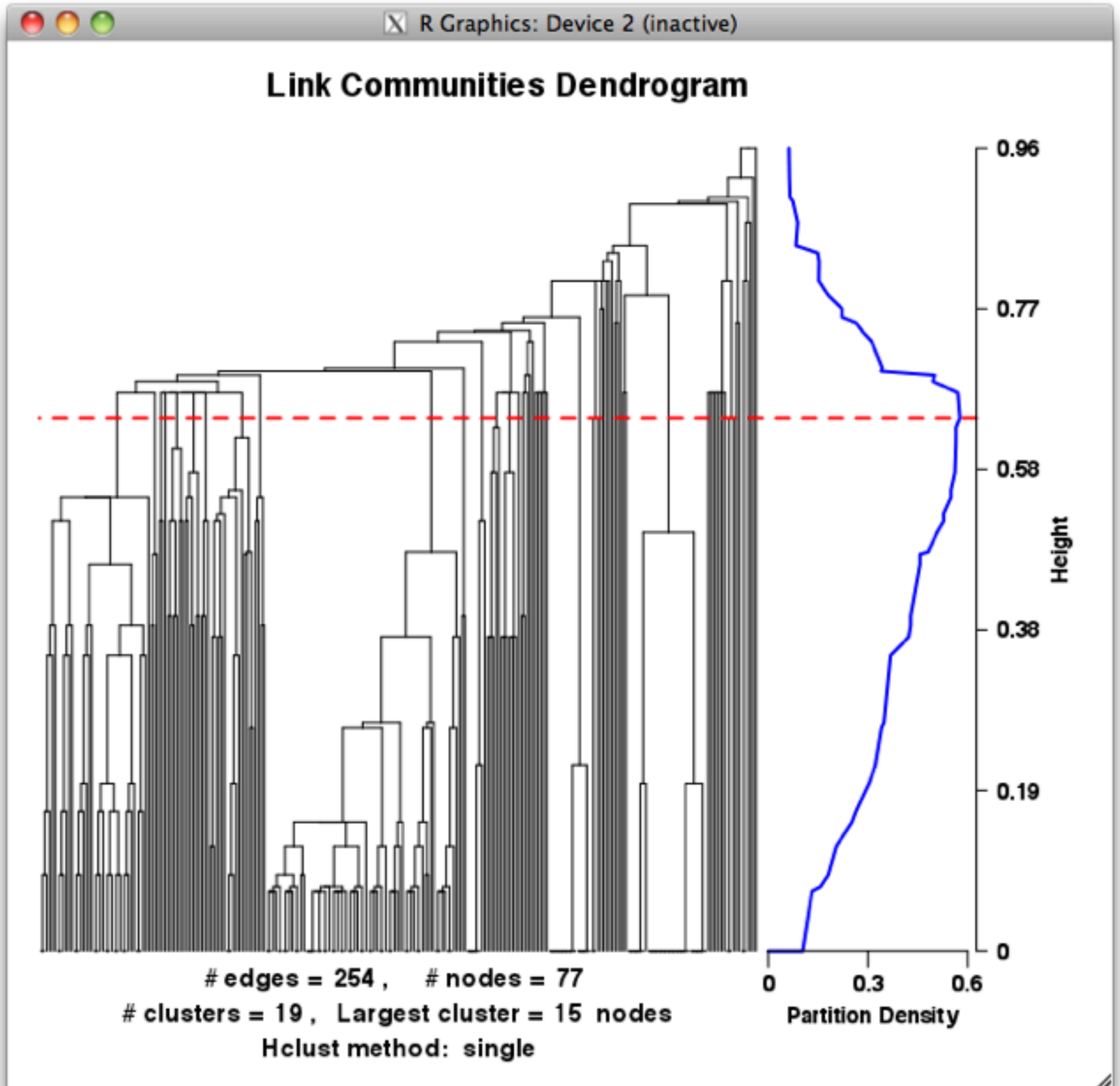




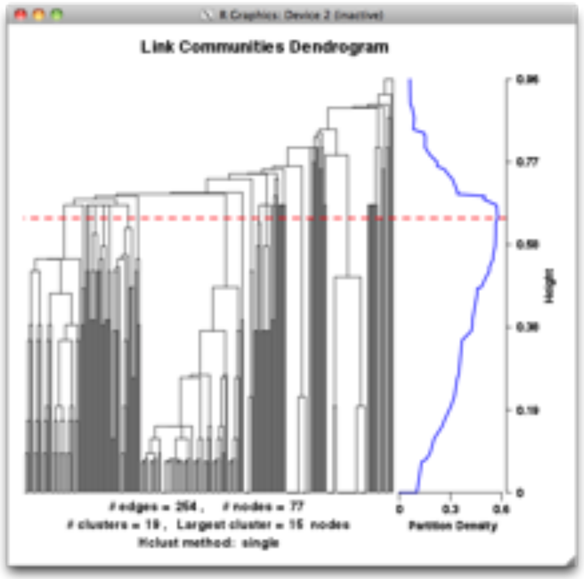


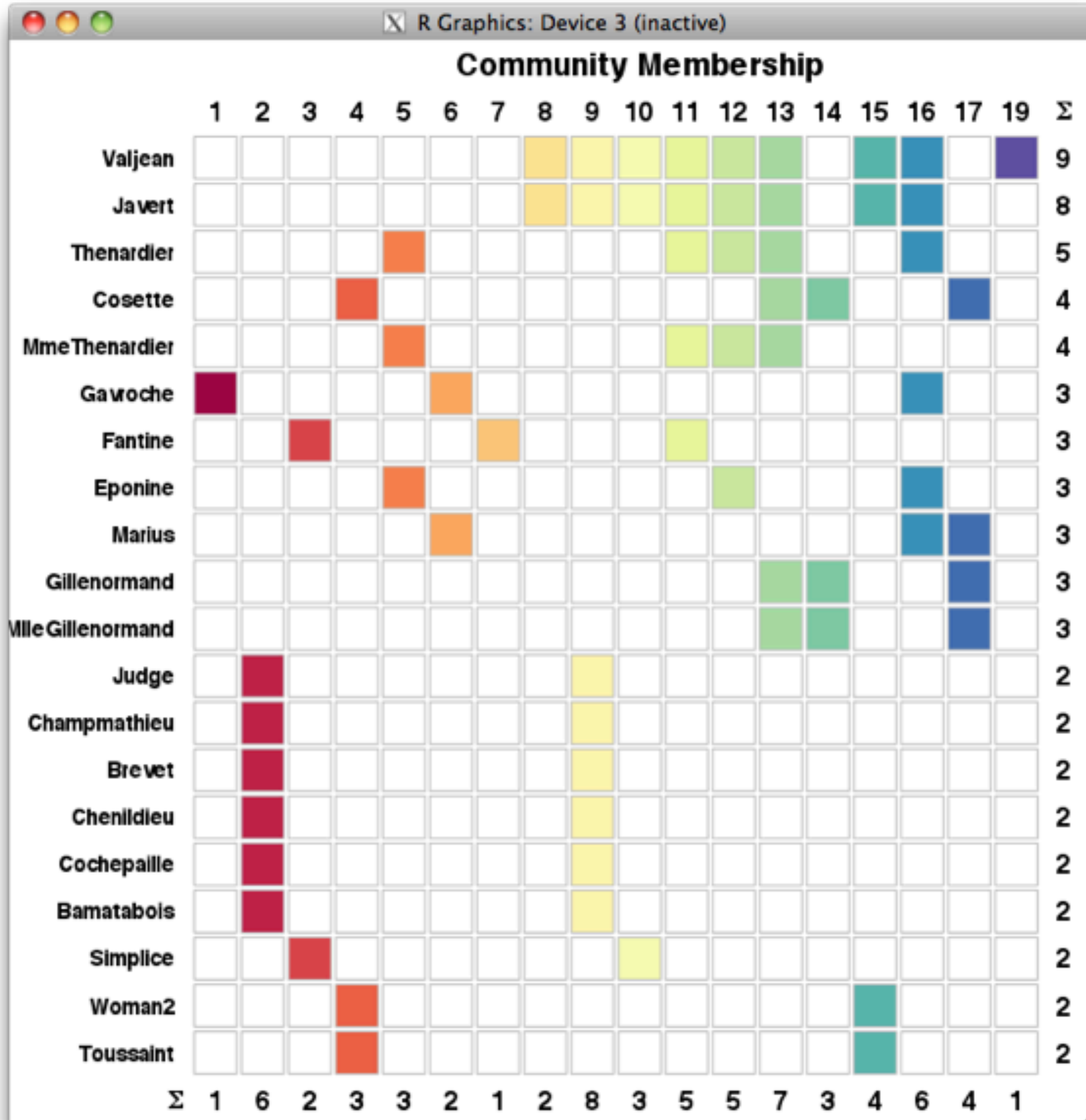
- **SUNSET**, SUNRISE, ORANGE
- **SUNSET**, SUNRISE, RED
- **SUNSET**, SUNRISE, PRETTY, BEAUTIFUL
- **SUNSET**, SUNRISE, MOON
- **SUNSET**, SUNRISE, BEACH
- **SUNSET**, SUNRISE, SUN, DAWN, DUSK, SUNSHINE
- **SUNSET**, SUNRISE, DAWN, DUSK, AFTERNOON, EVENING

LinkComm R Package by Alex T. Kalinka (Pavel Tomancak's group)

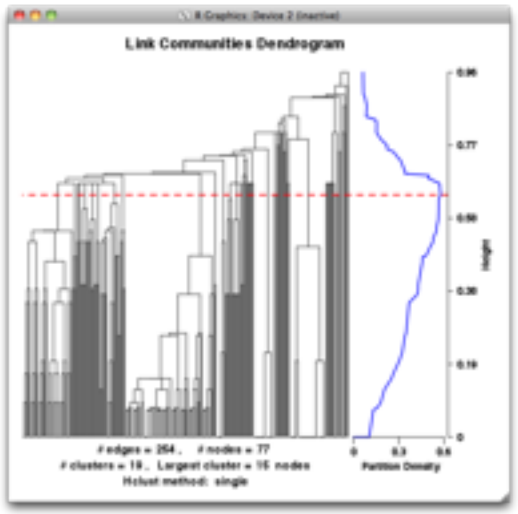


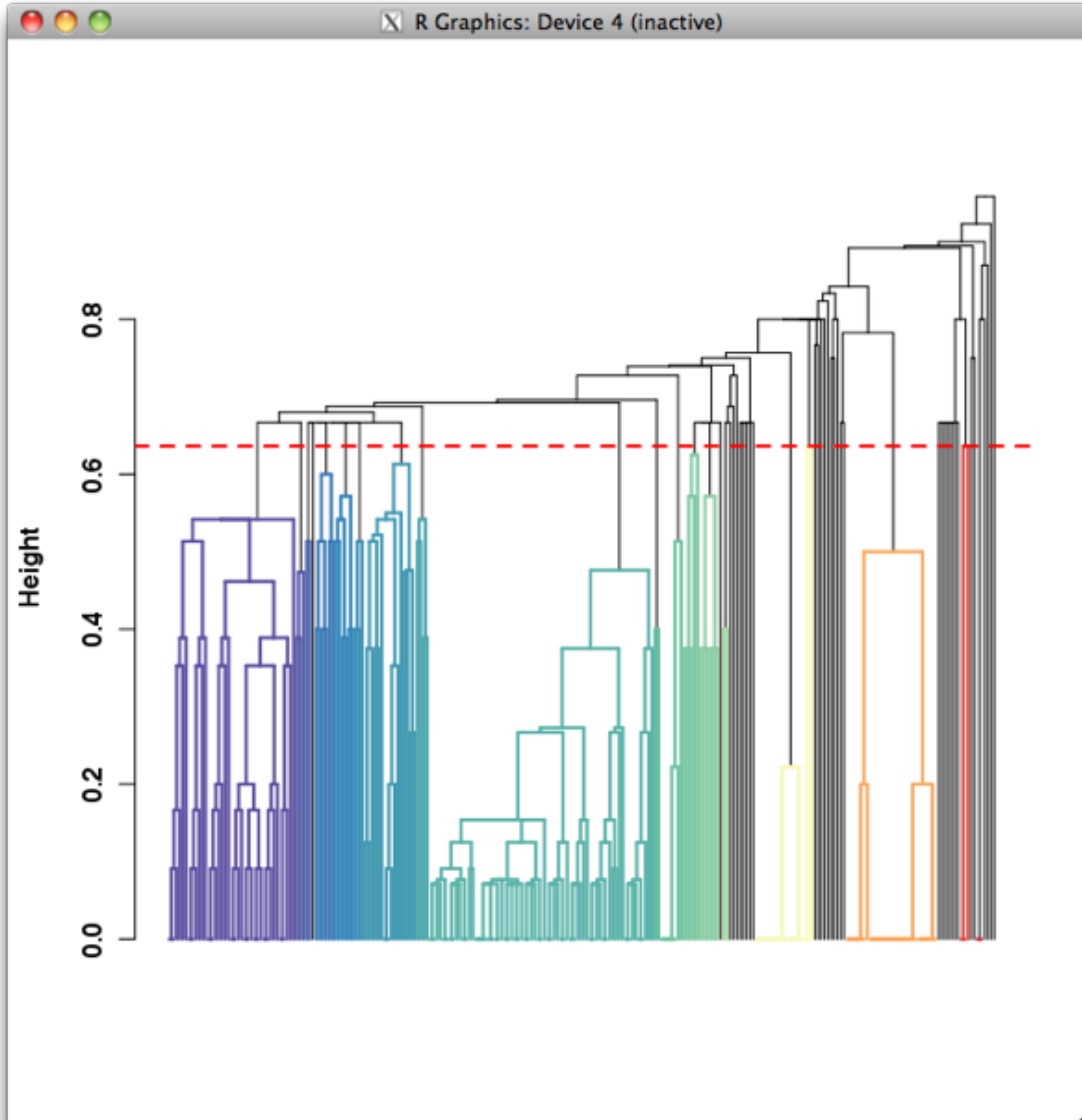
LinkComm R Package by Alex T. Kalinka (Pavel Tomancak's group)



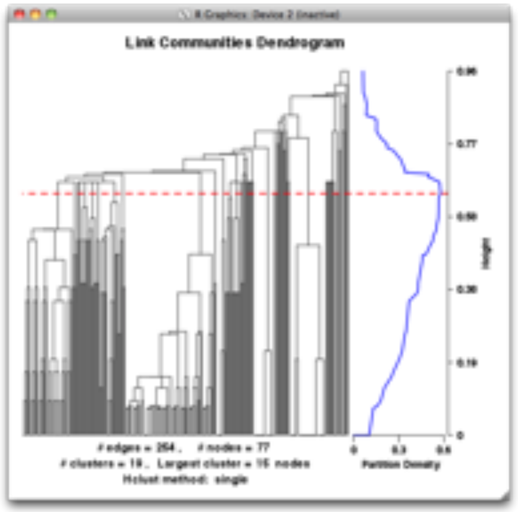


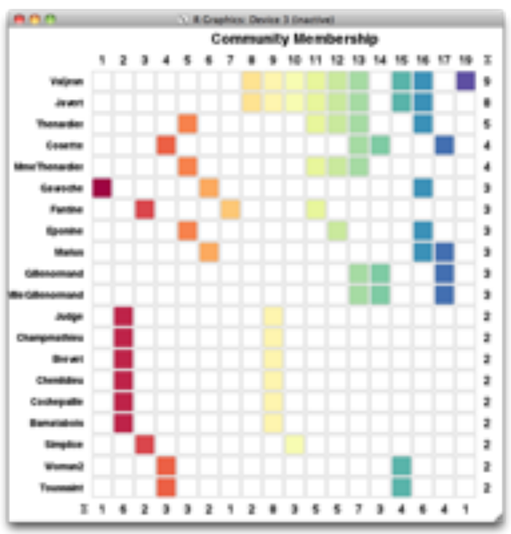
LinkComm R Package by Alex T. Kalinka (Pavel Tomancak's group)

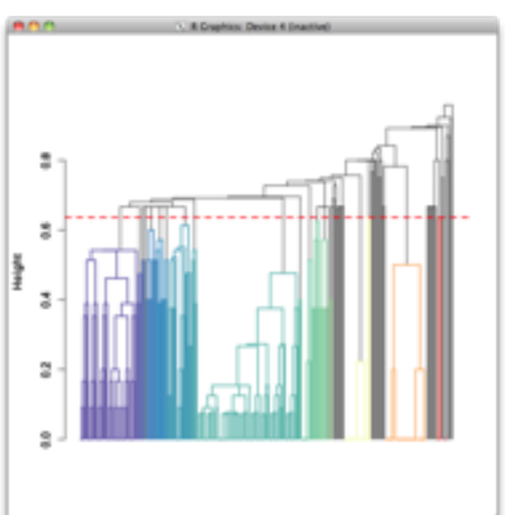


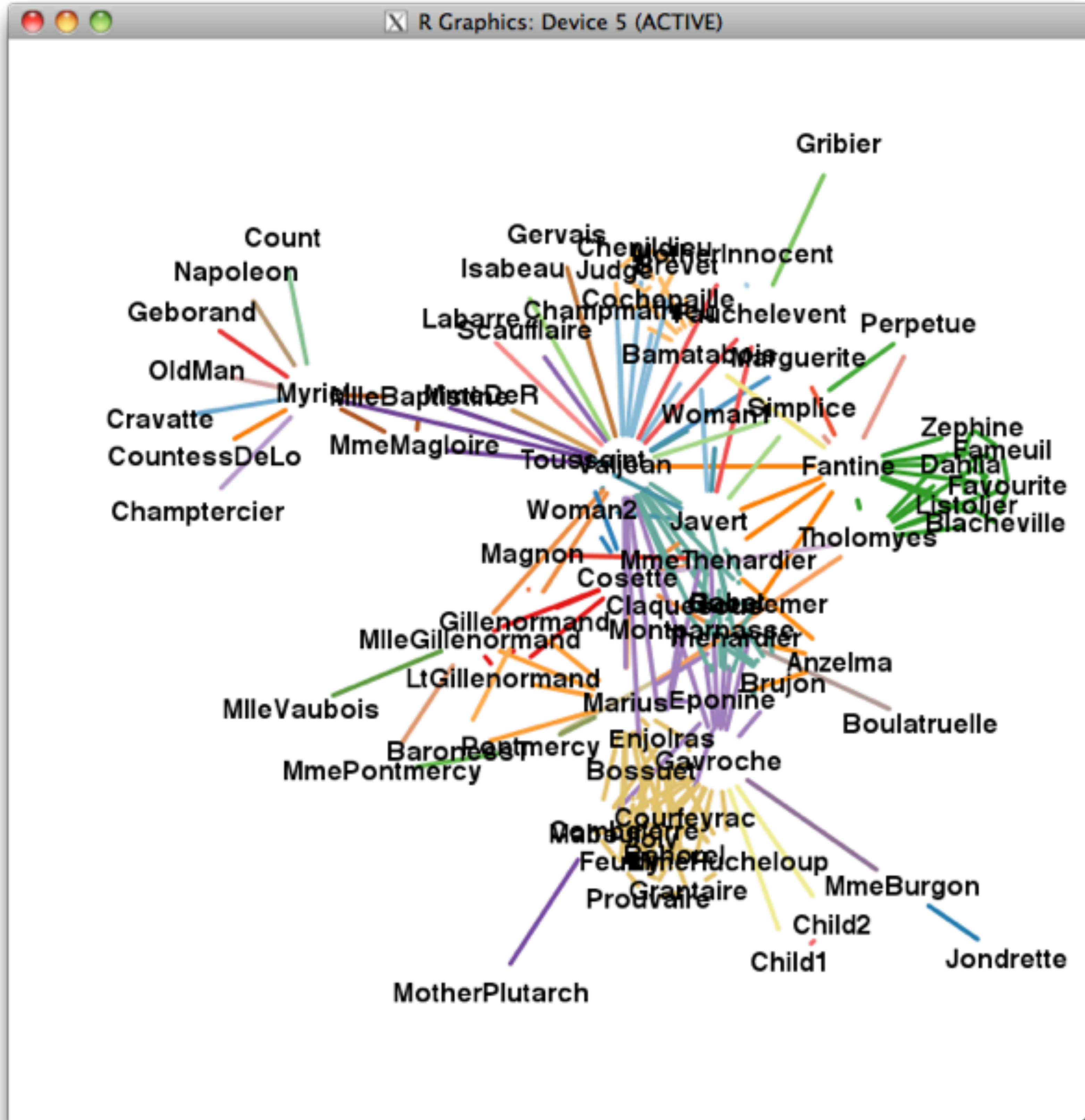


LinkComm R Package by Alex T. Kalinka (Pavel Tomancak's group)

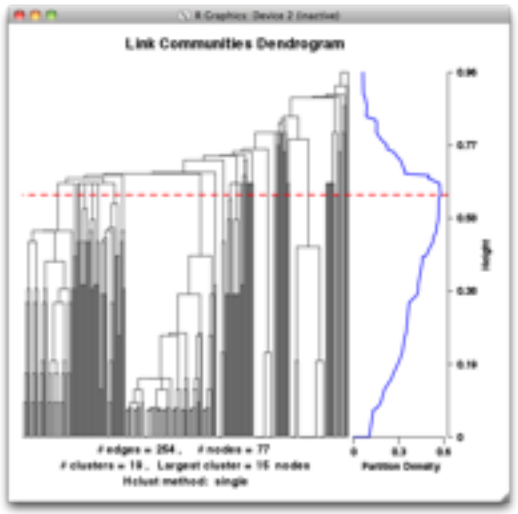


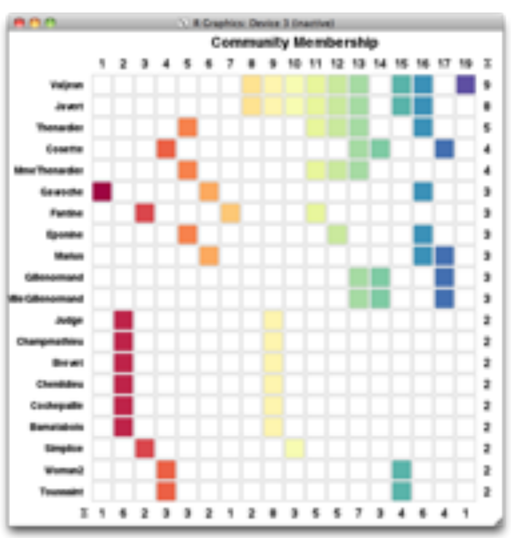


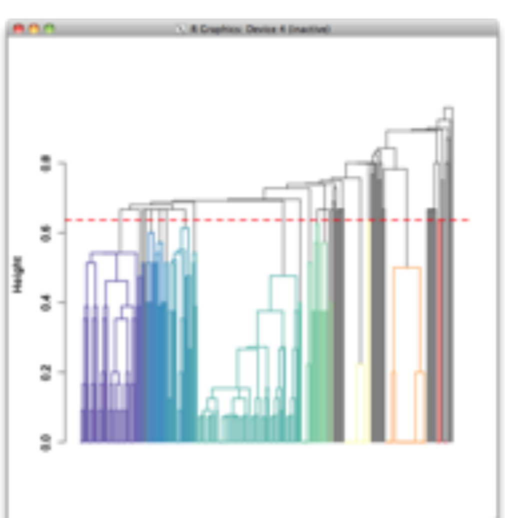


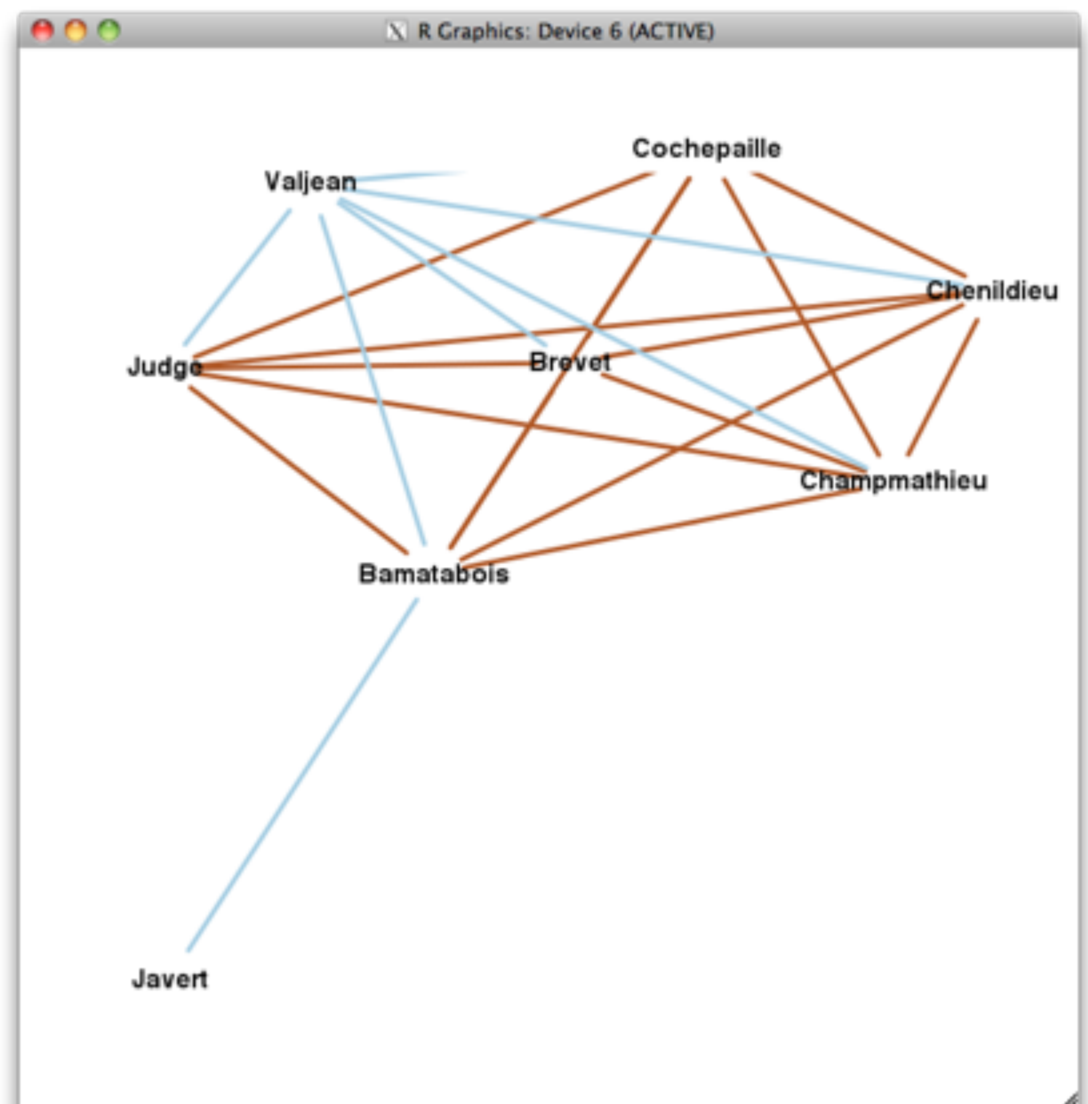


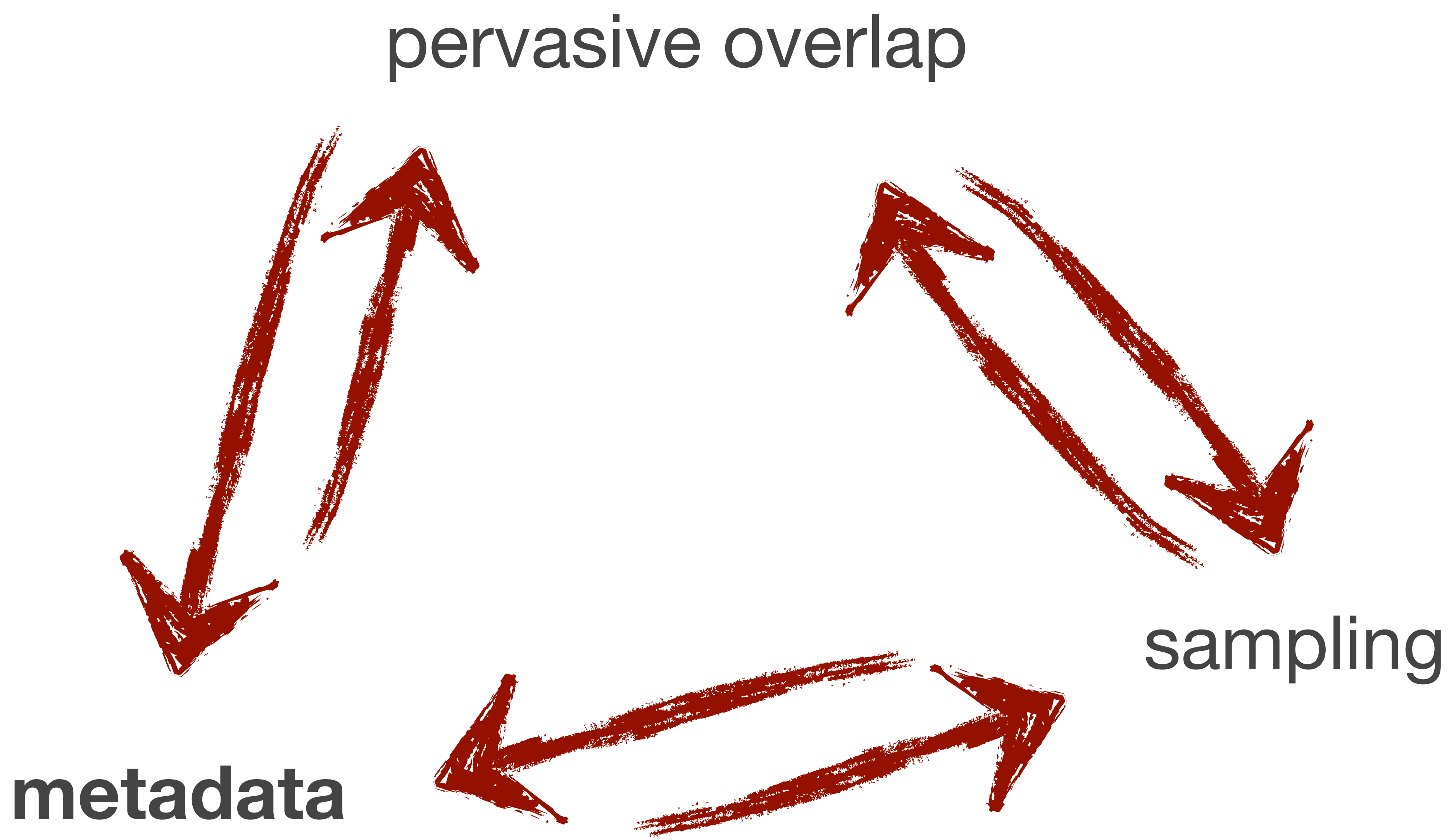
LinkComm R Package by Alex T. Kalinka (Pavel Tomancak's group)

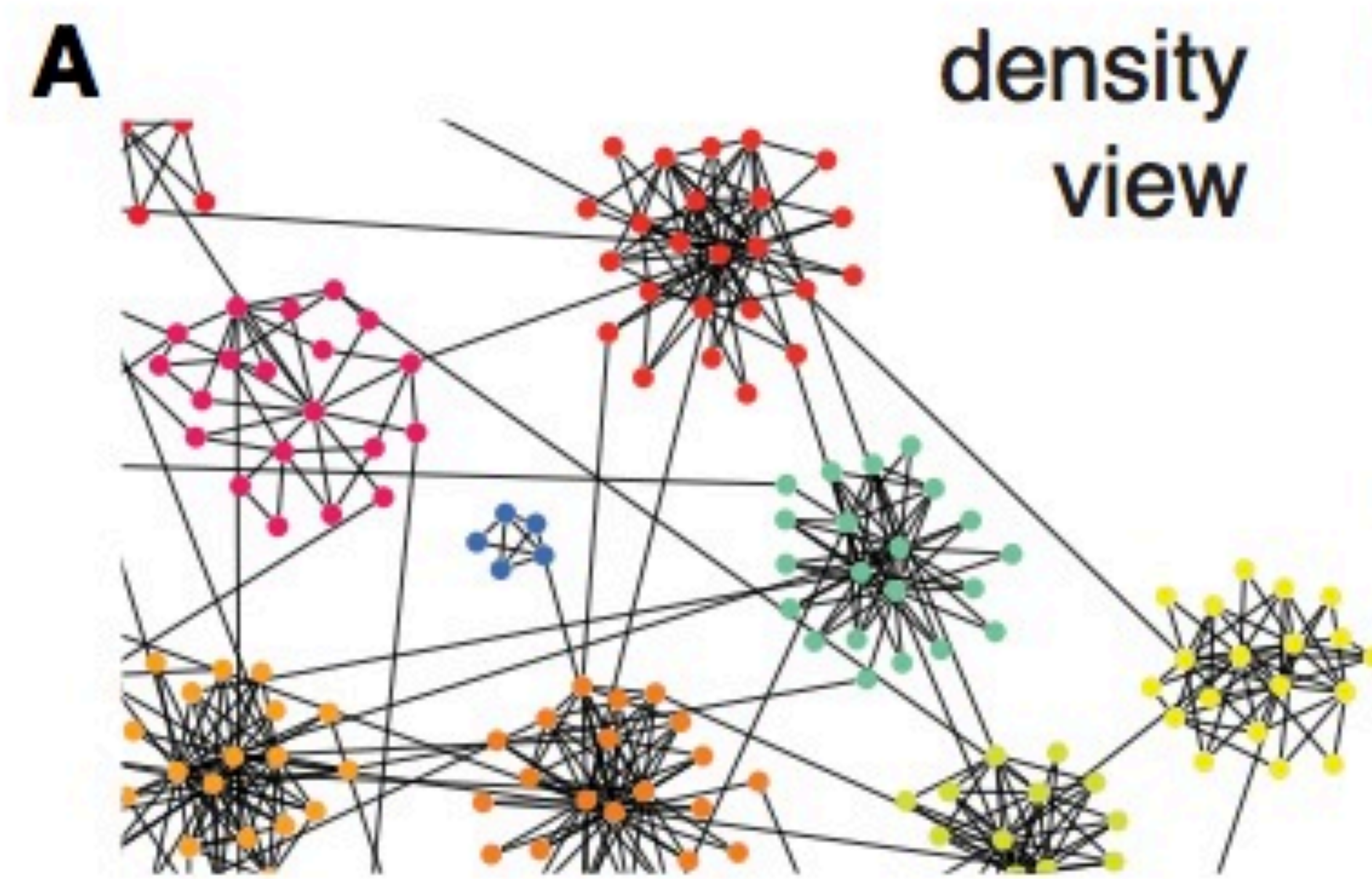




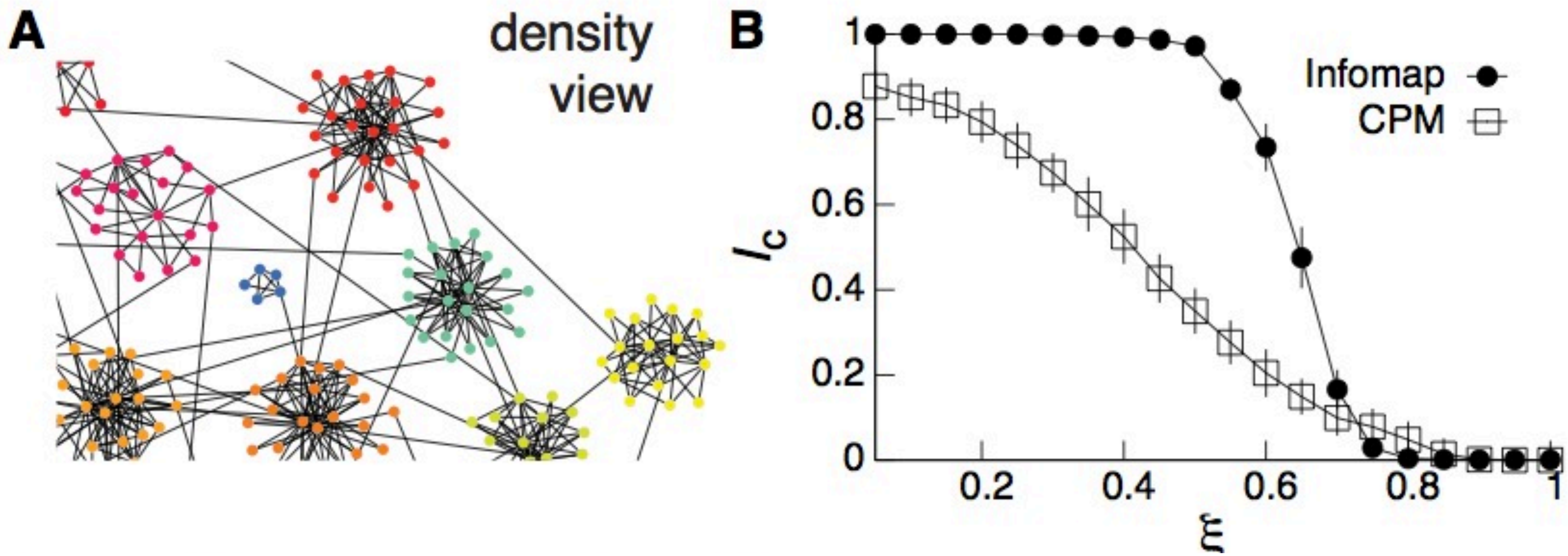




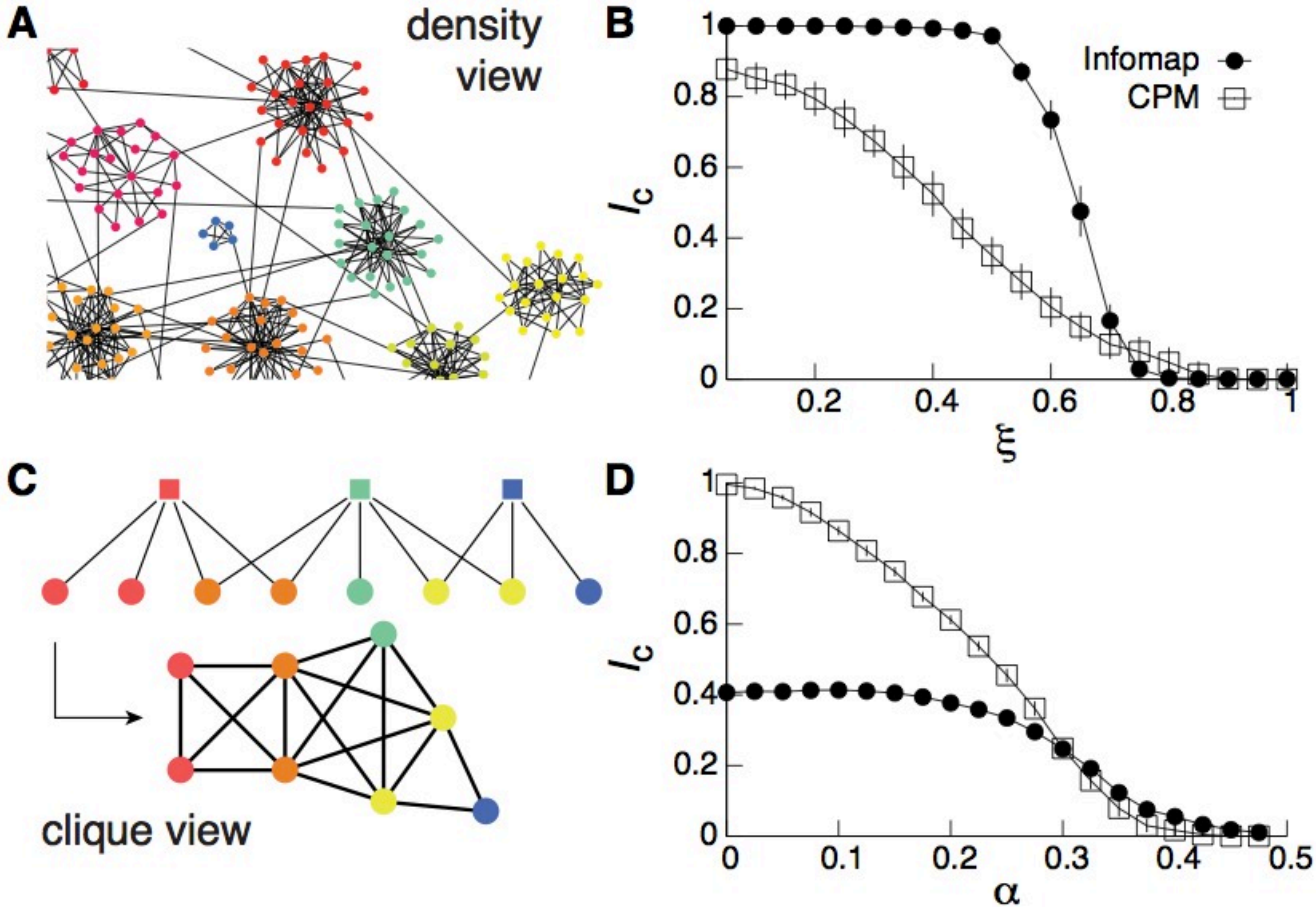

pervasive overlap
sampling
metadata

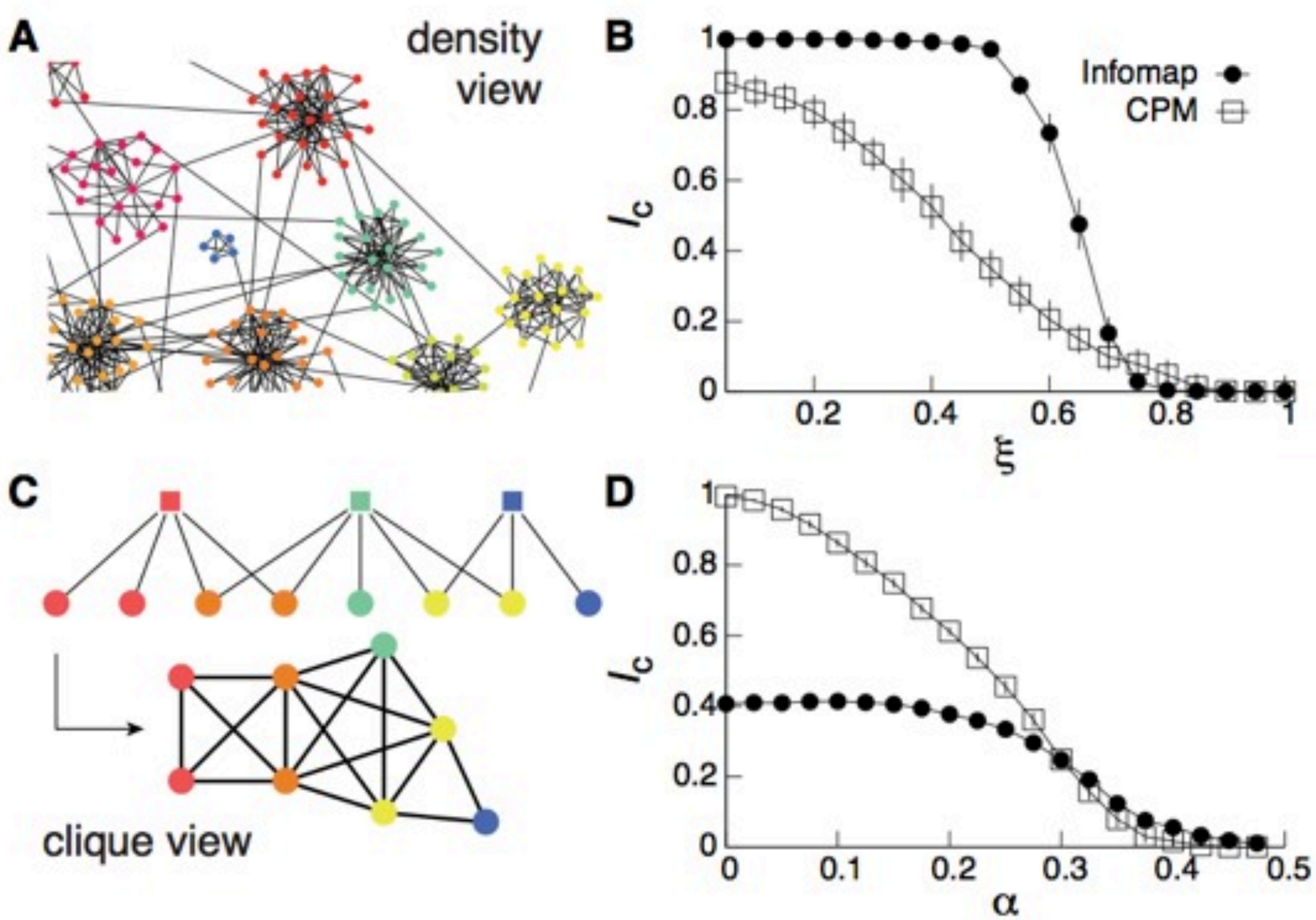


Building a (synthetic) benchmark graph assumes a model. A specific view of the community structure.



A
density
view
B
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
I_c
ξ
Infomap
CPM


A
density view
B
1
0.8
0.6
0.4
0.2
0
I_c
0.2
0.4
0.6
0.8
1
ξ
Infomap
CPM
C
clique view
D
1
0.8
0.6
0.4
0.2
0
I_c
0
0.1
0.2
0.3
0.4
0.5
α



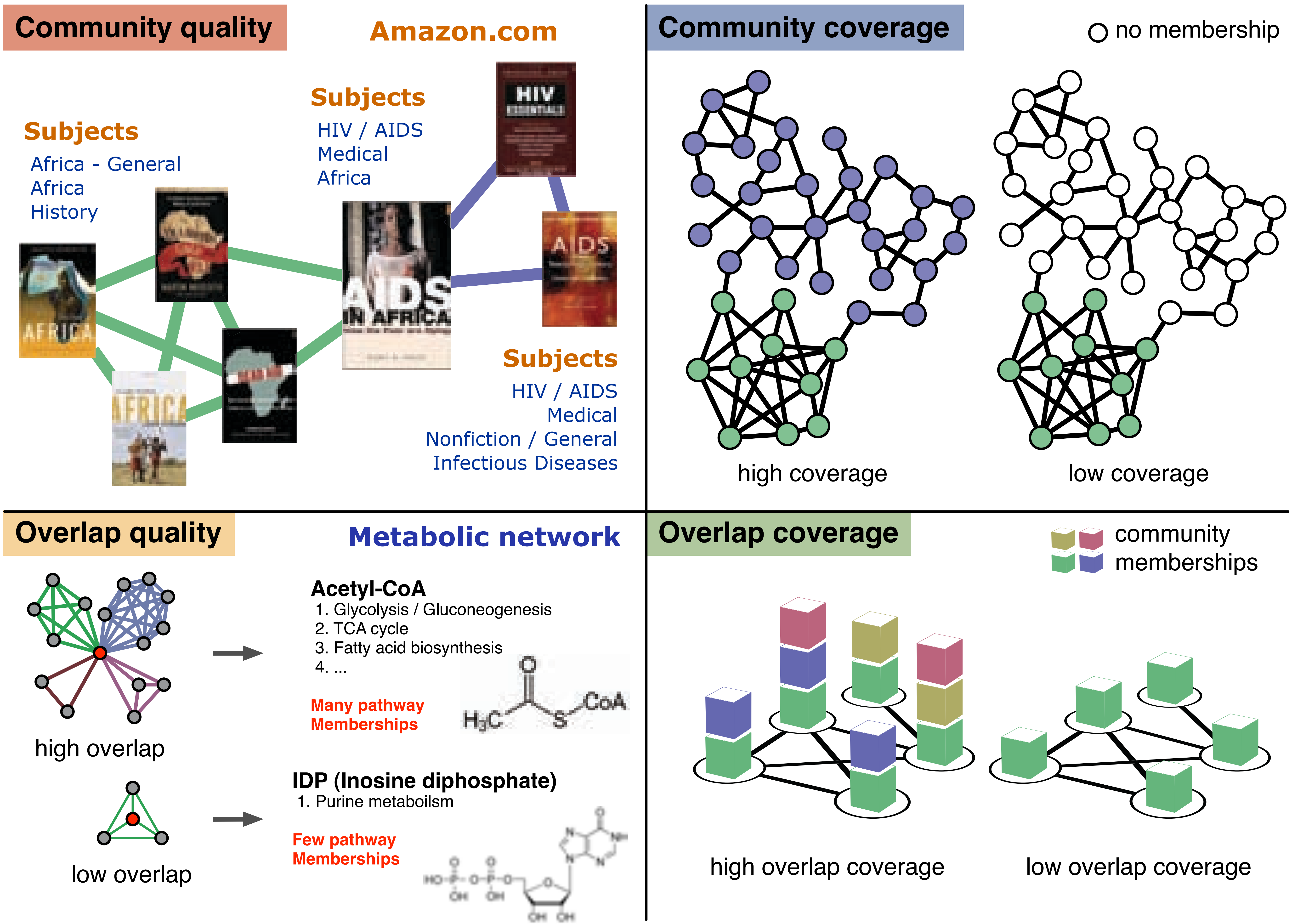
A synthetic graph cannot be used to compare methods with different *models* of community structure.

Problem:

A synthetic graph cannot be used to compare methods with different *models* of community structure.

Solution:

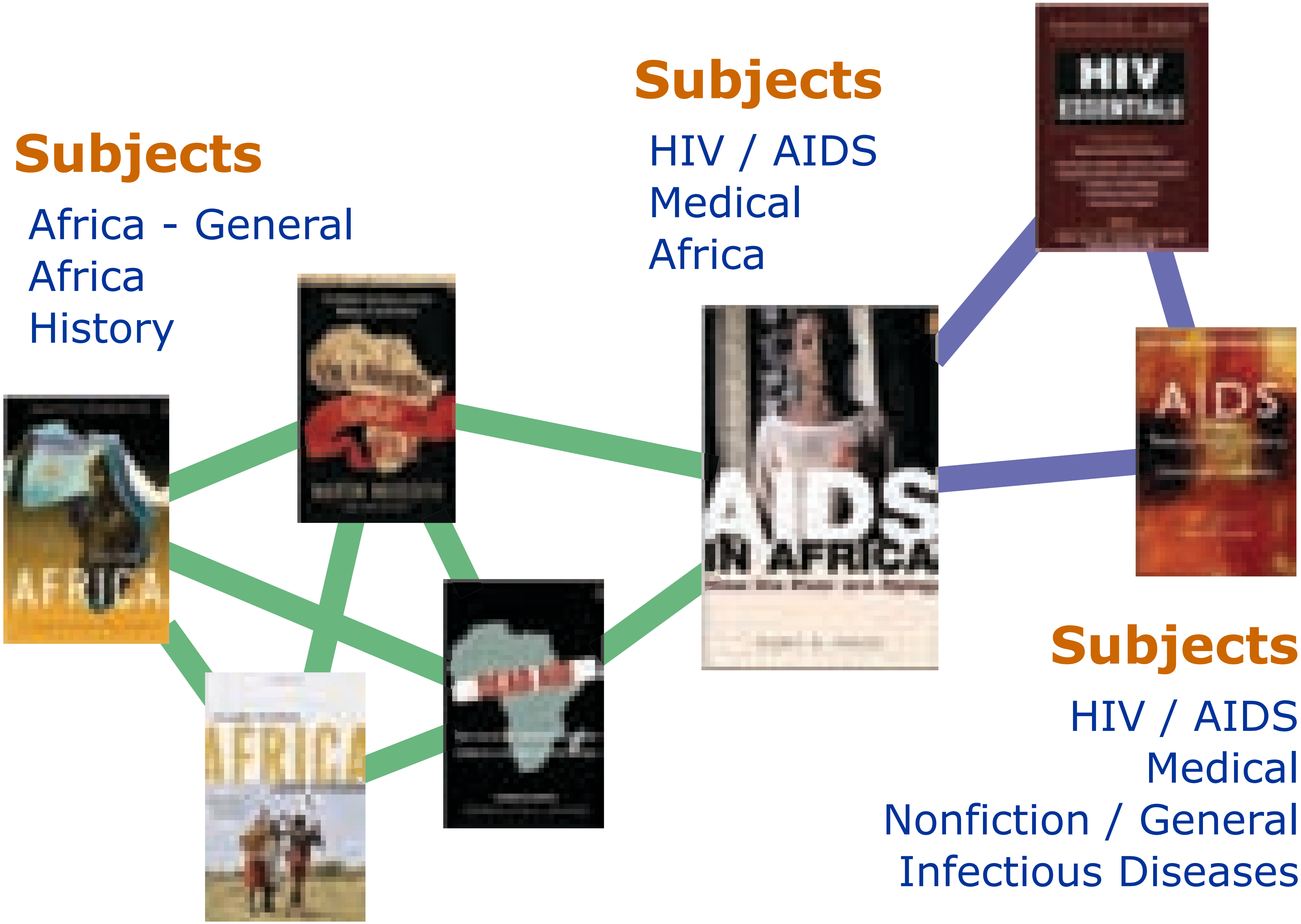
Use metadata to test the detected structure.



Community quality
Amazon.com
Subjects
Africa - General
Africa
History
Subjects
HIV / AIDS
Medical
Africa
HIV
AIDS IN AFRICA
AIDS
AFRICA
Subjects
HIV / AIDS
Medical
Nonfiction / General
Infectious Diseases
Community coverage
no membership
high coverage
low coverage
Overlap quality
Metabolic network
Acetyl-CoA
1. Glycolysis / Gluconeogenesis
2. TCA cycle
3. Fatty acid biosynthesis
4. ...
Many pathway Memberships
high overlap
IDP (Inosine diphosphate)
1. Purine metaboilsm
Few pathway Memberships
low overlap
Overlap coverage
community memberships
high overlap coverage
low overlap coverage

Community quality

Amazon.com



Subjects

Africa - General
Africa
History

Subjects

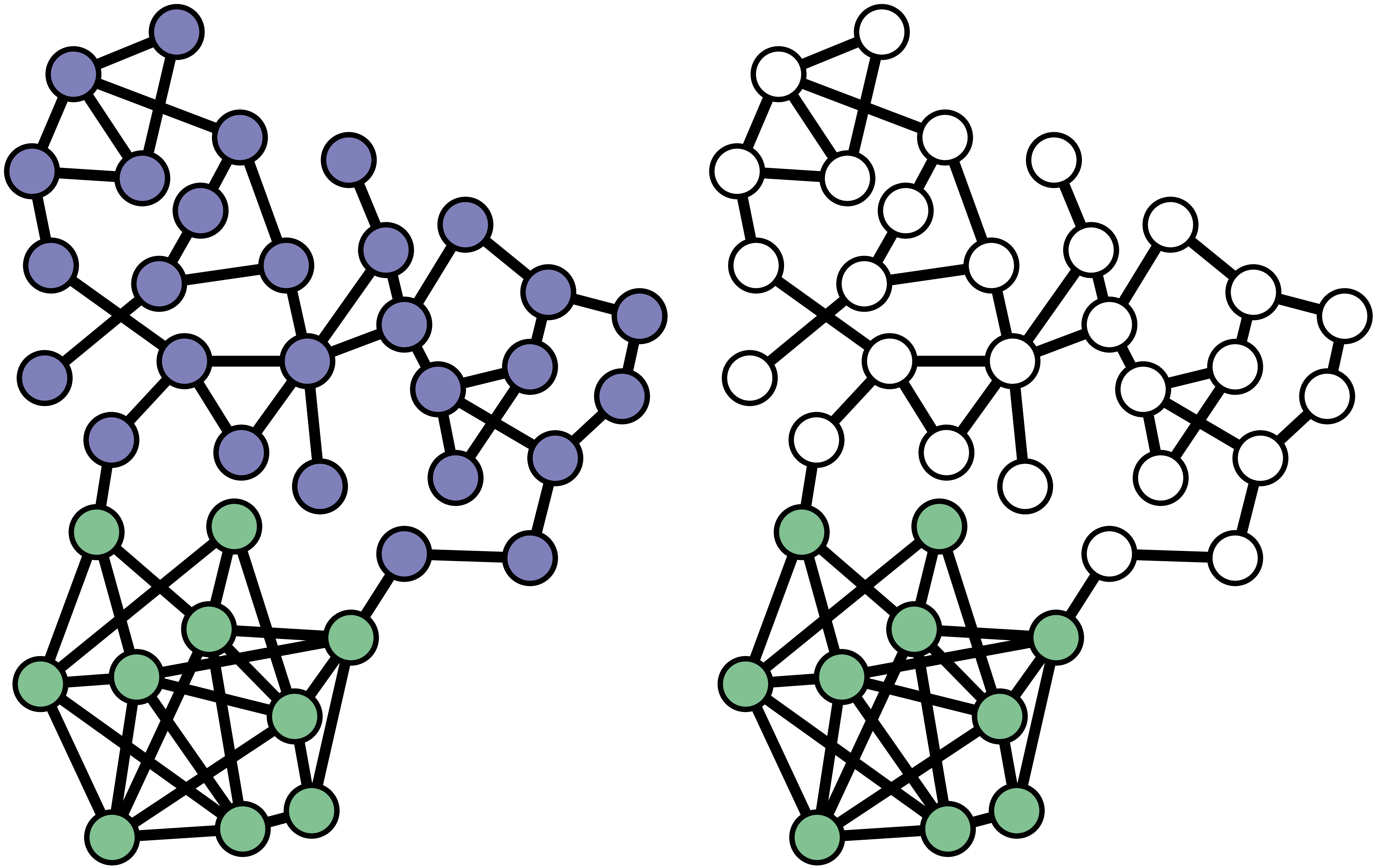
HIV / AIDS
Medical
Africa

Subjects

HIV / AIDS
Medical
Nonfiction / General
Infectious Diseases

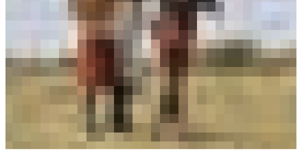
Community coverage

○ no membership



high coverage

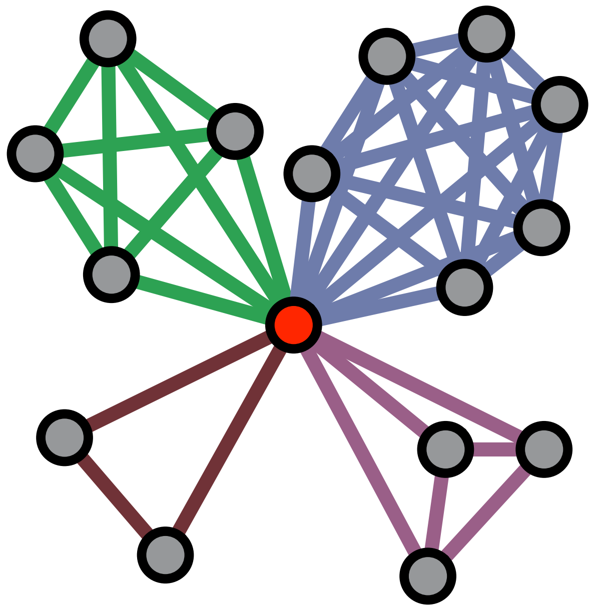
low coverage



Infectious Diseases

Overlap quality

Metabolic network

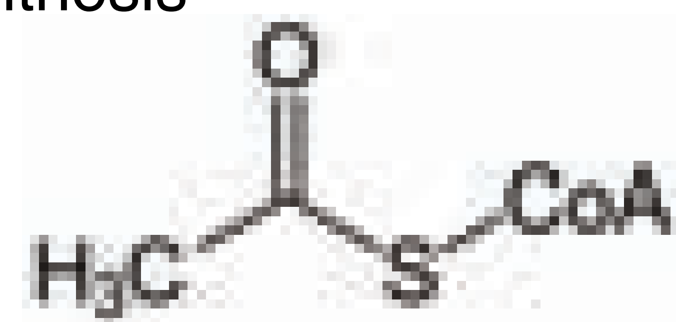


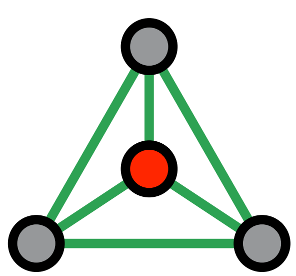
high overlap

Acetyl-CoA

1. Glycolysis / Gluconeogenesis
2. TCA cycle
3. Fatty acid biosynthesis
4. ...

Many pathway Memberships



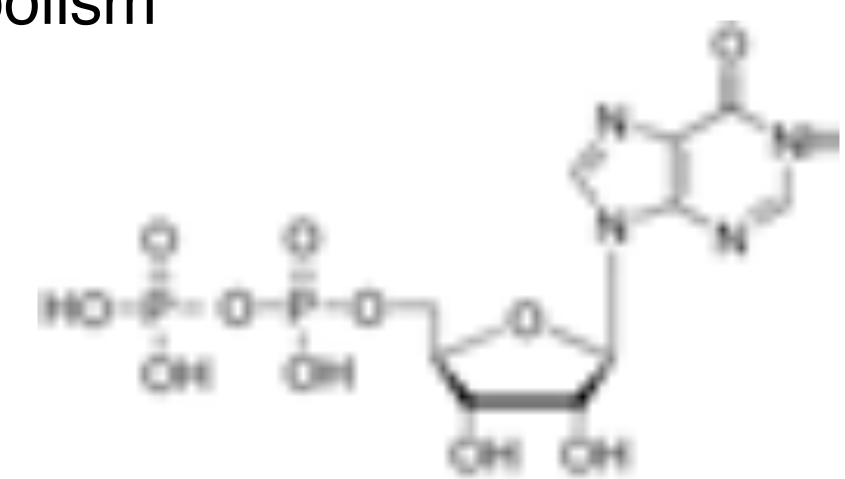


low overlap

IDP (Inosine diphosphate)

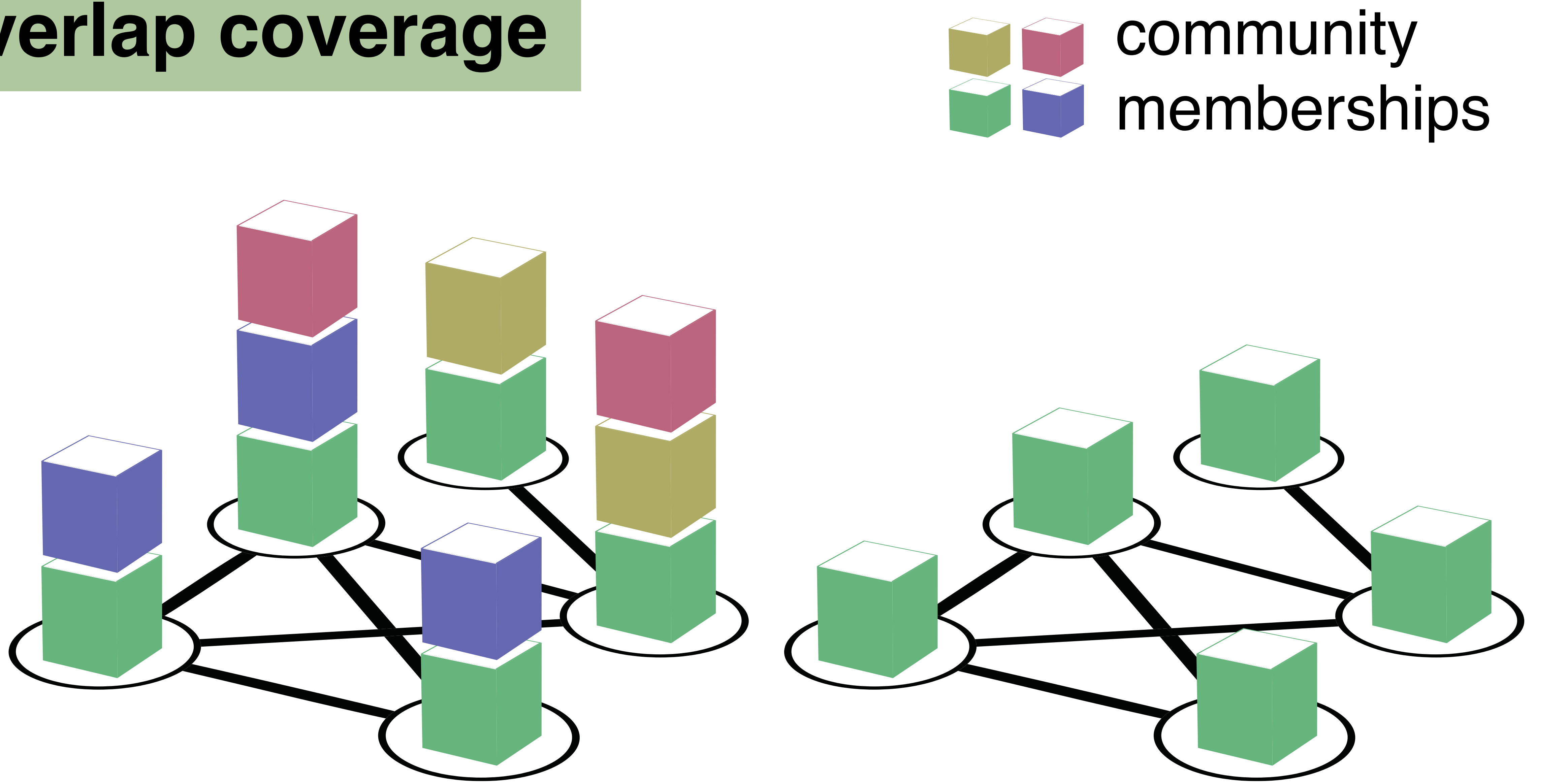
1. Purine metaboilsm

Few pathway Memberships



	high coverage	low coverage

Overlap coverage



high overlap coverage

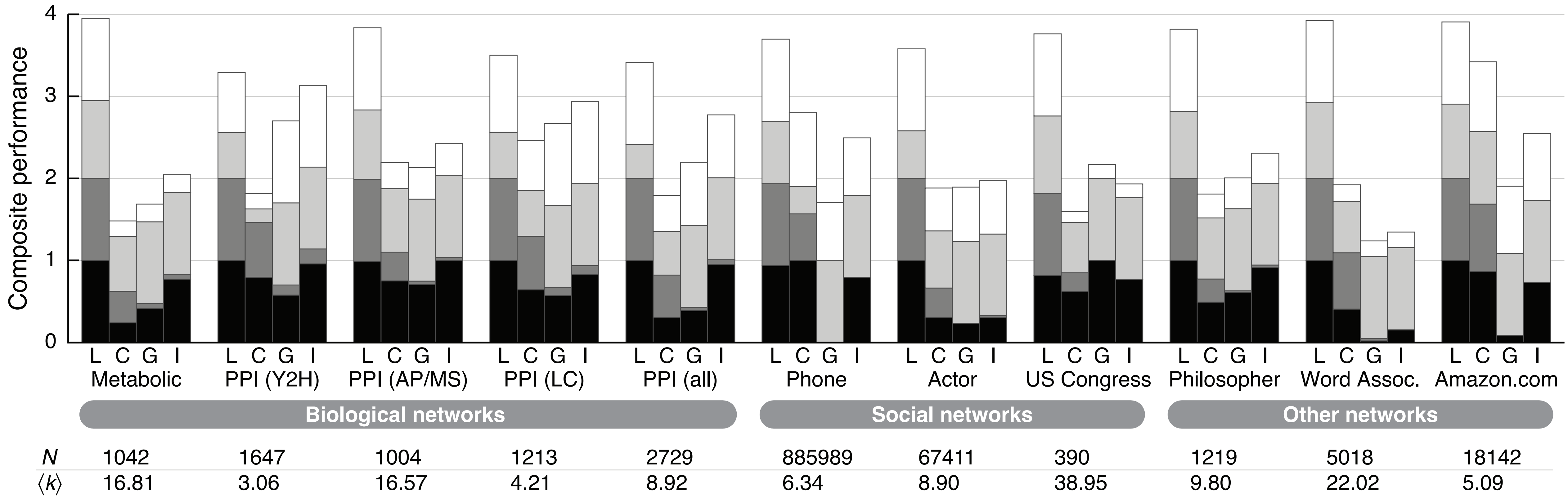
low overlap coverage

network	description	N	$\langle k \rangle$	metadata: community	metadata: overlap
PPI (Y2H)	PPI network of *S. cerevisiae* obtained by yeast two-hybrid (Y2H) experiment [41]	1647	3.06	Set of each protein's known functions (GO terms)[a]	The number of GO terms
PPI (AP/MS)	Affinity purification mass spectrometry (AP/MS) experiment	1004	16.57	GO terms	GO terms
PPI (LC)	Literature curated (LC)	1213	4.21	GO terms	GO terms
PPI (all)	Union of Y2H, AP/MS, and LC PPI networks	2729	8.92	GO terms	GO-terms
Metabolic	Metabolic network (metabolites connected by reactions) of *E. coli*	1042	16.81	Set of each metabolite's pathway annotations (KEGG)[b]	The number of KEGG pathway annotations
Phone	Social contacts between mobile phone users [44, 45, 46]	885989	6.34	Each user's most likely geographic location	Call activity (number of phone calls)
Actor	Film actors that appear in the same movies during 2000–2009 [47]	67411	8.90	Set of plot keywords for all of the actor's films	Length of career (year of first role)
US Congress	Congressmen who co-sponsor bills during the 108th US Congress [48, 49]	390	38.95	Political ideology, from the common space score [50, 51]	Seniority (number of congresses served)
Philosopher	Philosophers and their philosophical influences, from the English Wikipedia[c]	1219	9.80	Set of (wikipedia) hyperlinks exiting in the philosopher's page	Number of wikipedia subject categories
Word Assoc.	English words that are often mentally associated [52]	5018	22.02	Set of each word's *senses*, as documented by WordNet[d]	Number of senses
Amazon.com	Products that users frequently buy together	18142	5.09[e]	Set of each product's user tags (annotations)	Number of product categories

Palla, G., Derény, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814 (2005).

Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004).

Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105**, 1118–1123 (2008).

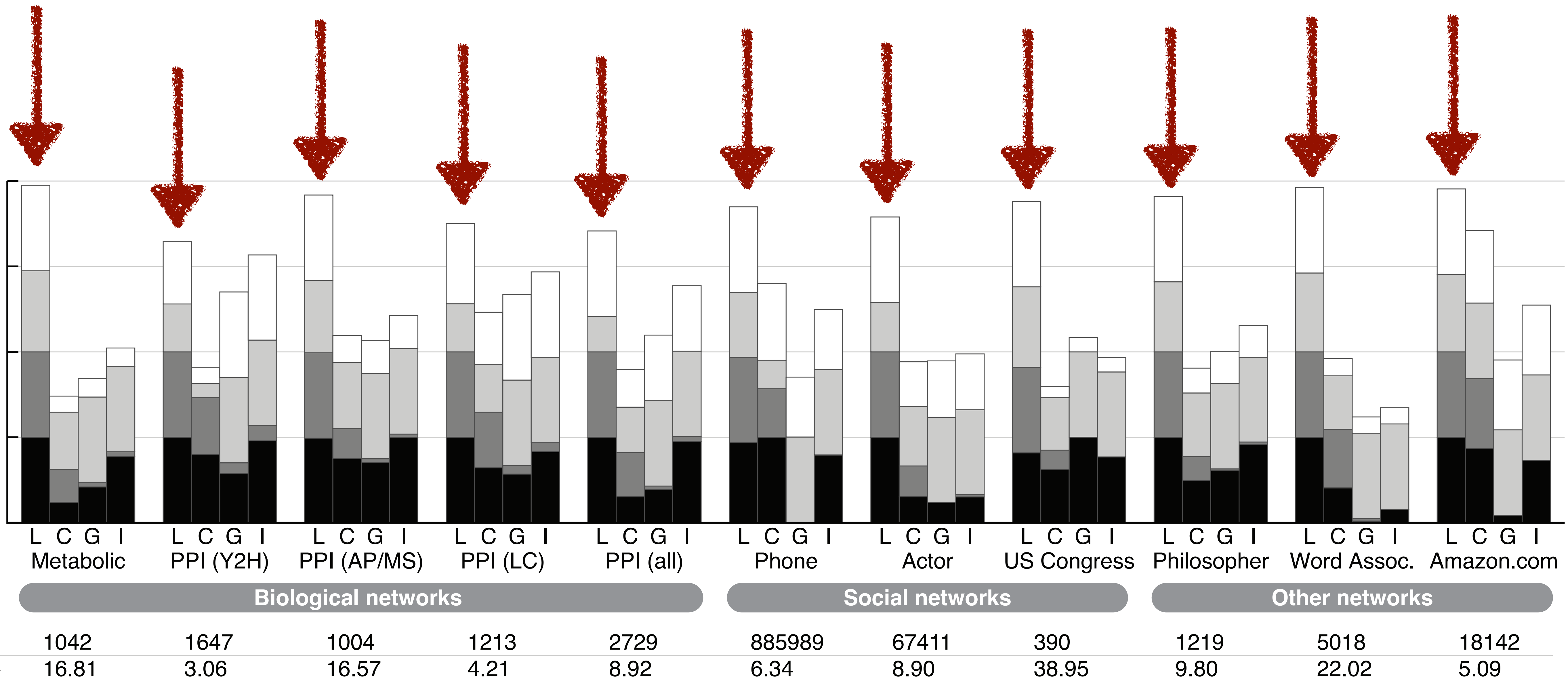


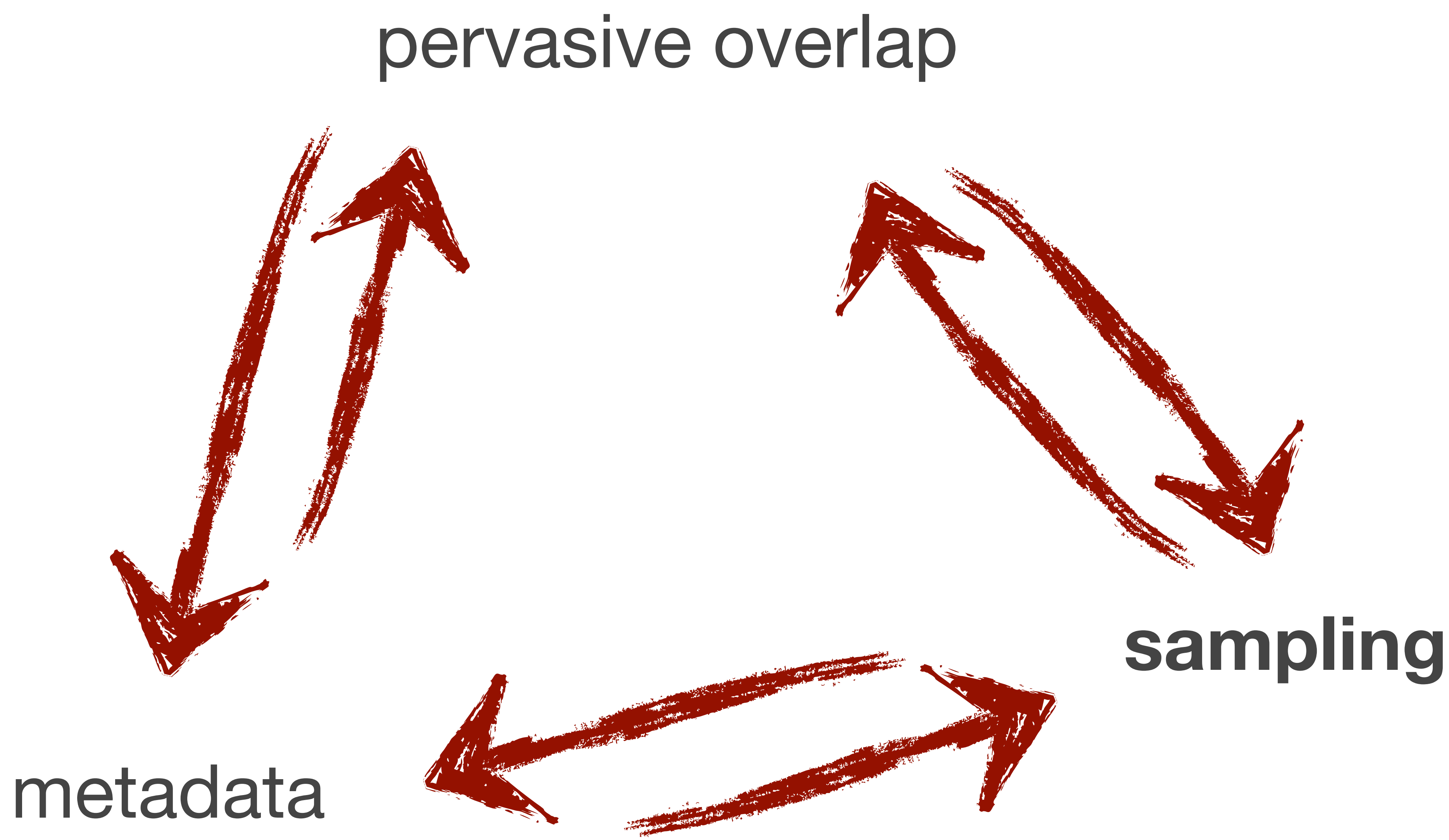
Measures

- overlap coverage
- community coverage
- overlap quality
- community quality

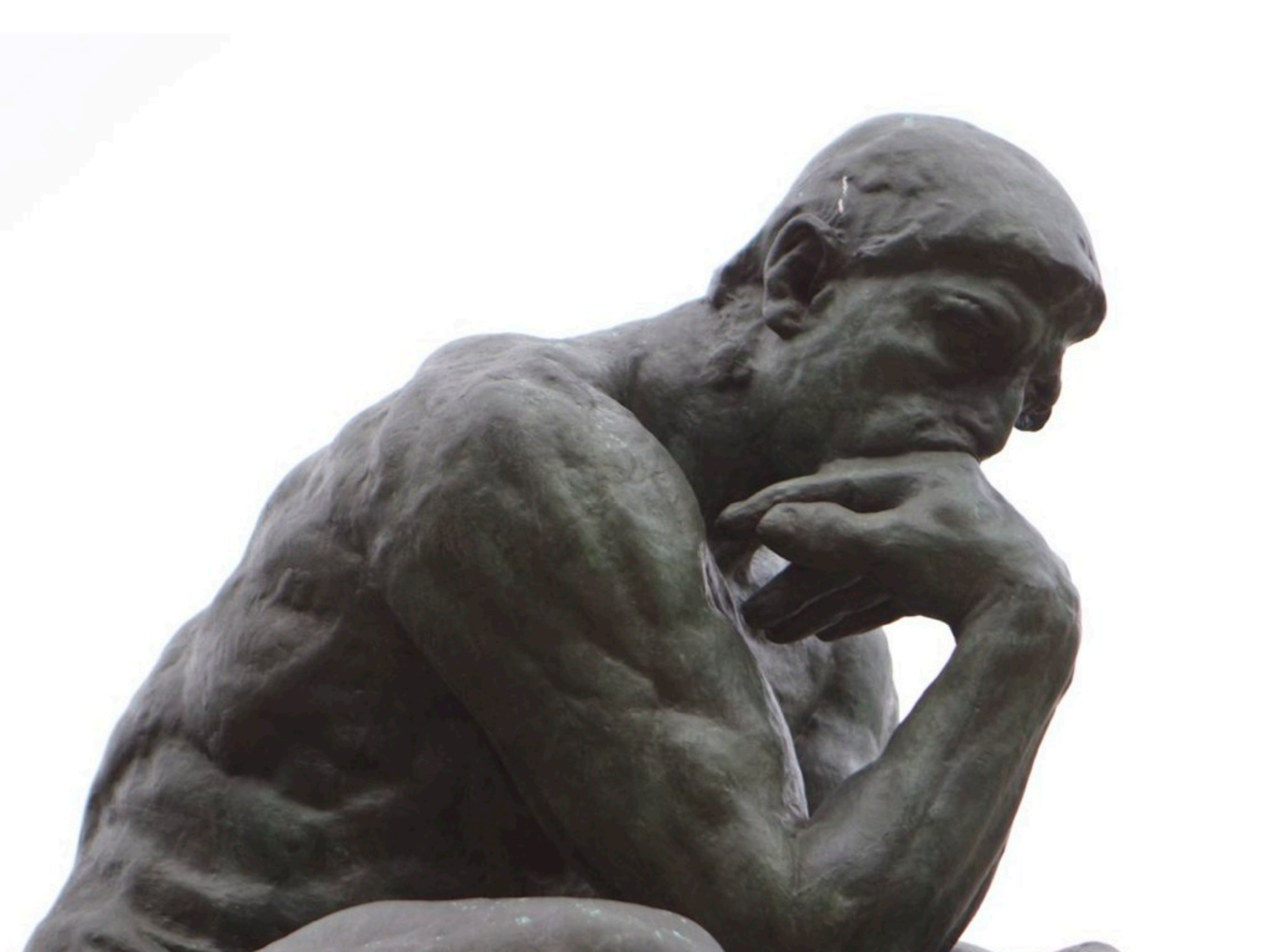
Methods

L – Links
C – Clique Percolation
G – Greedy Modularity
I – Infomap


L C G I
Metabolic
PPI (Y2H)
PPI (AP/MS)
PPI (LC)
PPI (all)
Phone
Actor
US Congress
Philosopher
Word Assoc.
Amazon.com
Biological networks
Social networks
Other networks
1042 1647 1004 1213 2729 885989 67411 390 1219 5018 18142
16.81 3.06 16.57 4.21 8.92 6.34 8.90 38.95 9.80 22.02 5.09

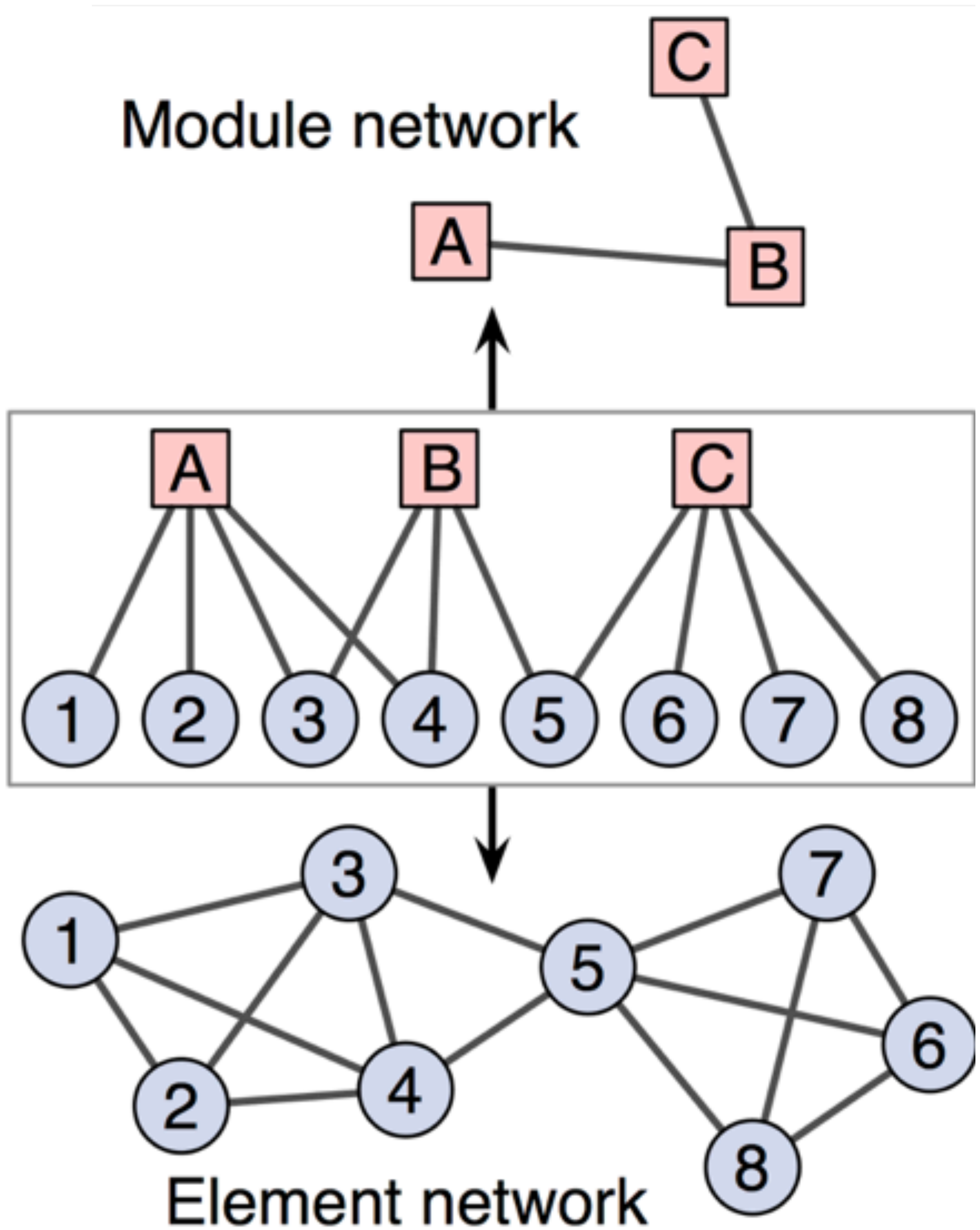

pervasive overlap
sampling
metadata


But Sune, we often find 'good' non-overlapping communities in networks that should possess pervasive overlap according to your argument.
ACID



hmm. could ***sampling*** cause networks with pervasive overlap to ***appear*** non-overlapping?

a simple model for pervasive overlap



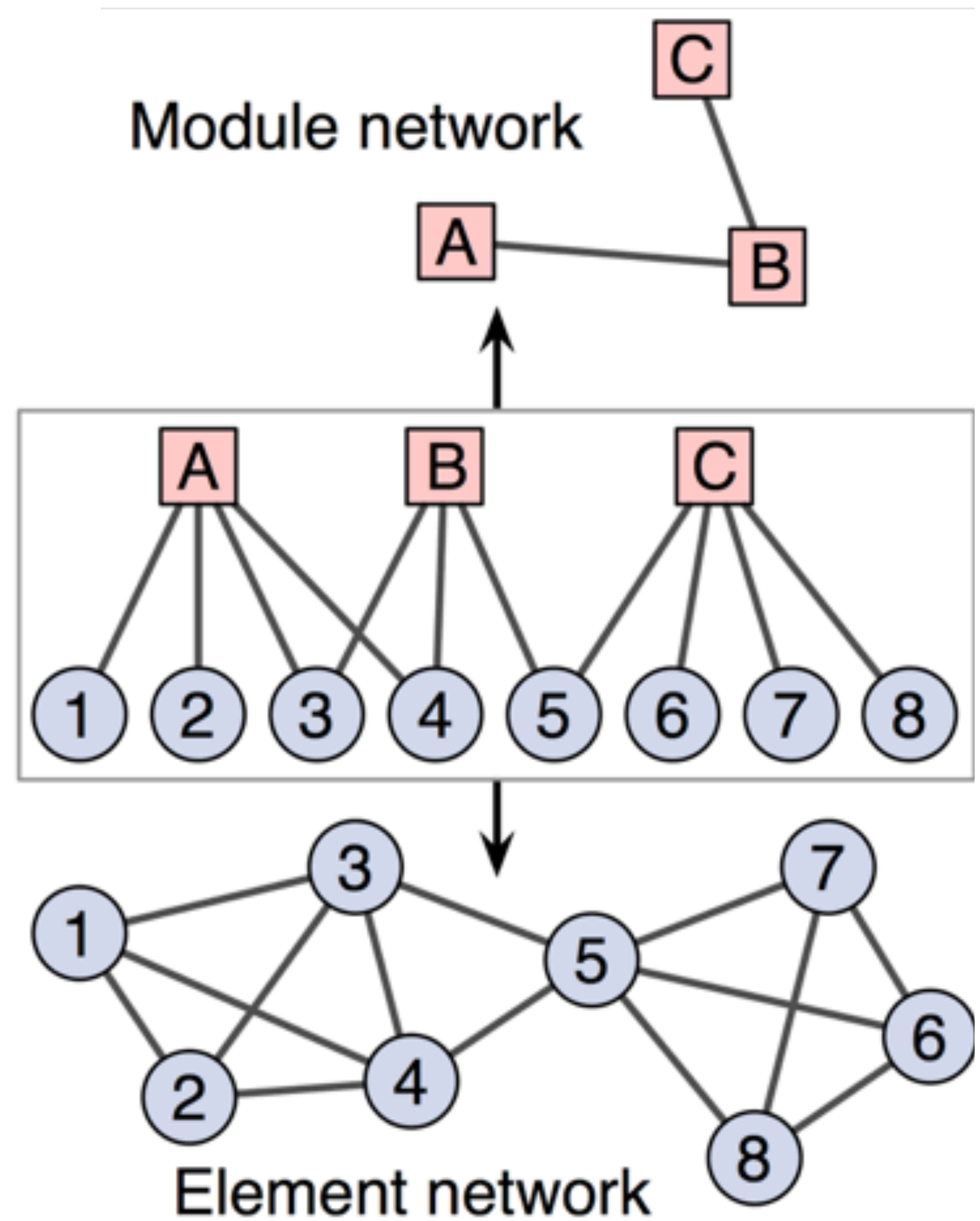
a simple model for pervasive overlap

pervasively overlapping network characterized by two **degree distributions** r_m and s_n

these determine the fraction of elements that belong to *m* modules and fraction of modules that contain *n* elements

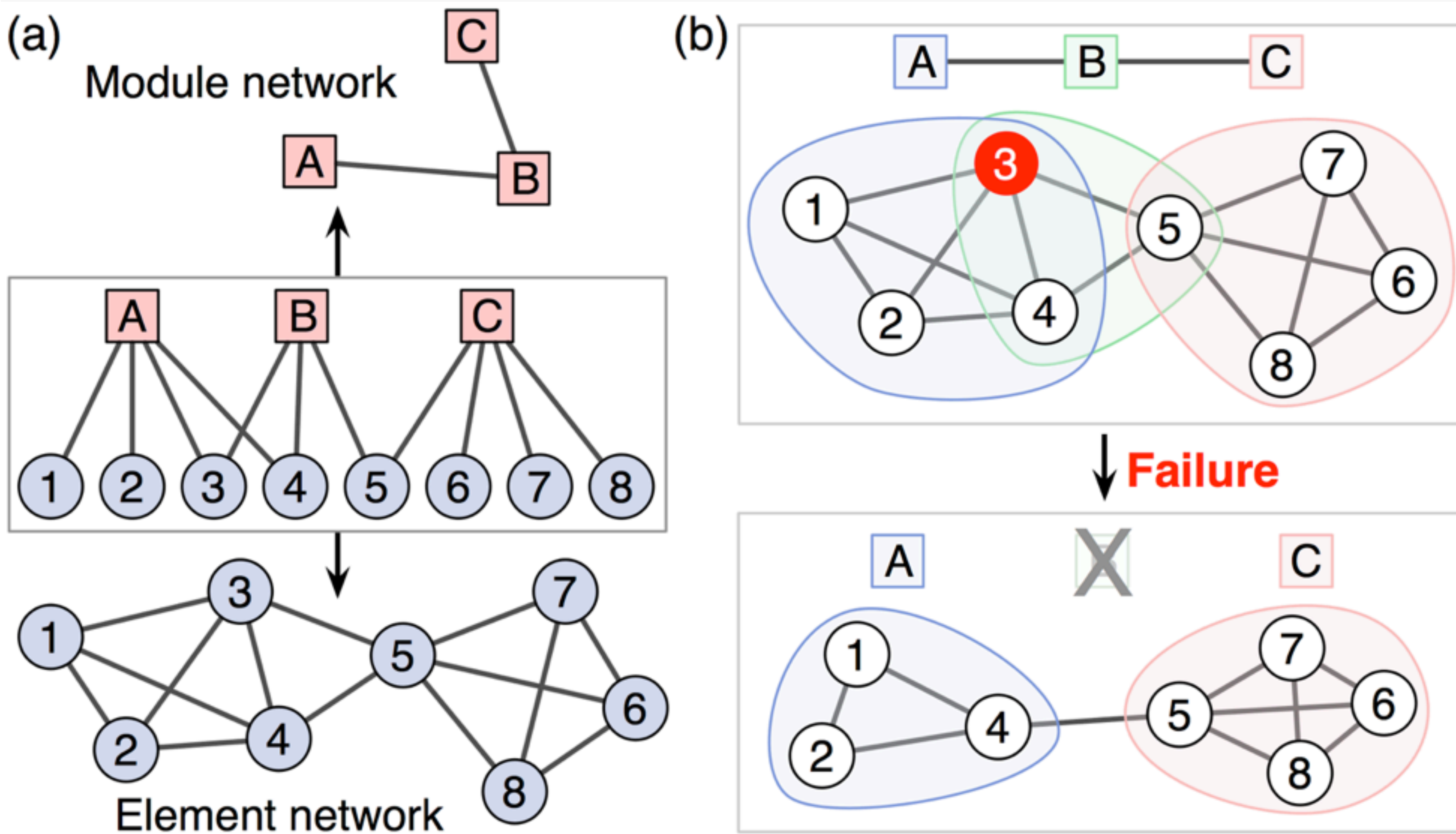
with averages

$$\mu \equiv \sum_m m r_m$$
$$\nu \equiv \sum_n n s_n.$$



projection provides module and element networks respectively

failure model

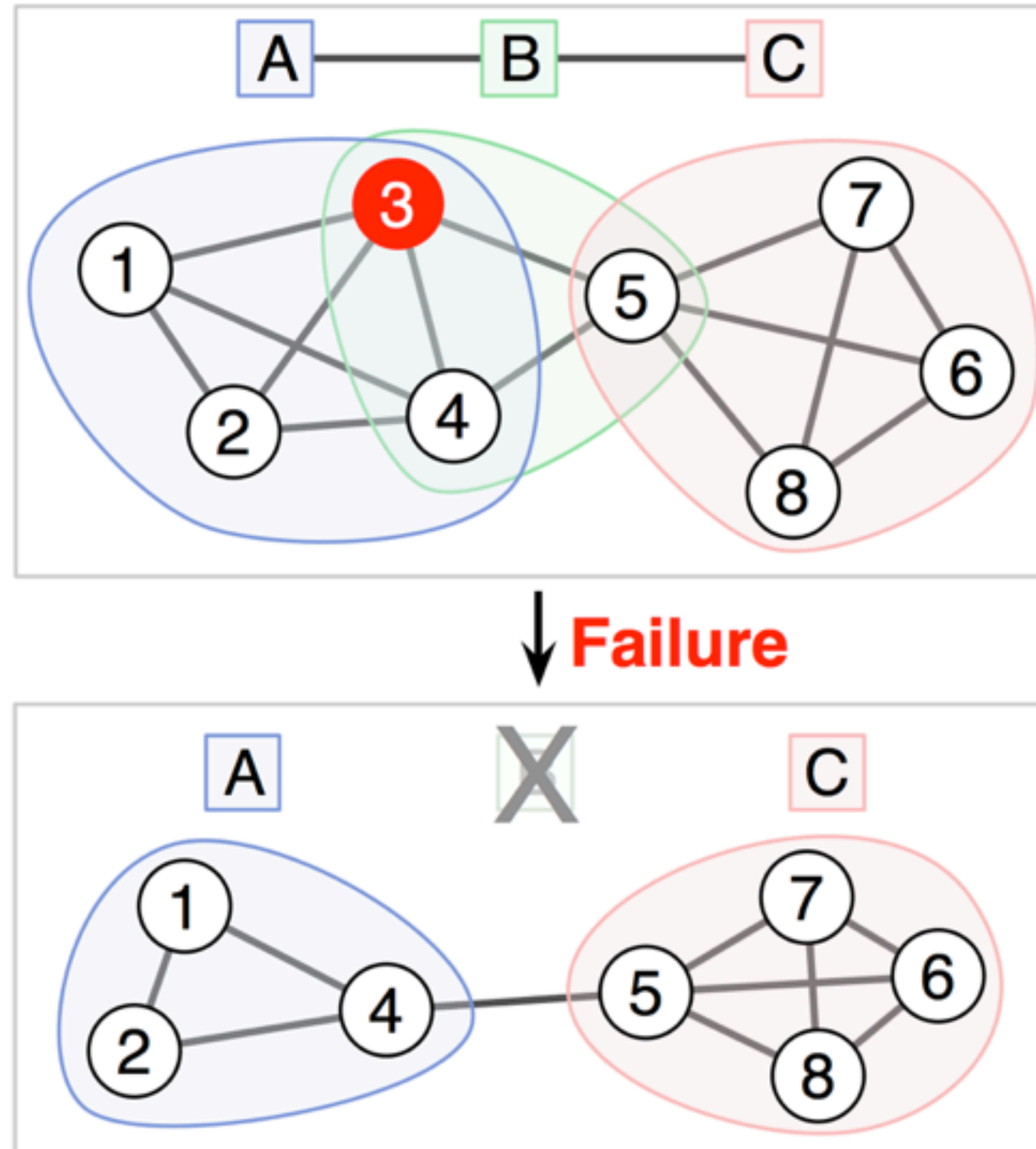


failure model

failures occurs on the element network. before projection, elements fail with probability (1 - *p*) and are removed from the network

we say that **modules fail** when fewer some critical f_c of the nodes in the module remain

failed modules are removed from the module network, but their elements remain in the element network

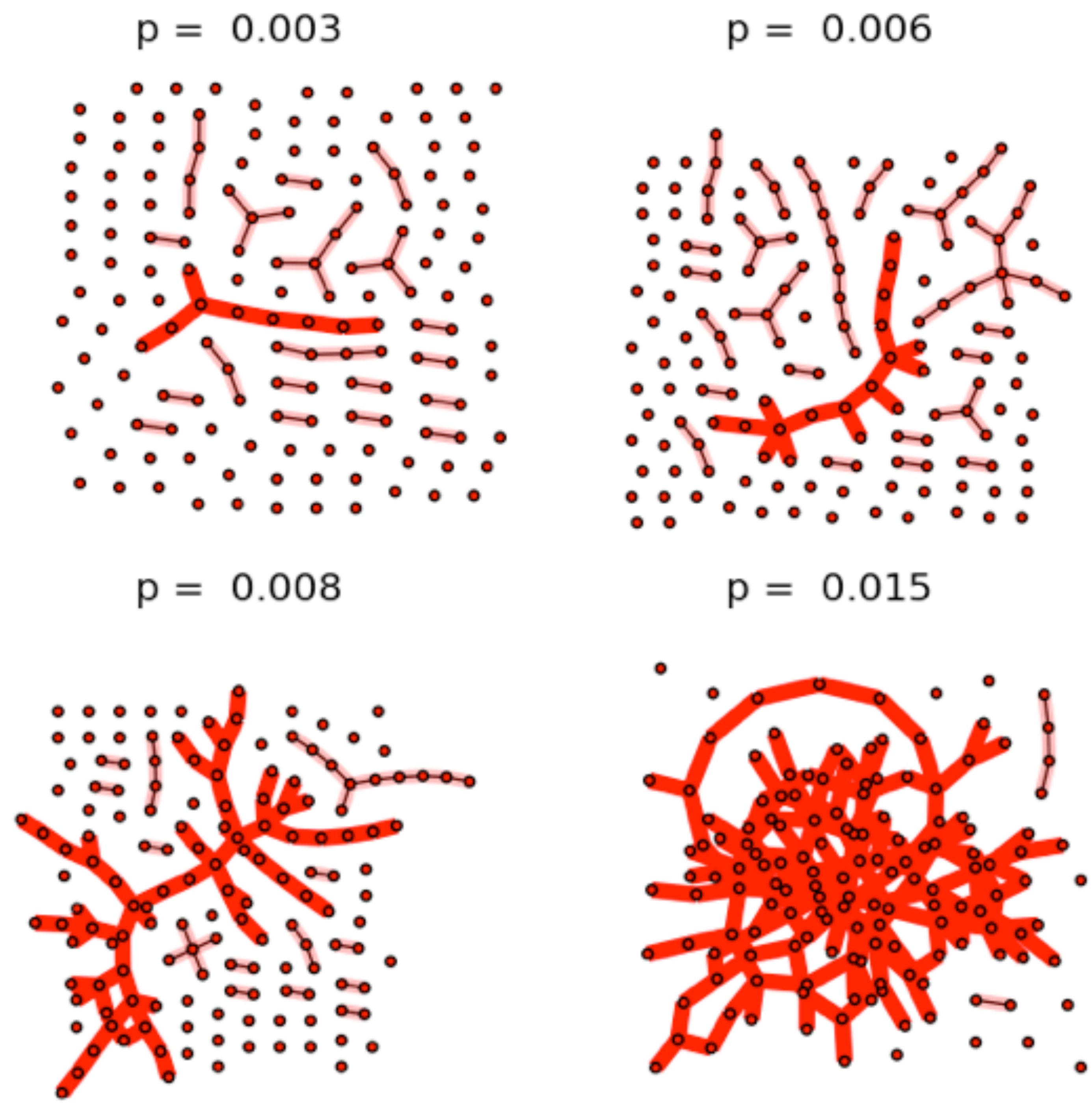


quantity of interest

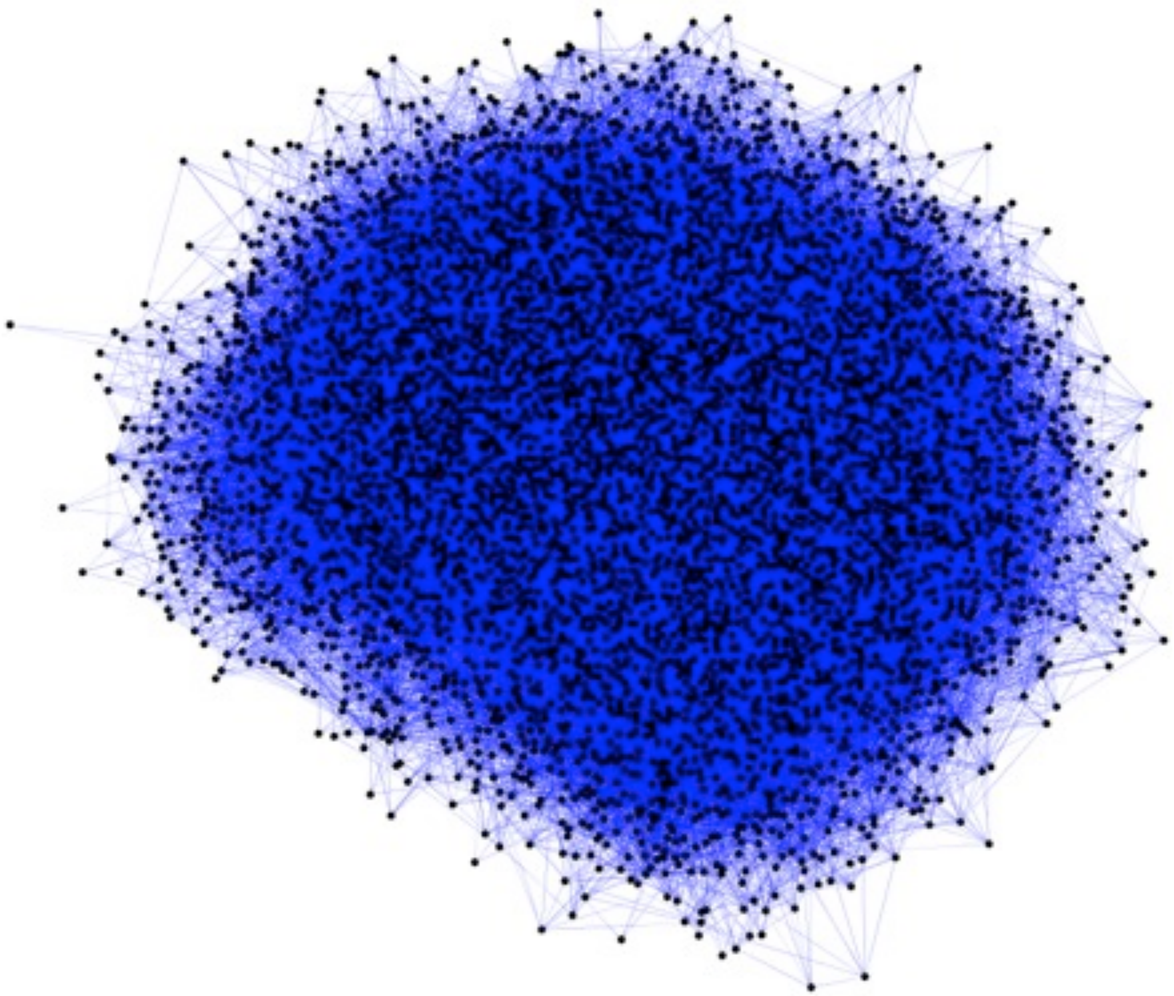
The giant component in the element network disappears when the network loses global connectivity.

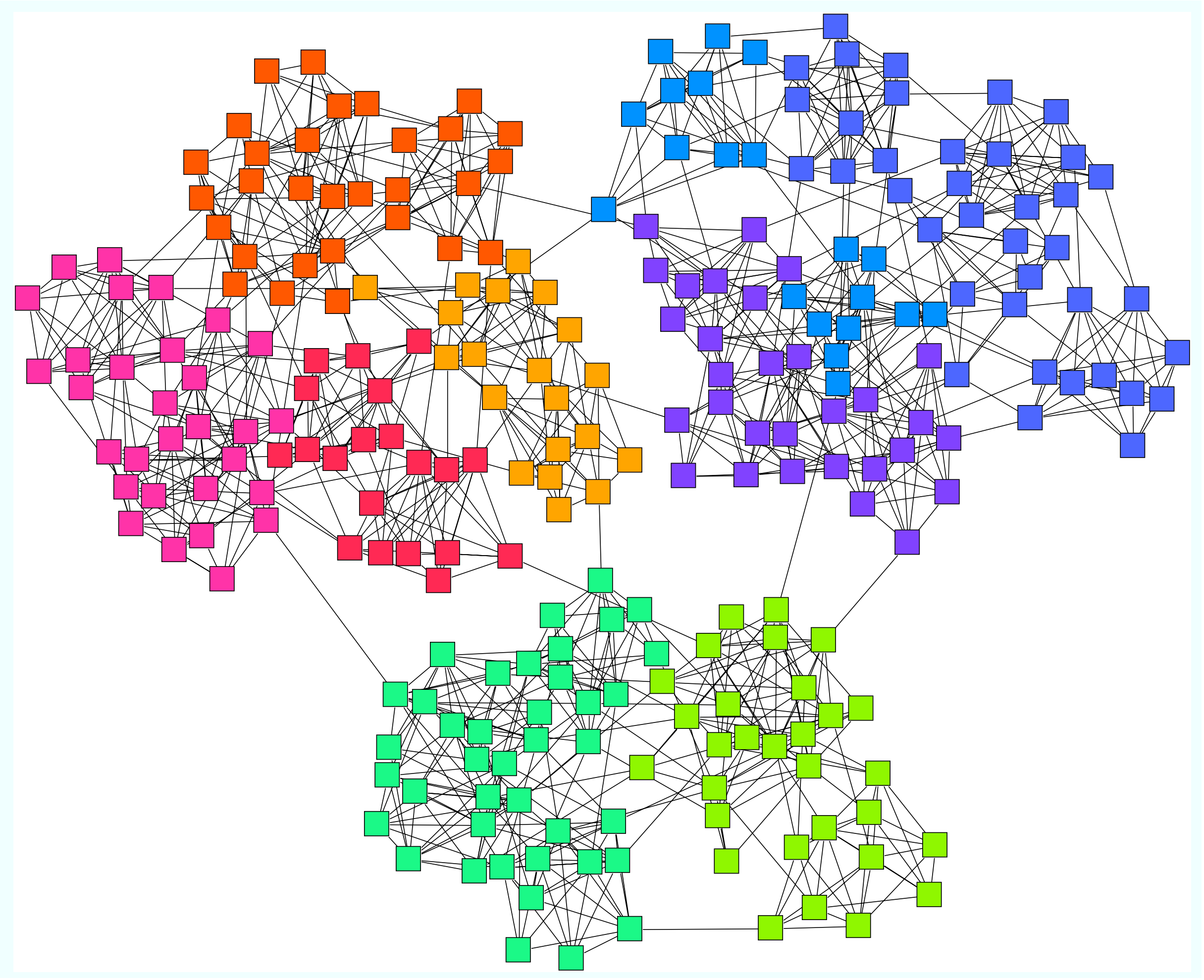
In the module network the giant connected component vanishes when the modules become uncoupled (non-overlapping)

We wish to determine *S*(*p*), the *fraction of remaining nodes within the giant component* as a function of *p*, for both the element and module networks



with these things in place, we can ***sharpen*** our question could we end up in a situation where the *element network remains globally connected*, but *module network has under gone a percolation transition*?





probability of a randomly chosen element to belong to m modules

probability that a random element within a randomly chosen module belongs to m other modules

$$f_0(z) = \sum_{m=0}^{\infty} r_m z^m, \quad f_1(z) = \frac{1}{\mu} \sum_{m=0}^{\infty} m r_m z^{m-1},$$

$$g_0(z) = \sum_{n=0}^{\infty} s_n z^n, \quad g_1(z) = \frac{1}{\nu} \sum_{n=0}^{\infty} n s_n z^{n-1}.$$

probability of a random module to contain n elements

prob that a random module of a randomly chosen element to contain n other elements

... a couple of pages of math

1 Element network

Consider a randomly chosen element A that belongs to a group of size n. Let $P(k|n)$ be the probability that A still belongs to a connected cluster of k nodes (including itself) in this group after failures occur:

$$P(k|n) = \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}. \tag{2}$$

The generating function for the number of other elements connected to A within this group is

$$h_n(z) = \sum_{k=1}^{n} P(k|n) z^{k-1} = (zp + 1 - p)^{n-1}. \tag{3}$$

Averaging over module size:

$$h(z) = \frac{1}{\nu} \sum_{n=0}^{\infty} n s_n h_n(z) = g_1(zp + 1 - p). \tag{4}$$

The total number of elements that A is connected to, from all modules it belongs to, is then generated by

$$G_0(z) = f_0(h(z)). \tag{5}$$

Likewise, the total number of elements that a randomly chosen neighbor of A is connected to is generated by

$$G_1(z) = f_1(h(z)). \tag{6}$$

Before determining S, we first identify the critical point p_c where the giant component emerges. This happens when the expected number of elements two steps away from a random element exceeds the number one step away, or

$$\partial_z G_0(G_1(z))\big|_{z=1} - \partial_z G_0(z)\big|_{z=1} > 0. \tag{7}$$

Substituting Eqs. (5) and (6) gives $f_0'(1)h'(1)[f_1'(1)h'(1) - 1] > 0$ or $f_1'(1)h'(1) > 1$. Finally, the condition for a giant component to exist, since $h'(1) = pg_1'(1)$, is

$$pf_1'(1)g_1'(1) > 1. \tag{8}$$

For the uniform case, $r_m = \delta(m, \mu)$ and $s_n = \delta(n, \nu)$, this gives $p(\mu - 1)(\nu - 1) > 1$. If $\mu = 3$ and $\nu = 3$, then the transition occurs at $p_c = 1/4$.

To find S, consider the probability u for element A to not belong to the giant component. A is not a member of the giant component only if all of A's neighbors are also not members, so u satisfies the self-consistency condition $u = G_1(u)$. The size of the giant component is then $S = 1 - G_0(u)$.

2 Module network

Consider a random module C and then a random member element A. Let $Q(\ell|m)$ be the probability that C is connected to ℓ modules, including itself, through element A, who was originally connected to m modules including C:

$$Q(\ell|m) = \binom{m-1}{\ell-1} q_1^{\ell-1} (1 - q_1)^{m-\ell}, \tag{9}$$

where

$$q_1 = \frac{1}{\nu} \sum_{n=0}^{\infty} n s_n \sum_{i=x}^{n} \binom{n-1}{i-1} p^{i-1}(1-p)^{n-i}. \tag{10}$$

(Notice that $q_1 = 1$ when $x(n) \equiv \lceil n f_c \rceil = 1$ for all n.) The generating function j_m for the number of modules that C is connected to, including itself, through A is

$$j_m(z) = \sum_{\ell=1}^{m} Q(\ell|m) z^{\ell-1} = (zq_1 + 1 - q_1)^{m-1}. \tag{11}$$

Once again, averaging j_m over memberships gives

$$j(z) = \frac{1}{\mu} \sum_{m=0}^{\infty} m r_m j_m(z) = f_1(zq_1 + 1 - q_1). \tag{12}$$

The total number of modules that C is connected to is *not* generated by $g_0(j(z))$ but by $\tilde{g}_0(j(z))$, where the $\tilde{g}_i$ are the generating functions for module size after elements fail:

$$\tilde{g}_0(z) = \sum_{n=0}^{\infty} \tilde{s}_n z^n, \qquad \tilde{g}_1(z) = \frac{\sum_{n=0}^{\infty} n\tilde{s}_n z^{n-1}}{\sum_{n=0}^{\infty} n\tilde{s}_n}. \tag{13}$$

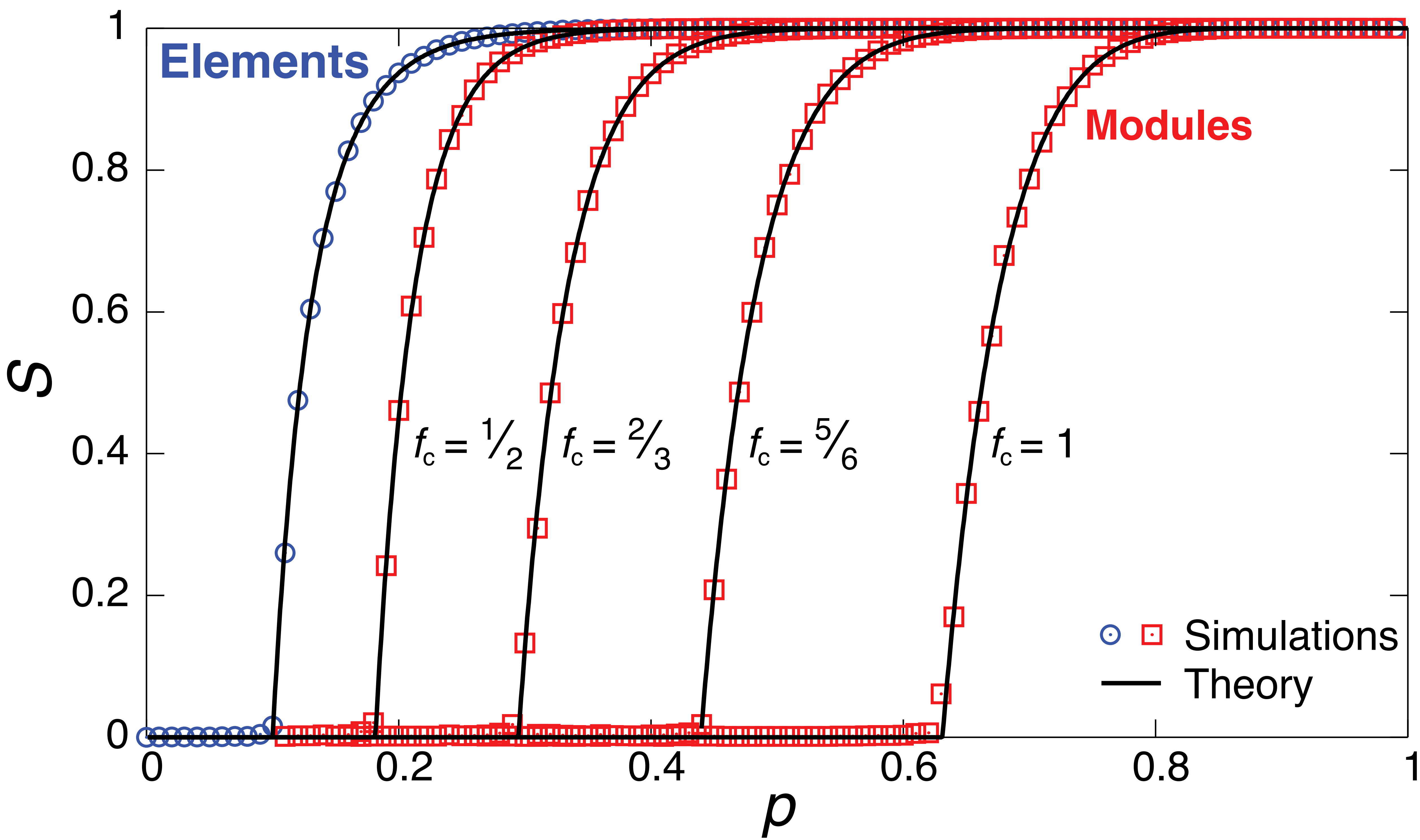
The probability $\tilde{s}_k$ to have k member elements remaining in a module after percolation is given by

$$\tilde{s}_k = \frac{\sum_n \binom{n}{k} p^k (1-p)^{n-k} s_n}{\sum_n \sum_{k'=x}^{n} \binom{n}{k'} p^{k'} (1-p)^{n-k'} s_n} \tag{14}$$

The denominator is necessary for normalization since we cannot observe modules with fewer than $\lceil n f_c \rceil$ members. Notice that $\tilde{s}_n = s_n$ when $s_n = \delta(n, \nu)$ and $\lceil n f_c \rceil = n = \nu$.

Finally, the total number of modules connected to C through any member elements is generated by $F_0(z) = \tilde{g}_0(j(z))$ and the total number of modules connected to a random neighbor of C is generated by $F_1(z) = \tilde{g}_1(j(z))$. As before, the module network has a giant component when $\partial_z F_0(F_1(z))|_{z=1} - \partial_z F_0(z)|_{z=1} > 0$ and $S = 1 - F_0(u) = 1 - \tilde{g}_0(j(u))$, where u satisfies $u = F_1(u) = \tilde{g}_1(j(u))$.

For the uniform case with $\mu = 3$, $\nu = 3$, and $f_c > 2/3$, the critical point for the module network is $p_c = 1/2$, a considerably higher threshold than for the element network ($p_c = 1/4$). In Fig. 2 we show S for $\mu = 3$ and $\nu = 6$. The "robustness gap" between the element and module networks widens as the module failure cutoff increases, covering a significant range of p for the larger values of f_c.



$$r_m = \delta(m, \mu) \qquad \nu = 6$$
$$s_n = \delta(n, \nu) \qquad \mu = 3$$

scale free networks

It is known that scale-free networks are robust to random failures when $2 < \lambda < 3$ (meaning that $p_c \rightarrow 0$).

(This result requires max value of distribution, *K,* to be large*.)*

Error and attack tolerance of complex networks

Réka Albert, Hawoong Jeong & Albert-László Barabási

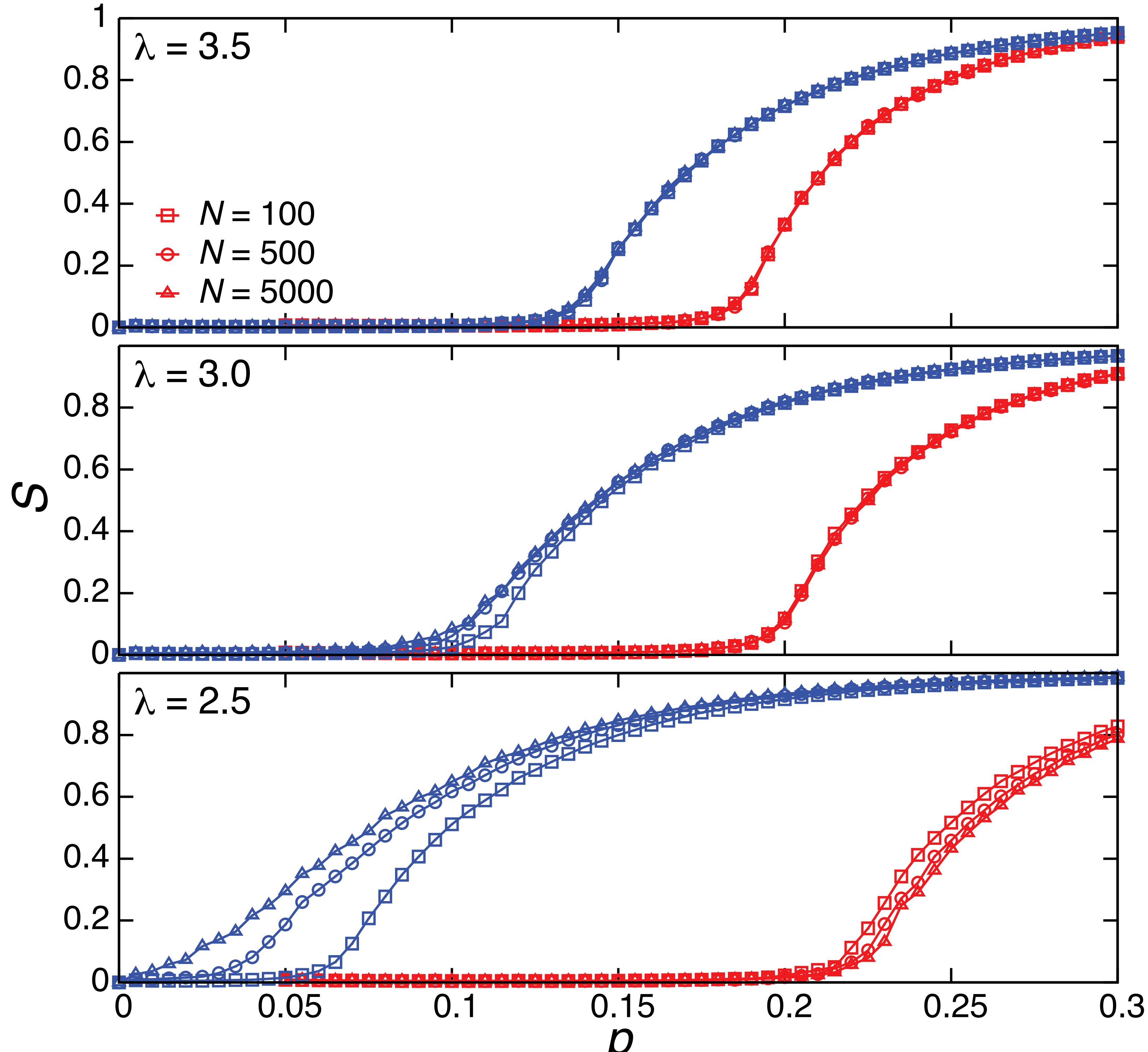
Department of Physics, 225 Nieuwland Science Hall, University of Notre Dame, Notre Dame, Indiana 46556, USA

Many complex systems display a surprising degree of tolerance against errors. For example, relatively simple organisms grow, persist and reproduce despite drastic pharmaceutical or environmental interventions, an error tolerance attributed to the robustness of the underlying metabolic network[1]. Complex communication networks[2] display a surprising degree of robustness: although key components regularly malfunction, local failures rarely lead to the loss of the global information-carrying ability of the network. The stability of these and other complex systems is often attributed to the redundant wiring of the functional web defined by the systems' components. Here we demonstrate that error tolerance is not shared by all redundant systems: it is displayed only by a class of inhomogeneously wired networks,

scale free networks

Here we take $r_m = \delta(m, \mu)$ as before, but now $s_n \sim n^{-\lambda}$, with $\lambda \geq 2$

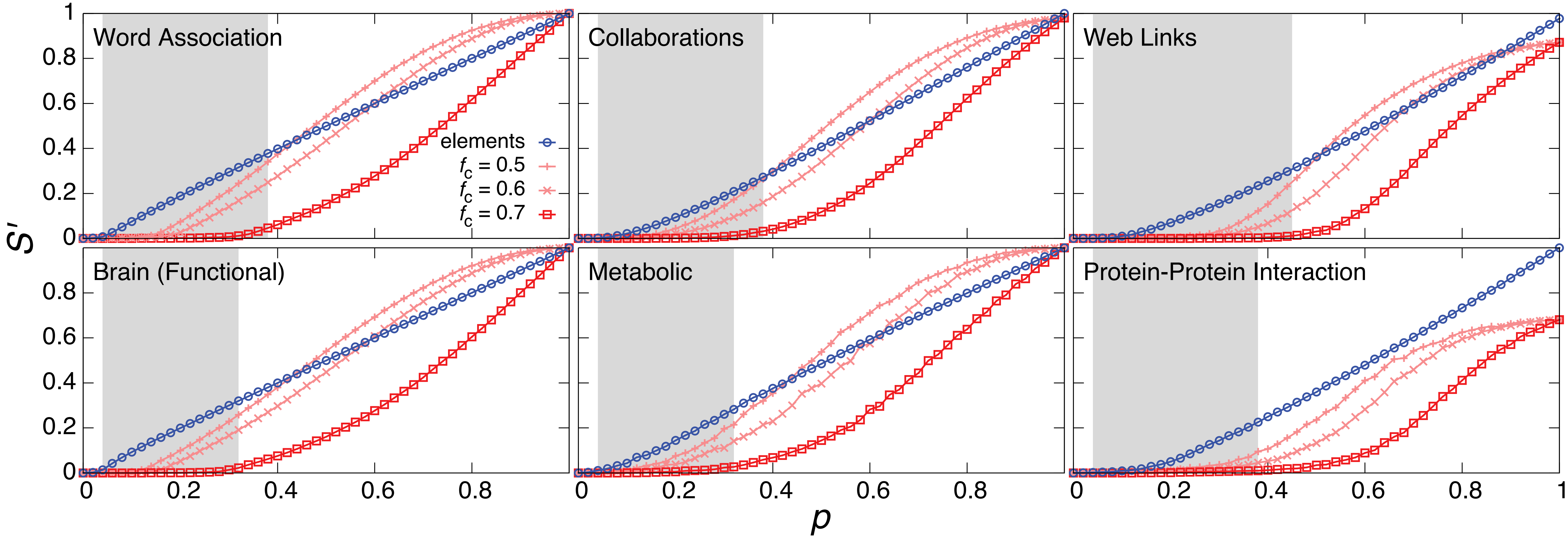
As we lower λ (increasing *K*), the elements become more robust (as expected), but the module network becomes *less* robust.



For modular networks, it may not be feasible to build extremely large modules. Interestingly, enforcing on s_n a maximum module size cutoff $N = \max\{n \mid s_n > 0\}$ only improves element robustness.

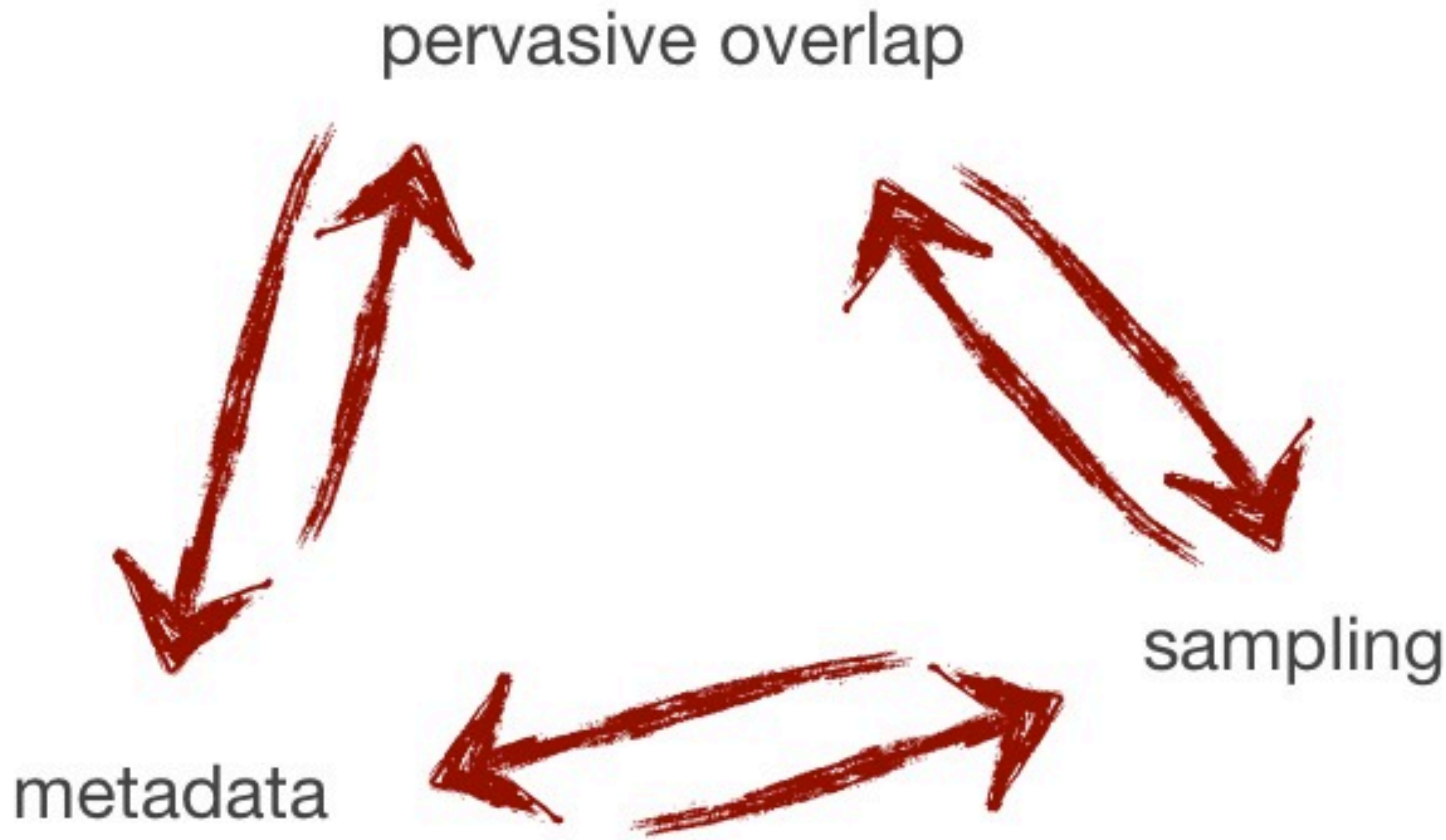
Figure 3: Robustness of scale-free networks. Here $r_m = \delta(m, 3)$, $s_n \sim n^{-\lambda}$, $f_c = 1/2$, and $N \equiv \max\{n \mid s_n > 0\}$. Increasing N and decreasing λ, measures known to improve

real world networks



$S'(p)$ the fraction of *original* nodes in the giant connected component

Shaded regions provide a guide to the eye for the robustness gap ($f_c = 0.7$).


pervasive overlap
sampling
metadata