NETWORK COMMUNITY ANALYSIS

Yong-Yeol “YY” Ahn

SCHOOL OF INFORMATICS AND COMPUTING

INDIANA UNIVERSITY

Bloomington

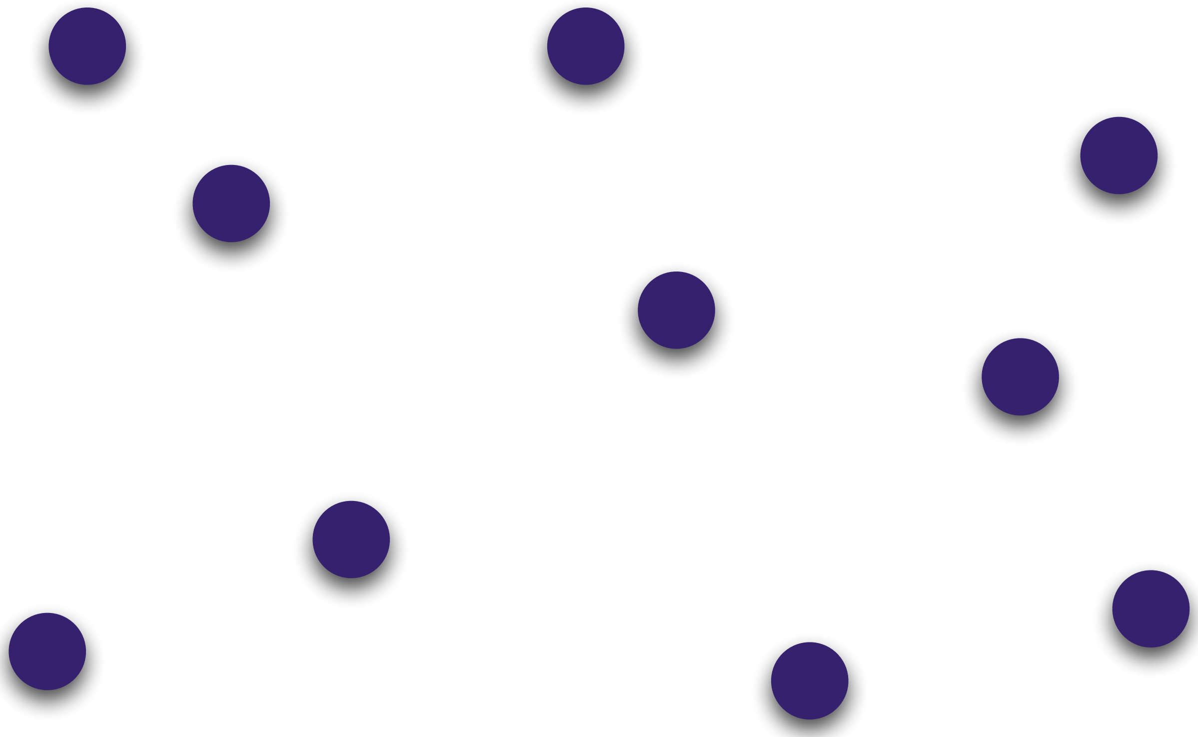
Disclaimer:

We're just scratching the surface.
There are so many cool studies that
I cannot cover here!

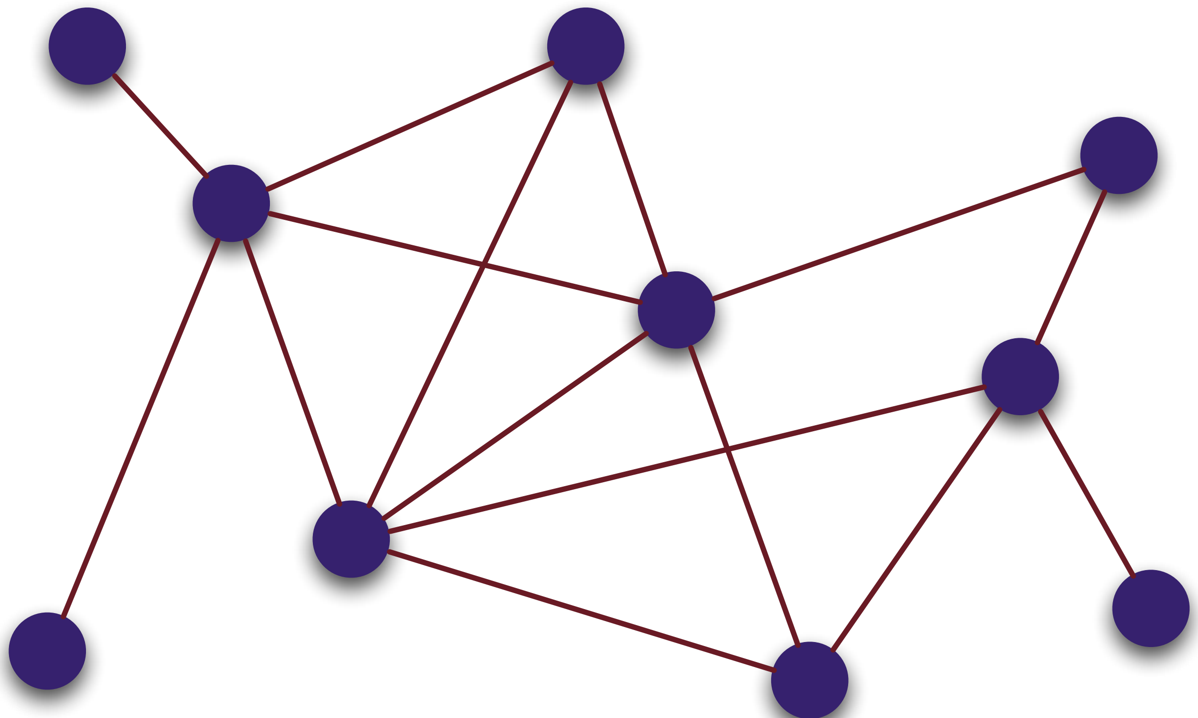
Visit:

http://goo.gl/opEfTp

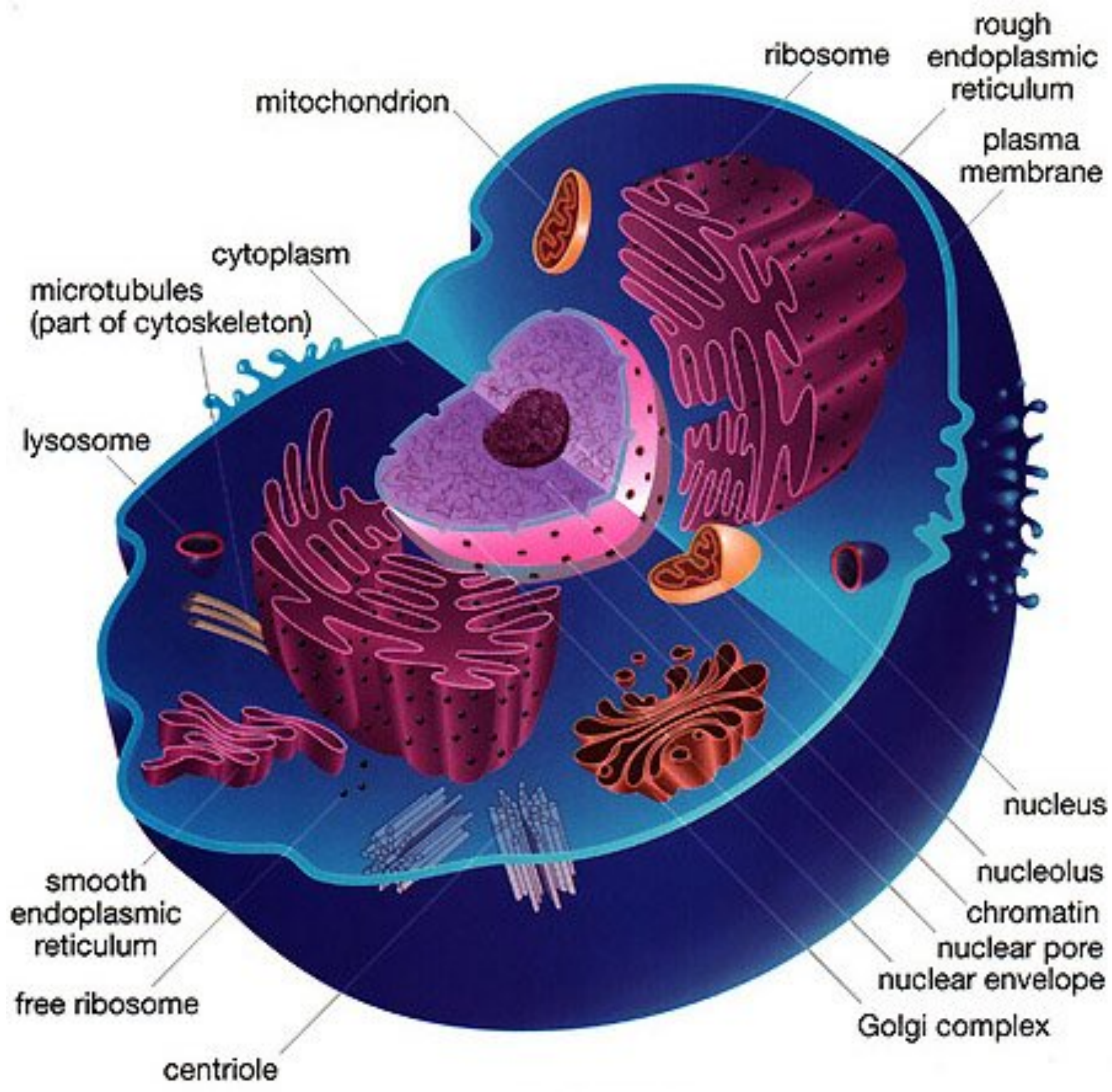
Network community analysis



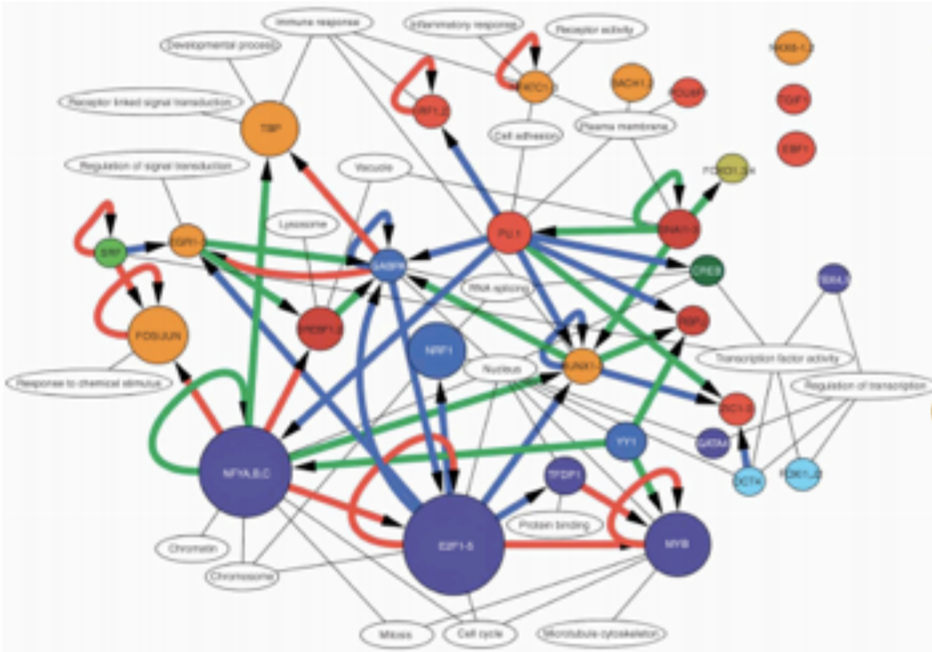
Nodes

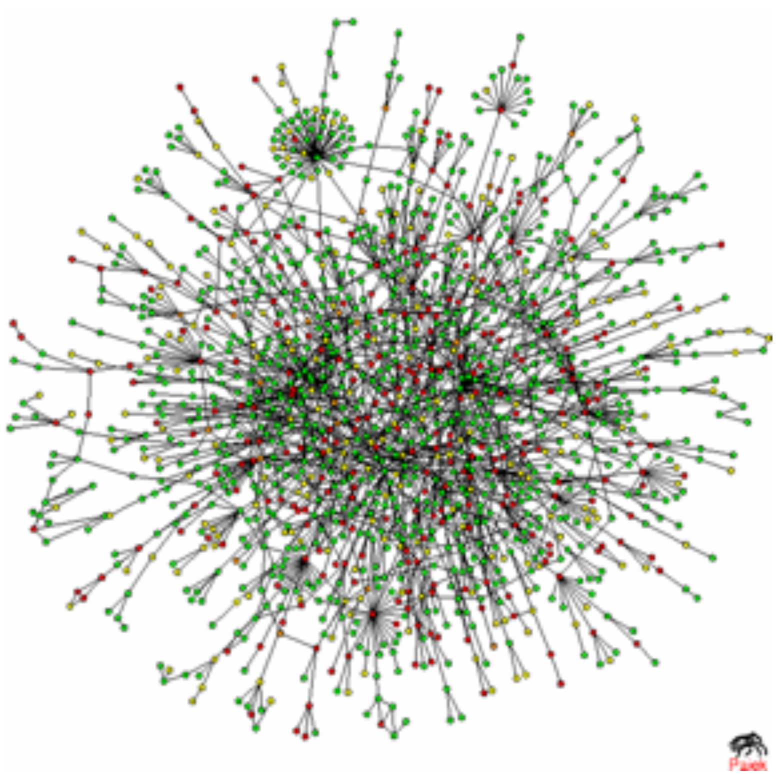


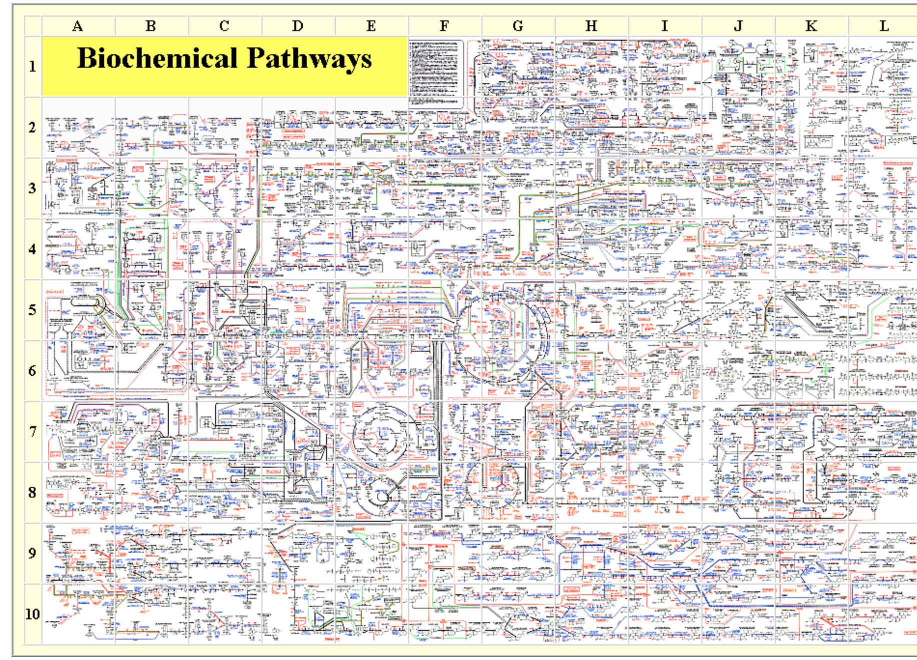
Links (edges) between nodes

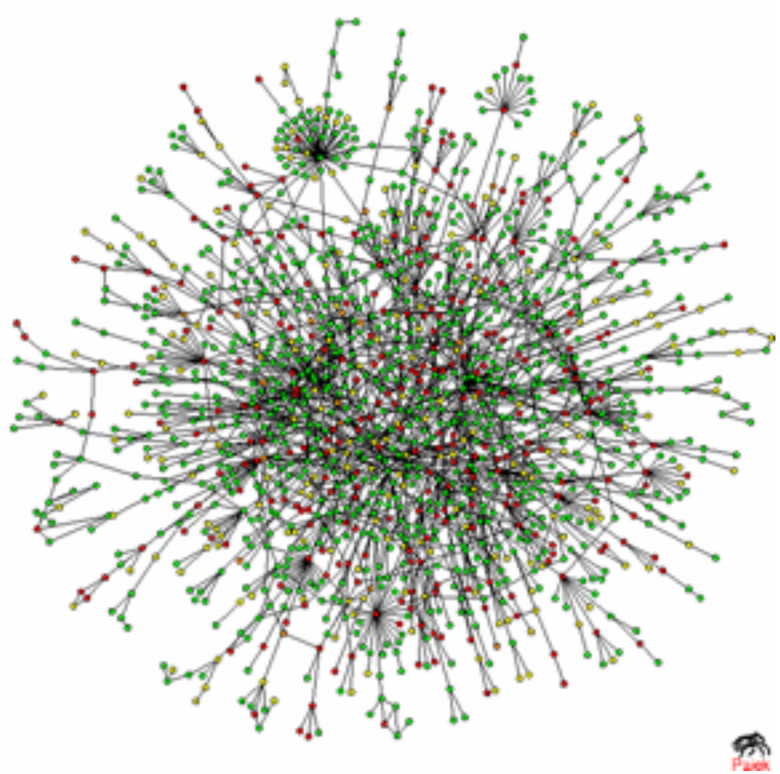

ribosome
rough endoplasmic reticulum
mitochondrion
plasma membrane
cytoplasm
microtubules (part of cytoskeleton)
lysosome
nucleus
nucleolus
chromatin
nuclear pore
nuclear envelope
smooth endoplasmic reticulum
free ribosome
Golgi complex
centriole


Immune response
Developmental process
Inflammatory response
Receptor activity
Receptor linked signal transduction
TBP
Cell adhesion
Plasma membrane
Regulation of signal transduction
Vacuole
Lysosome
PU.1
CREB
RNA splicing
NRF1
Nucleus
FOS/JUN
Response to chemical stimulus
Transcription factor activity
Regulation of transcription
YY1
NFYA,B,C
TFDP1
E2F1-5
MYB
Protein binding
Chromatin
Chromosome
Mitosis
Cell cycle
Microtubule cytoskeleton

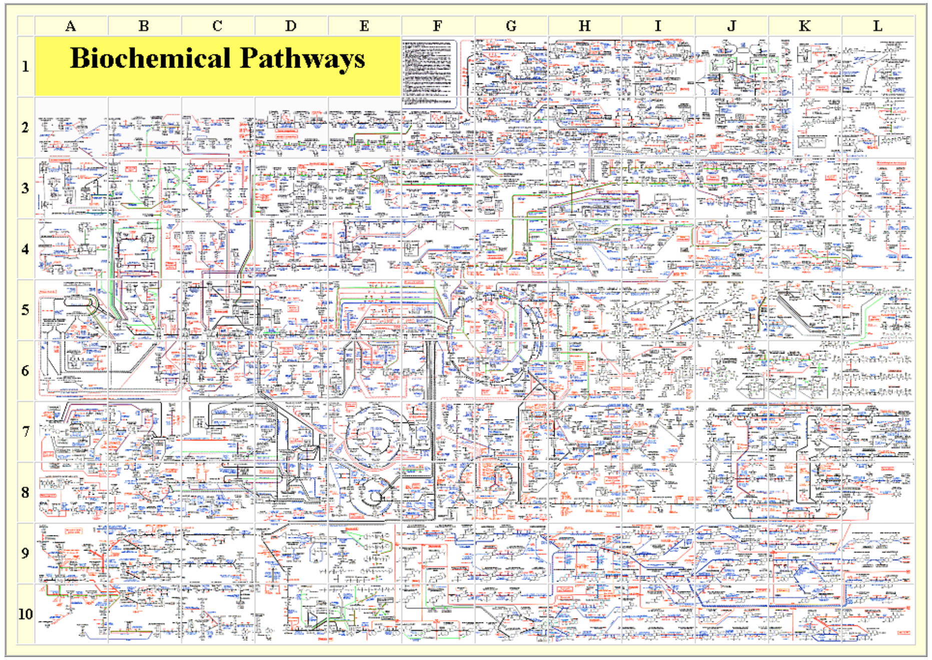

Immune response
Developmental process
Inflammatory response
Receptor activity
Receptor linked signal transduction
TBP
Cell adhesion
Plasma membrane
Regulation of signal transduction
Vacuole
Lysosome
PU.1
RNA splicing
FOS/JUN
NRF1
Nucleus
Transcription factor activity
Regulation of transcription
Response to chemical stimulus
YY1
NFYA,B,C
E2F1-5
MYB
Protein binding
Chromatin
Chromosome
Mitosis
Cell cycle
Microtubule cytoskeleton

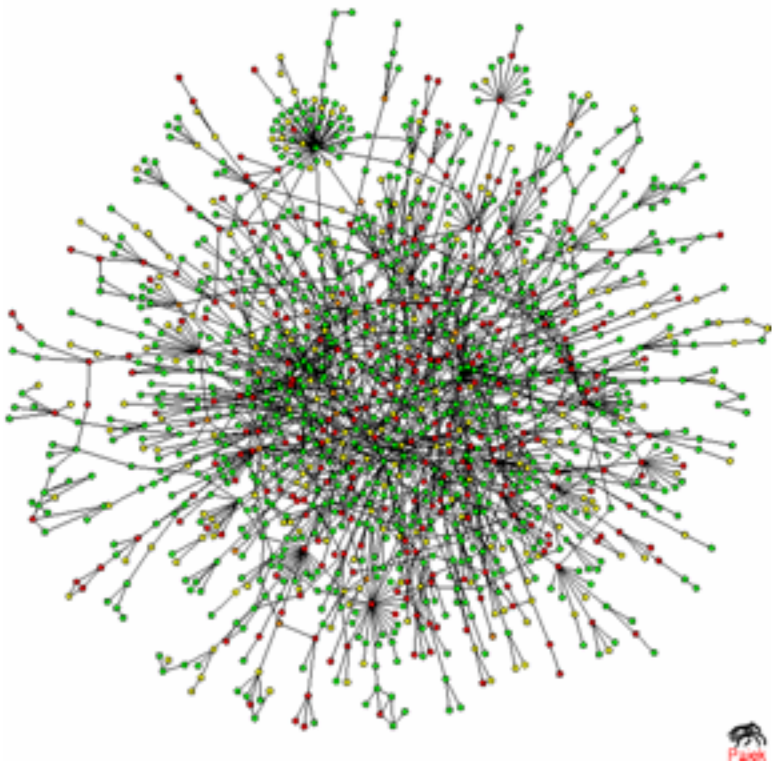

Pajek

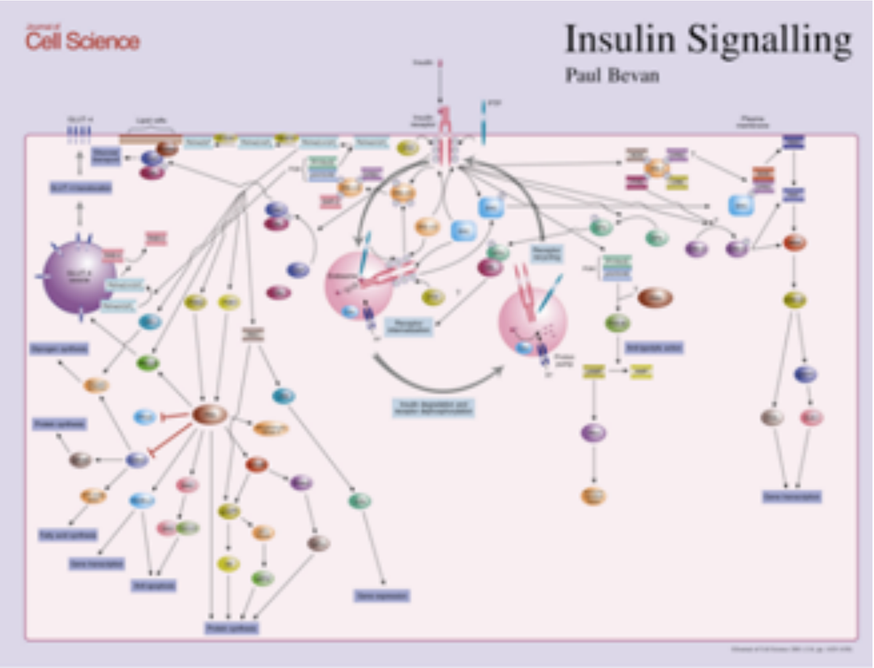

Biochemical Pathways

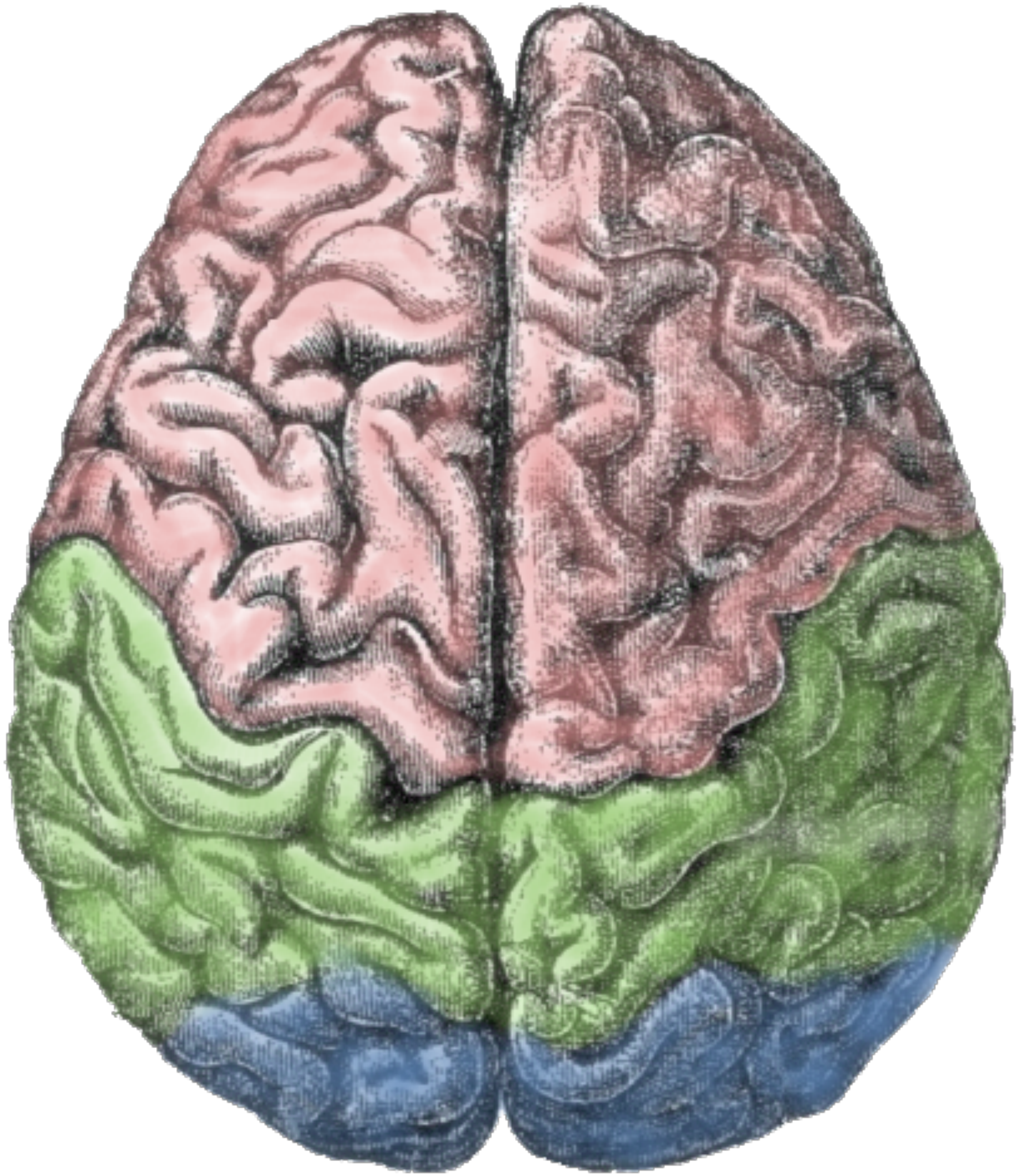



Immune response
Developmental process
Inflammatory response
Receptor activity
Receptor linked signal transduction
Regulation of signal transduction
Cell adhesion
Plasma membrane
Vacuole
Lysosome
RNA splicing
Nucleus
Transcription factor activity
Regulation of transcription
Response to chemical stimulus
Protein binding
Chromatin
Chromosome
Mitosis
Cell cycle
Microtubule cytoskeleton
TBP
PU.1
NRF1
YY1
TFDP1
FOS/JUN
NFYA,B,C
E2F1-5
MYB


Biochemical Pathways
A
B
C
D
E
F
G
H
I
J
K
L
1
2
3
4
5
6
7
8
9
10


Pajek


Cell Science
Insulin Signalling
Paul Bevan




facebook
December 2010

Networks = The maps of complex systems

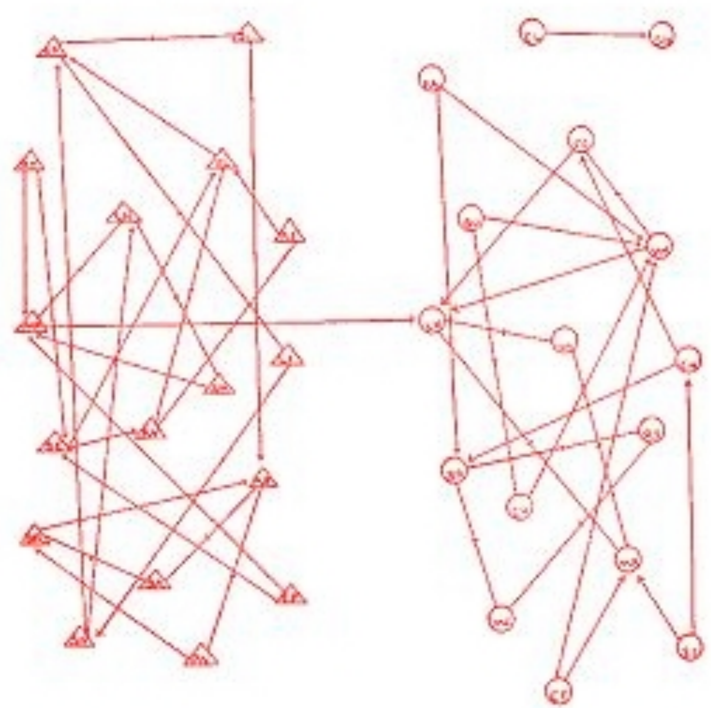
Network **community** analysis

What is a network community?

EMOTIONS MAPPED
BY NEW GEOGRAPHY

Charts Seek to Portray the
Psychological Currents of
Human Relationships.

New York Times
April 3, 1933

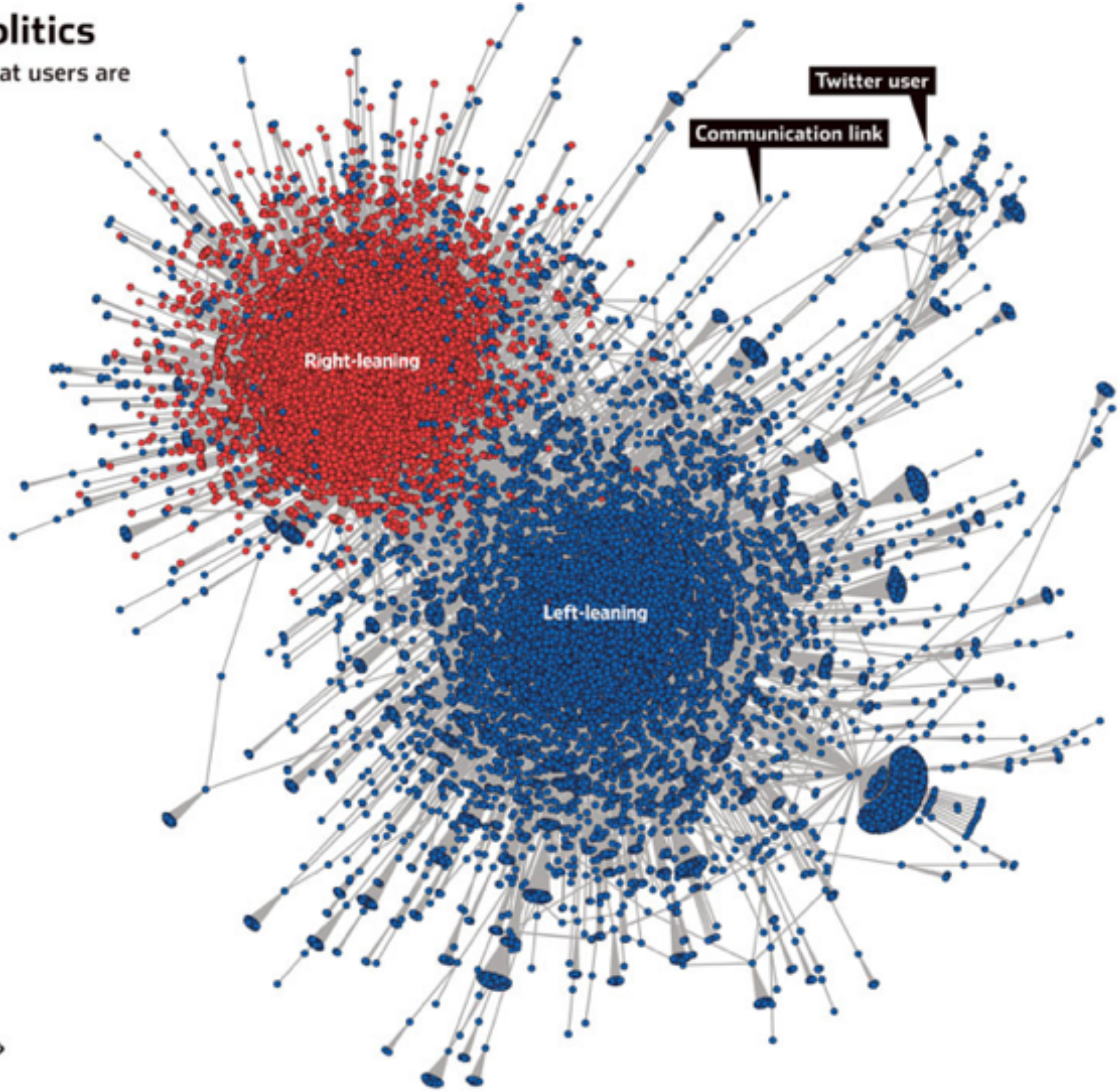


Moreno’s “sociogram”

Twitter's Divided Politics

Political Twitter traffic reveals that users are polarized along party lines.*

Researchers at Indiana University analyzed 250,000 Twitter messages on political topics exchanged by 45,000 people during the 2010 mid-term congressional elections. This chart of 'retweets'—in which one user forwards another's message—shows that, though there were more left-leaning users, right-leaning users were more densely connected to one another. (Each dot is a Twitter user, and the lines show retweets.) Even so, as the chart illustrates, lines of communication do sometimes reach across the political divide.

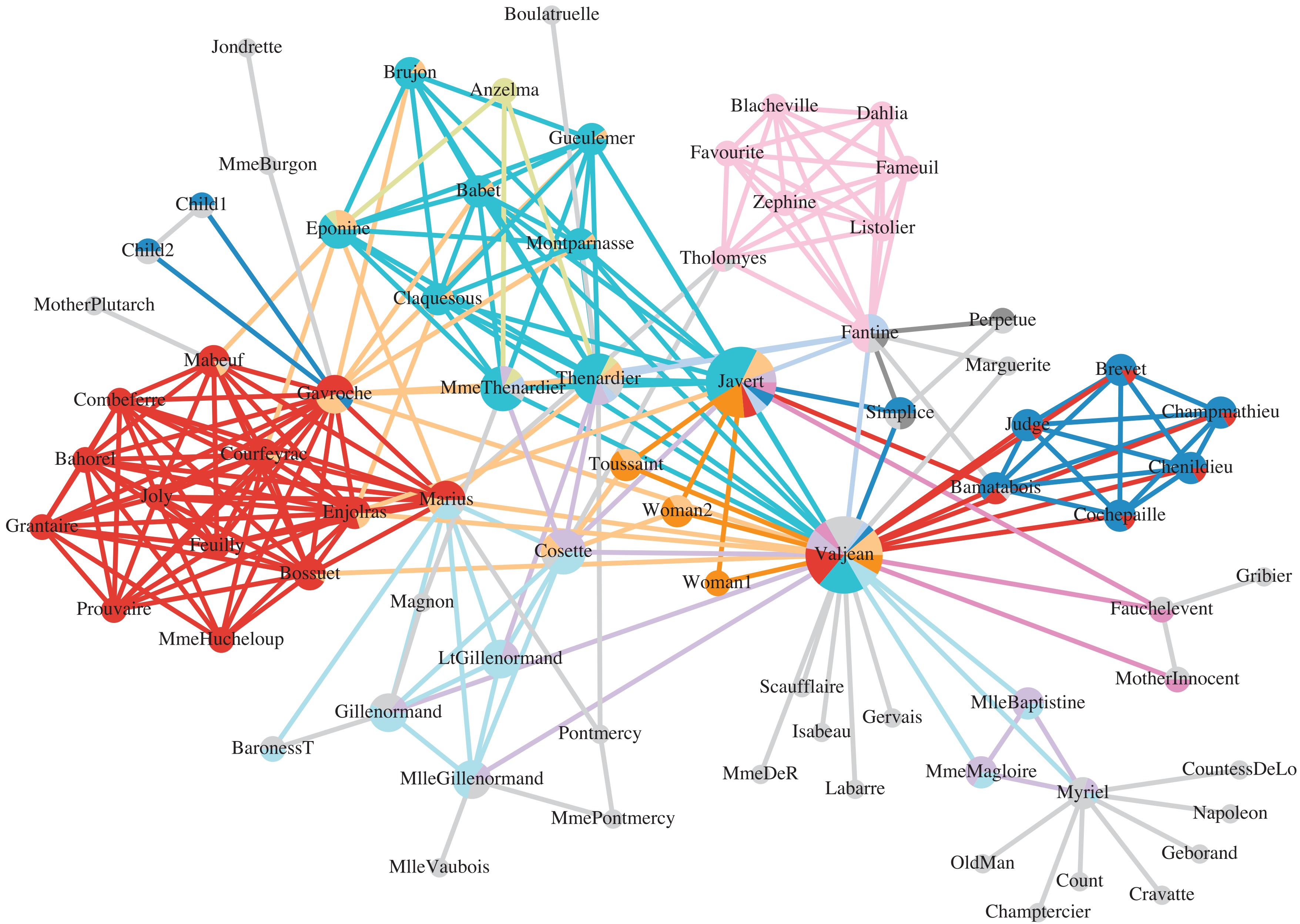


The most popular hashtags (short codes signaling the message's content), shown by number of tweets. Researchers found that users on the left and right use each other's hashtags.

Hashtags: tcot, top conservatives on Twitter; p2, progressives 2.0; sgp, smart girl politics; tlot, top libertarians on Twitter.

*Data show 'retweets' of other users' messages. Political leaning designations are based on algorithmically-determined communities of users which correlate with political affiliation.

Source: Center for Complex Networks and Systems Research, Indiana University



Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* (2010)

Cohesiveness

Separation

Group **cohesiveness**

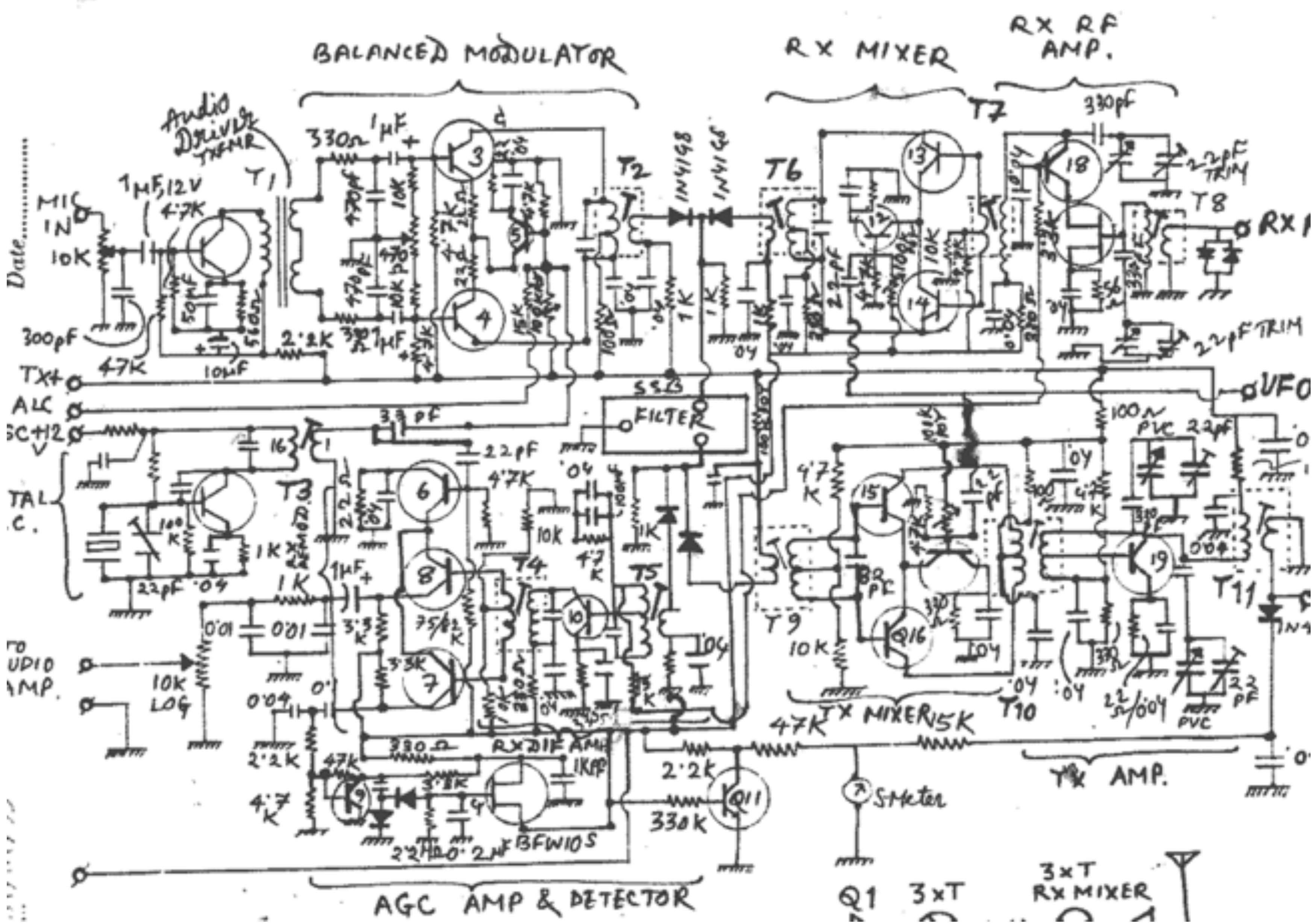
(Moreno & Jennigs 1938, Festinger 1950, Gross & Martin 1952)

Graph **partitioning**

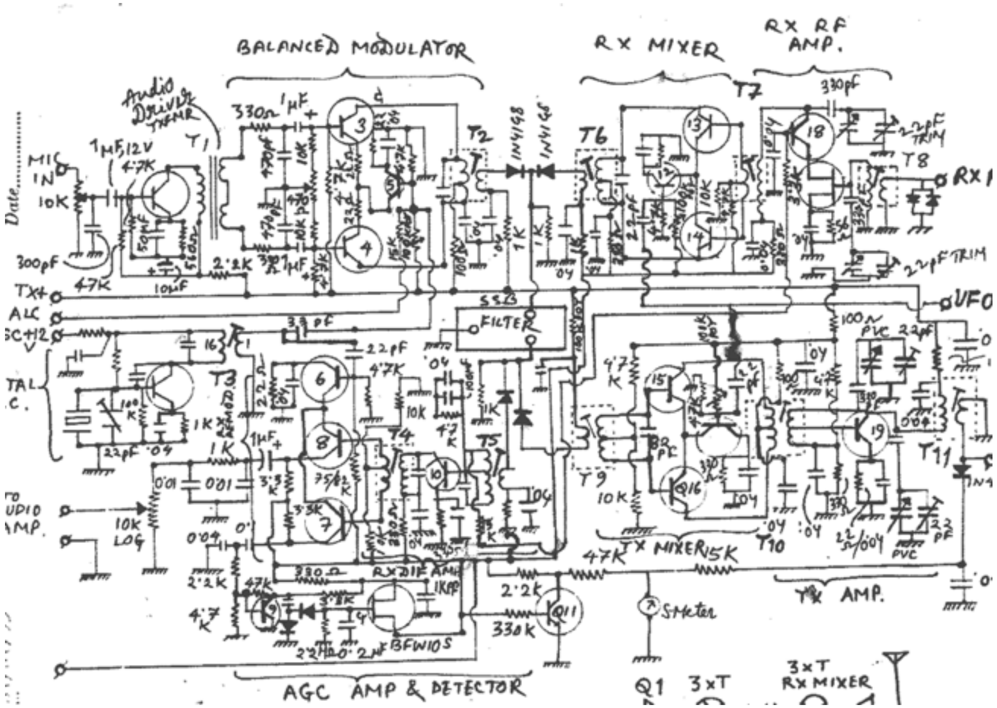
(Kernighan & Lin 1970)

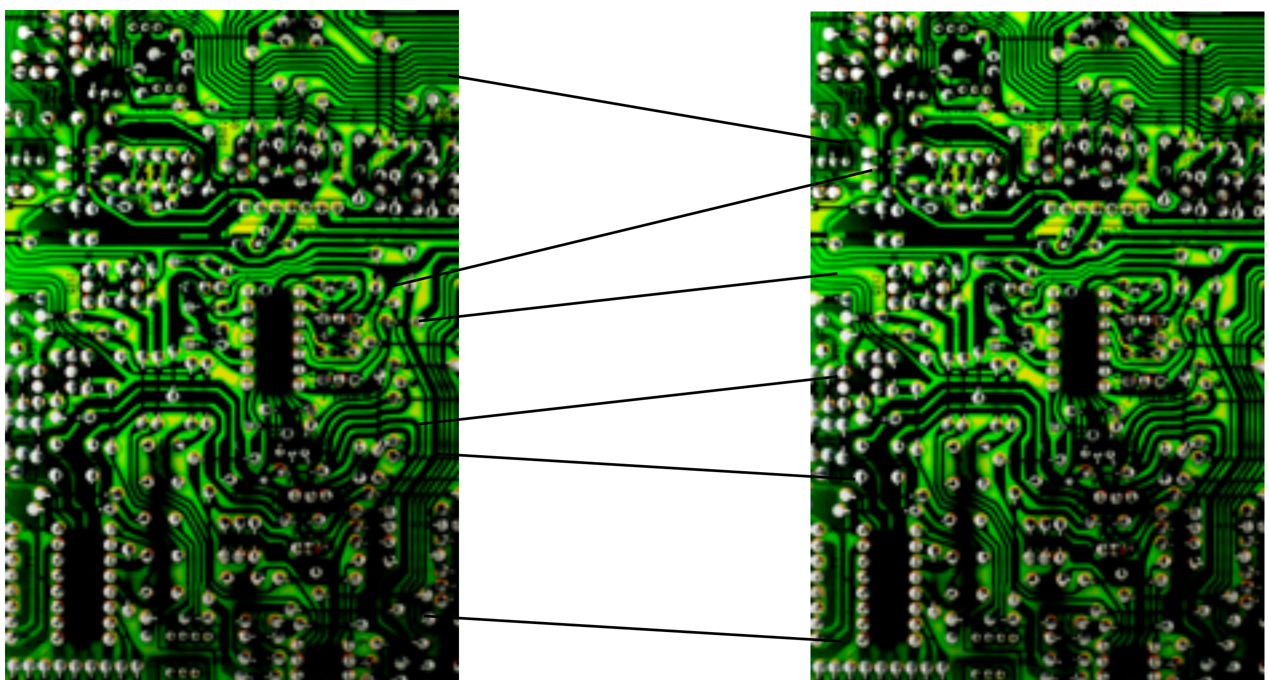
Why do we care?

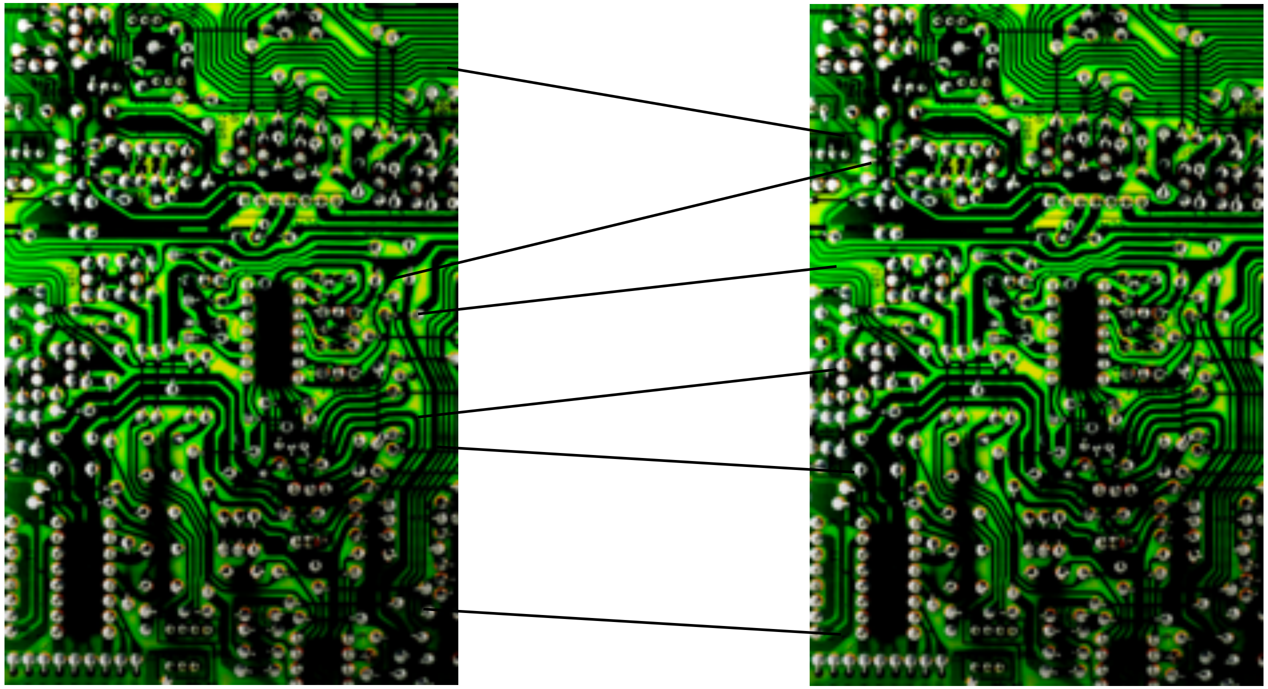
Original motivation:
Computation



BALANCED MODULATOR
RX MIXER
RX RF AMP.
Audio Driver TXFMR
MIC IN
10K
1MF,12V
4.7K
300pF
47K
10MF
2.2K
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
1N4148
SSB FILTER
TX+
ALC
VFO
330pF
22pF TRIM
10K LOG
AGC AMP & DETECTOR
RX DIF AMP
BFW10S
TX MIXER
S-Meter
330K
Q1
3xT
3xT RX MIXER
Date......


BALANCED MODULATOR
RX MIXER
RX RF AMP.
Audio Driver TXFMR
MIC IN
TX+
ALC
FILTER
SSB
RX IF AMP
TX MIXER
TX AMP.
AGC AMP & DETECTOR
S-meter
BFW10S
VFO
Q1
3xT
3xT RX MIXER





How to minimize the number of wires?

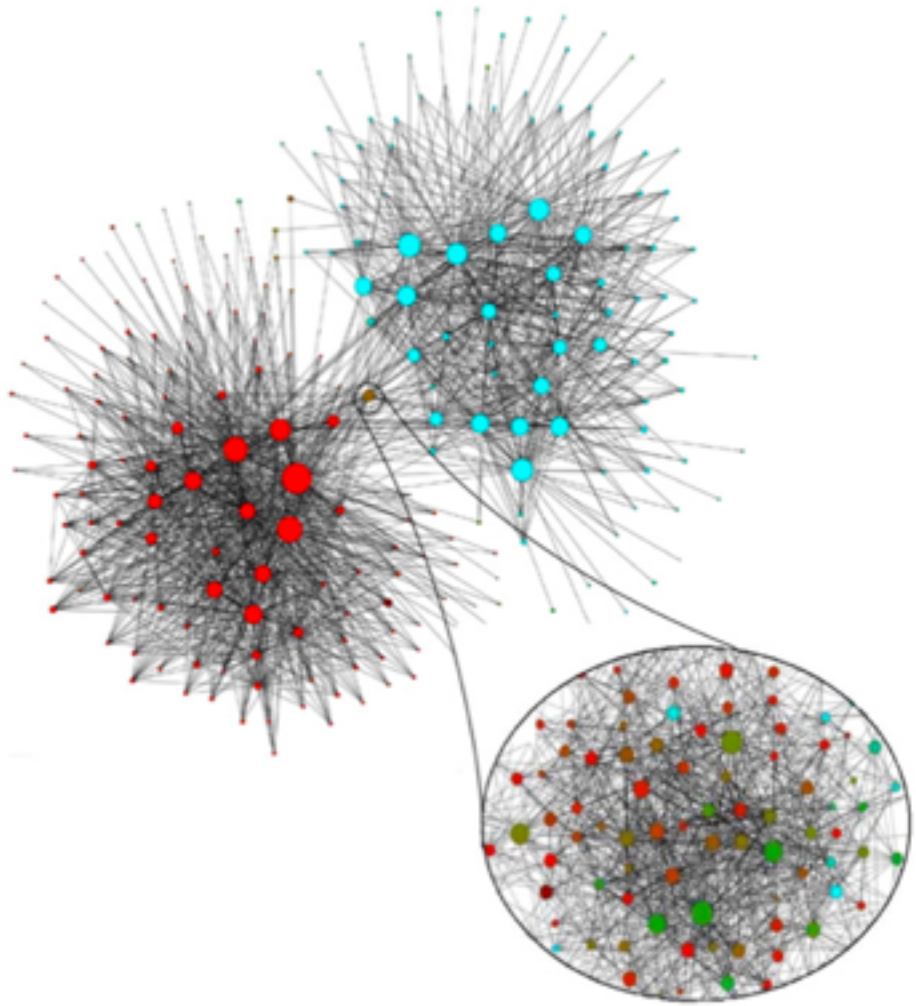
How to minimize the communication between computers?

Circuits, Communication between softwares ~ **Networks**

Functional modules ~ **Communities**

Correspondence to **functional**, **structural units**

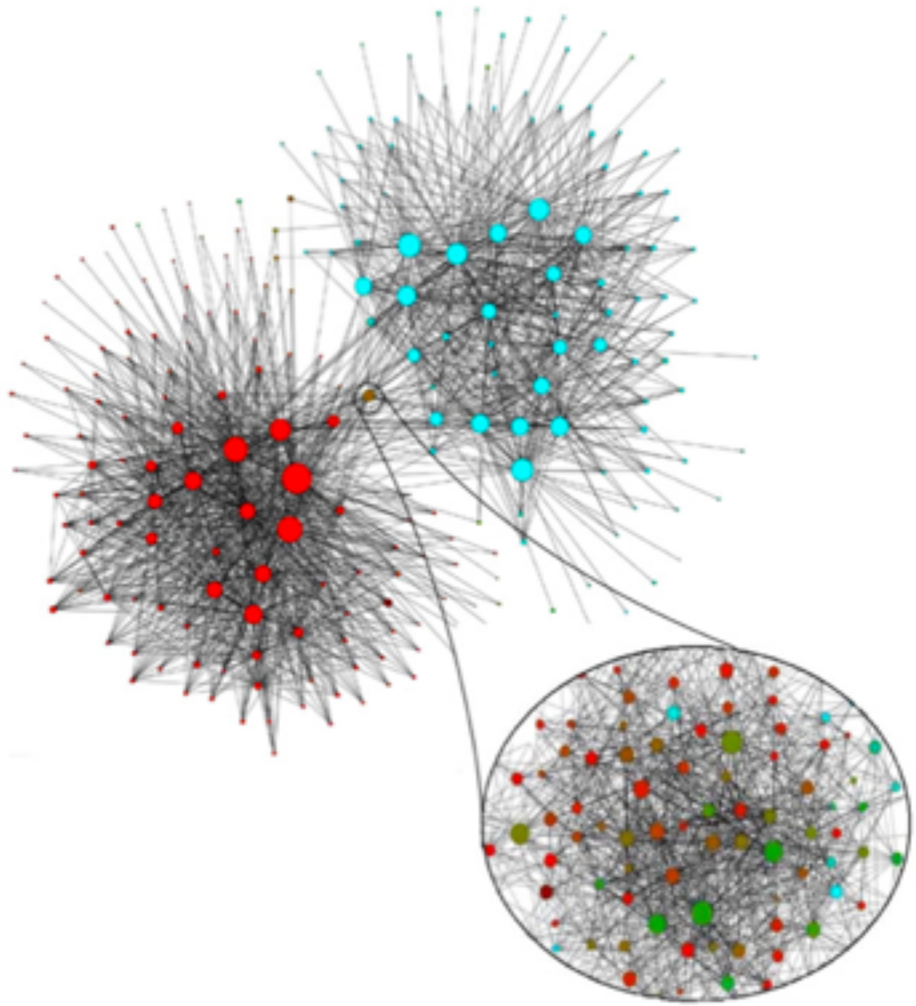
Belgian communication network



V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech.* (2008)

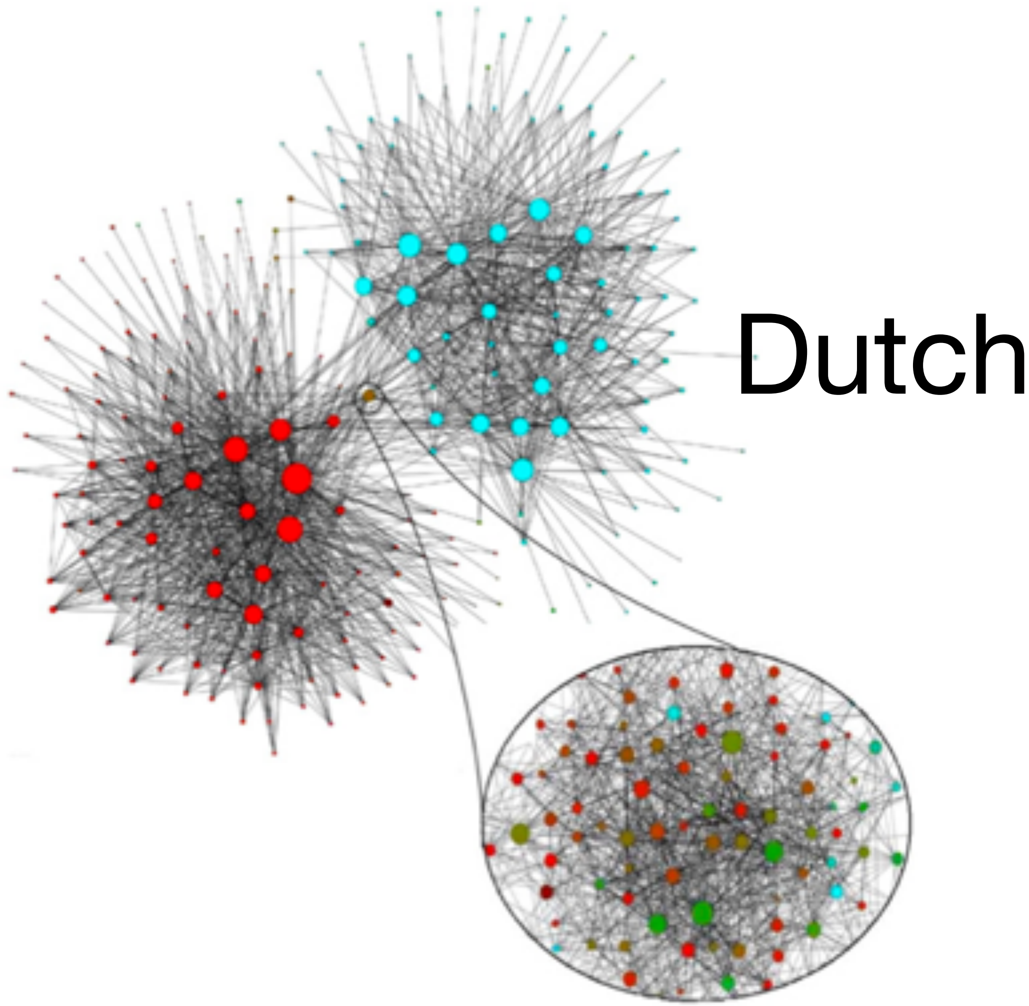
Belgian communication network

French



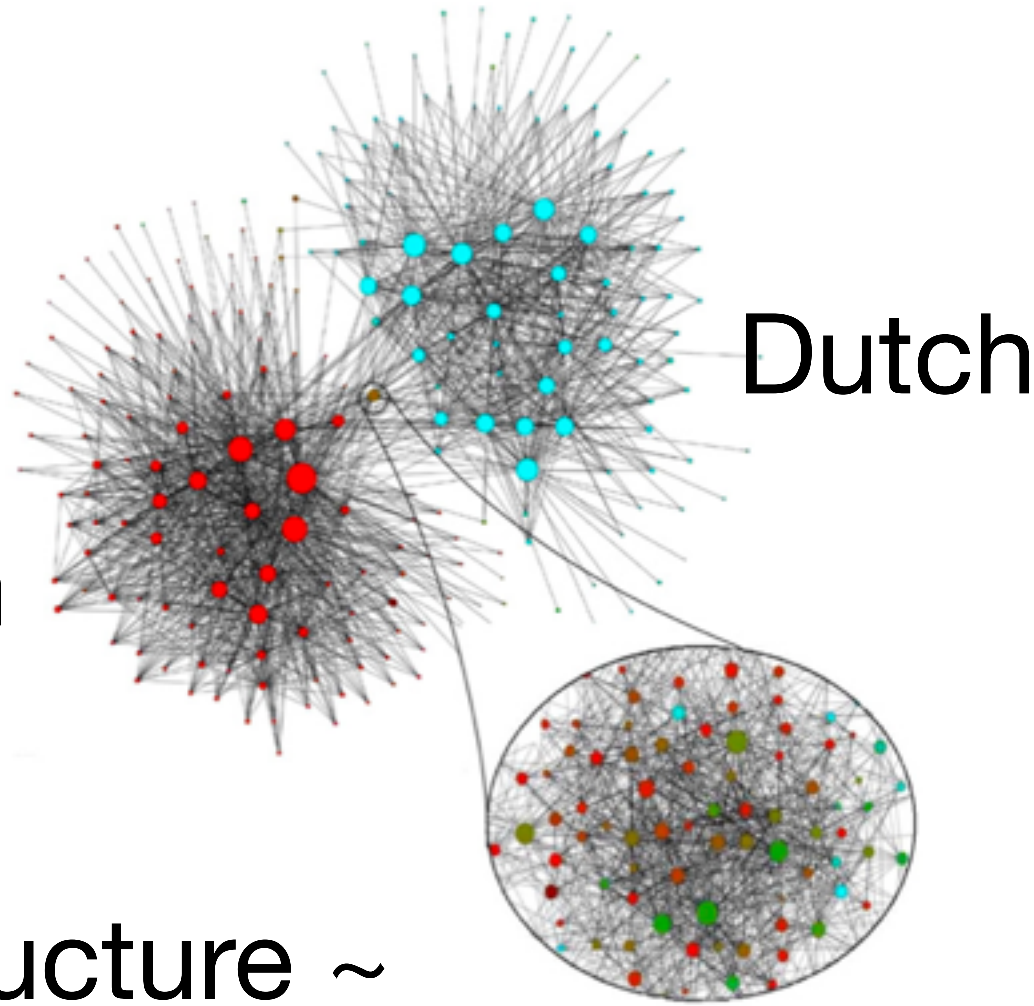
V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech.* (2008)

Belgian communication network



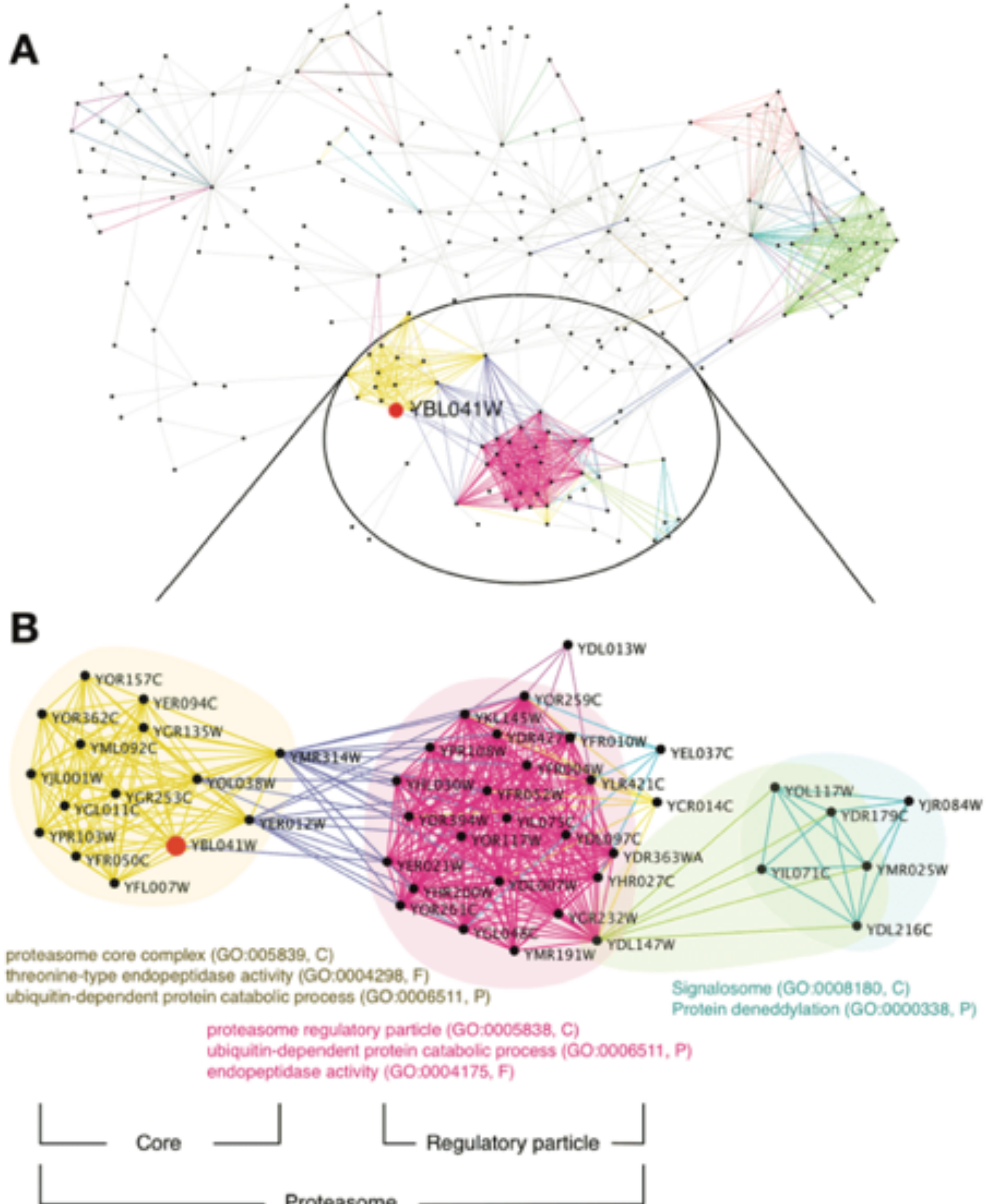
V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech.* (2008)

Belgian communication network

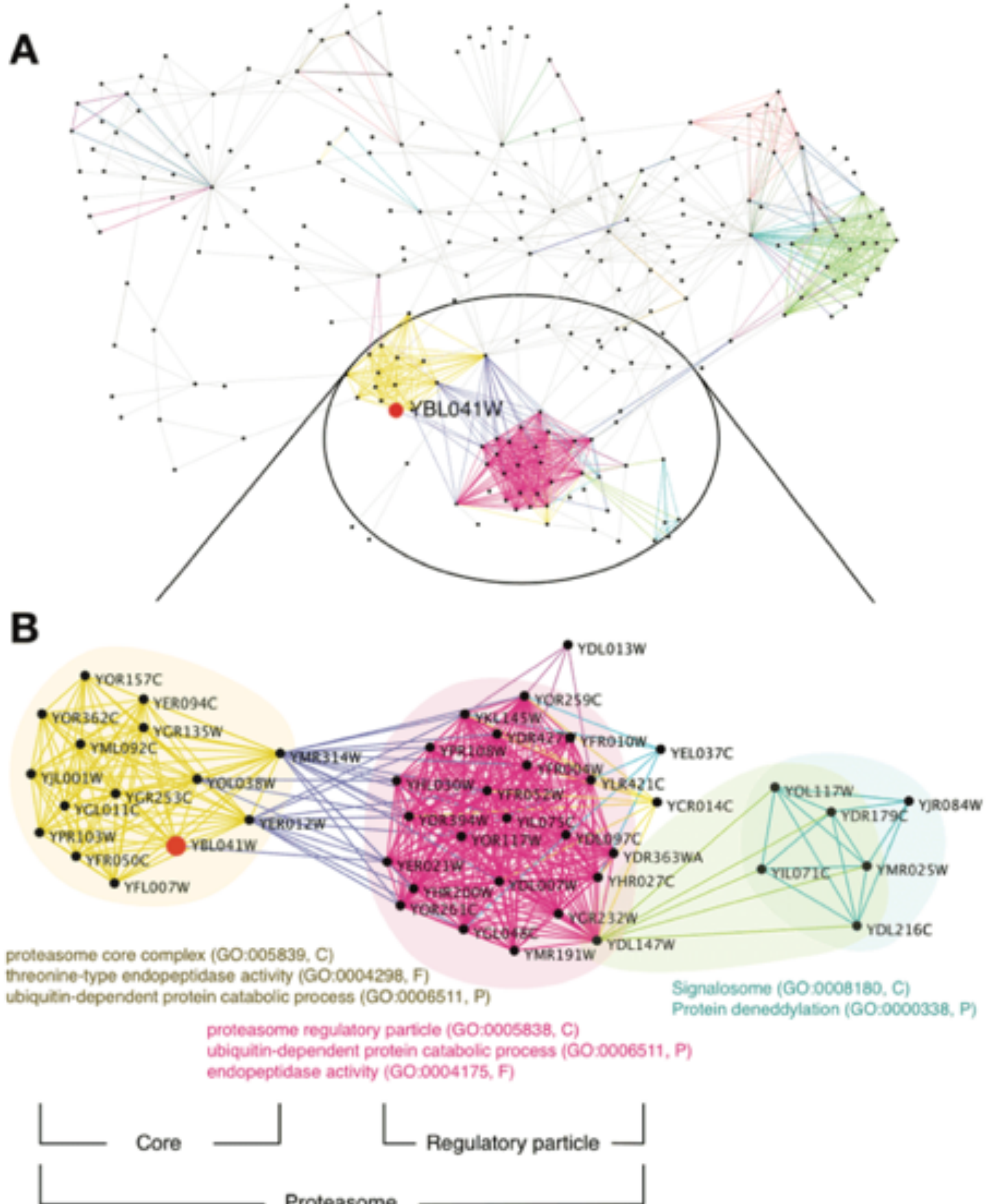


Social structure ~
Communities

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech.* (2008)



Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* (2010)

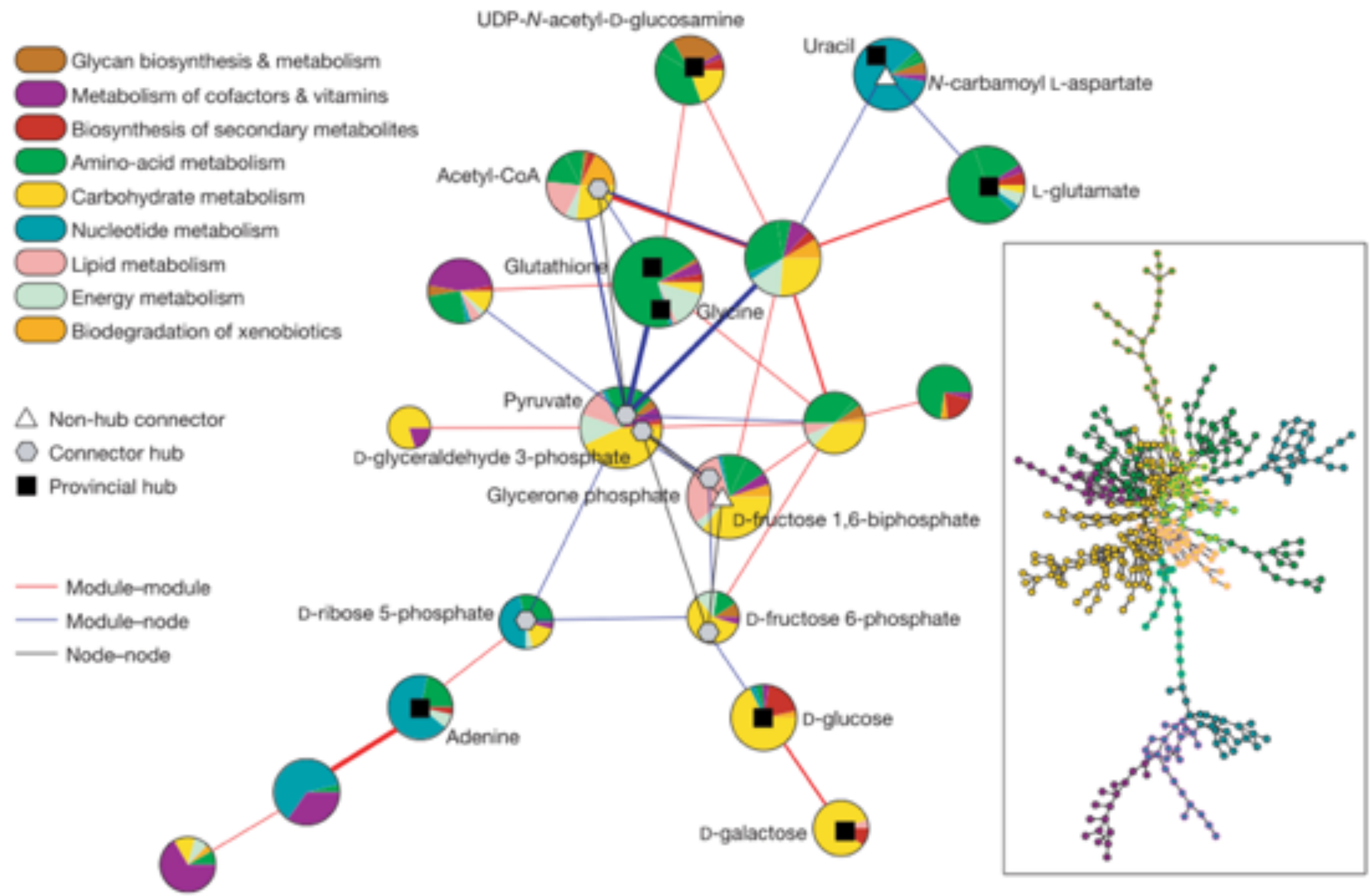


Protein complexes

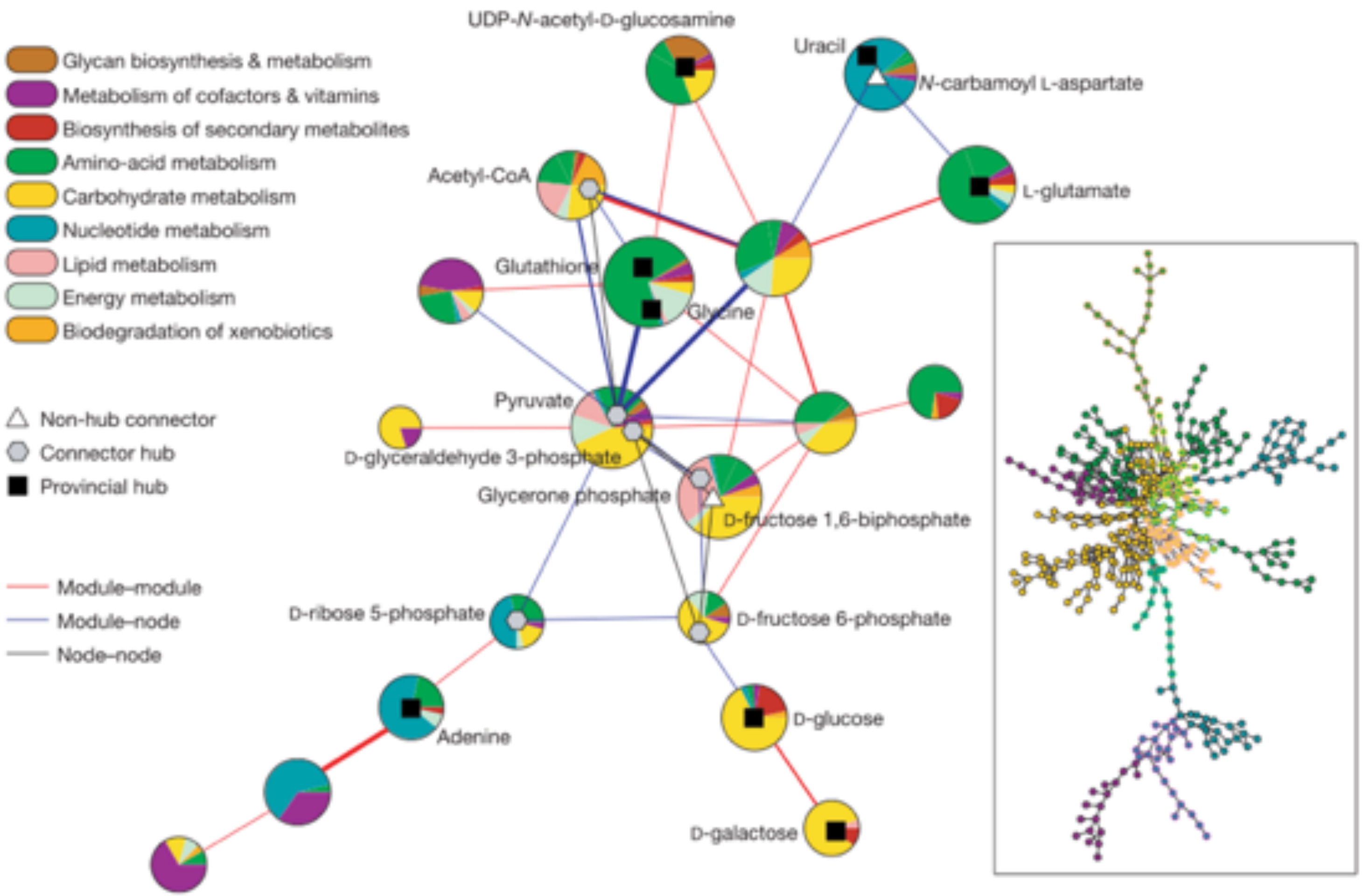
~

communities

Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* (2010)

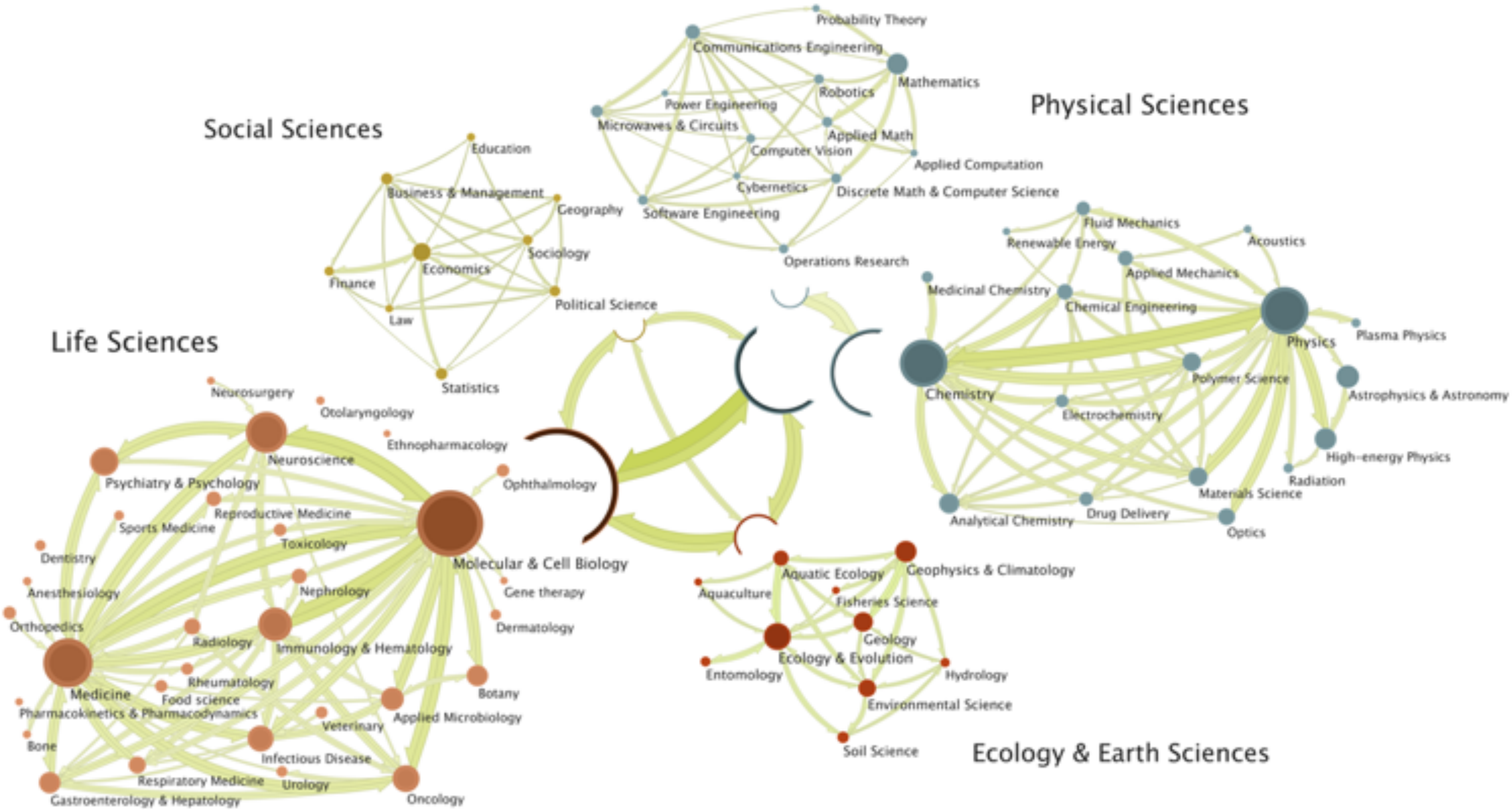


R. Guimerà & L. A. N. Amaral, *Nature* (2005)

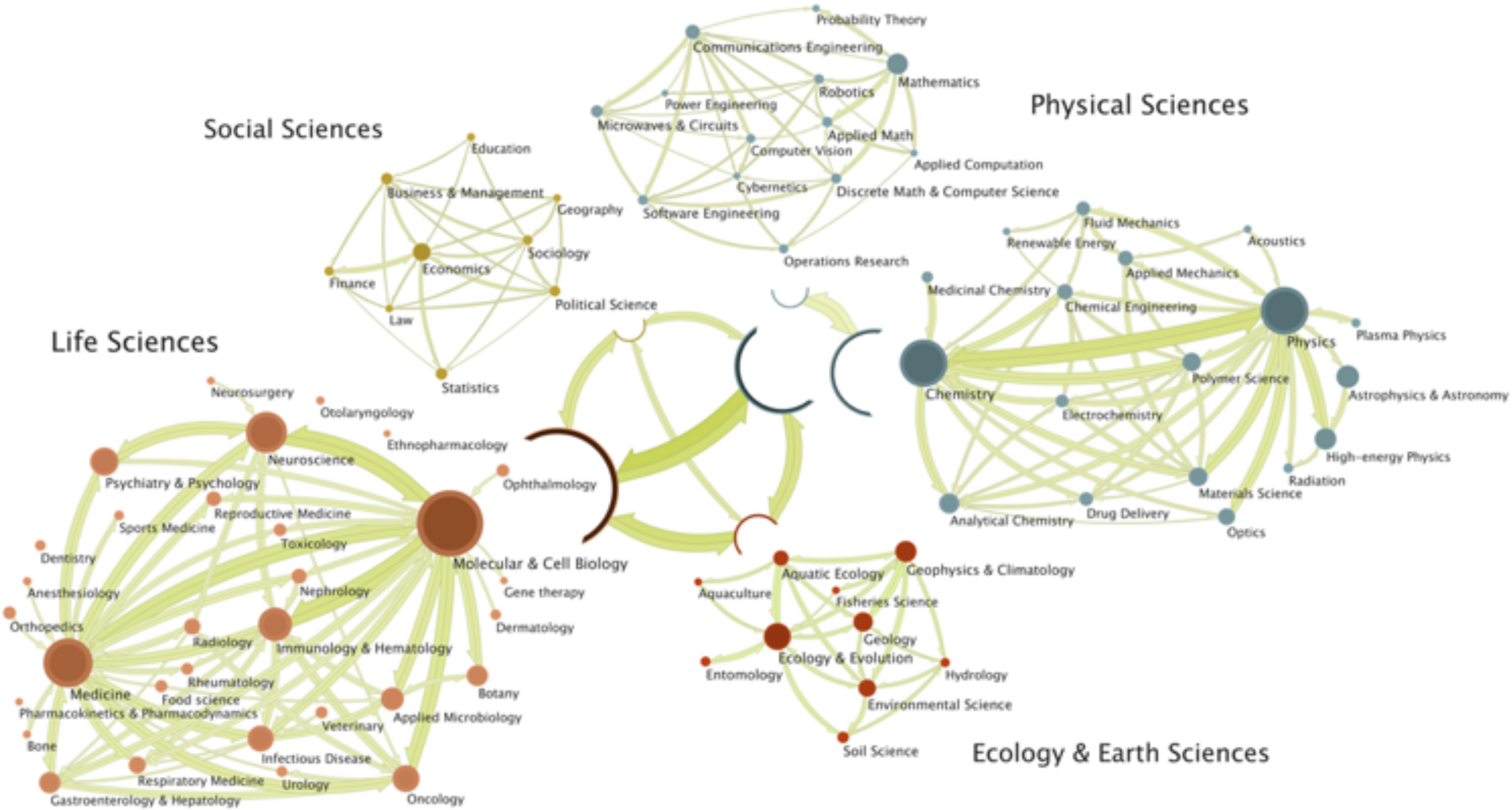


Metabolic pathways ~ **communities**

R. Guimerà & L. A. N. Amaral, *Nature* (2005)



M. Rosvall, C. T. Bergstrom, PLoS One (2011)

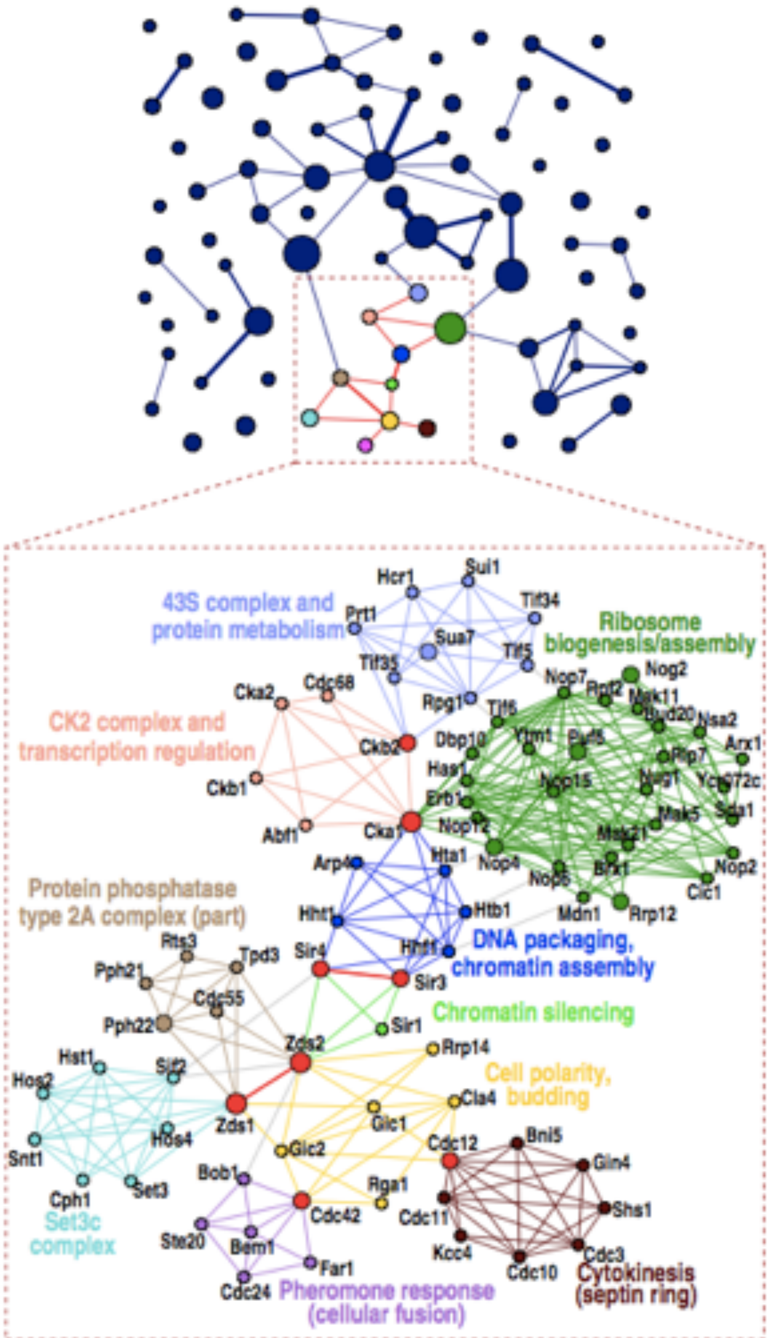


Disciplines ~ **communities**

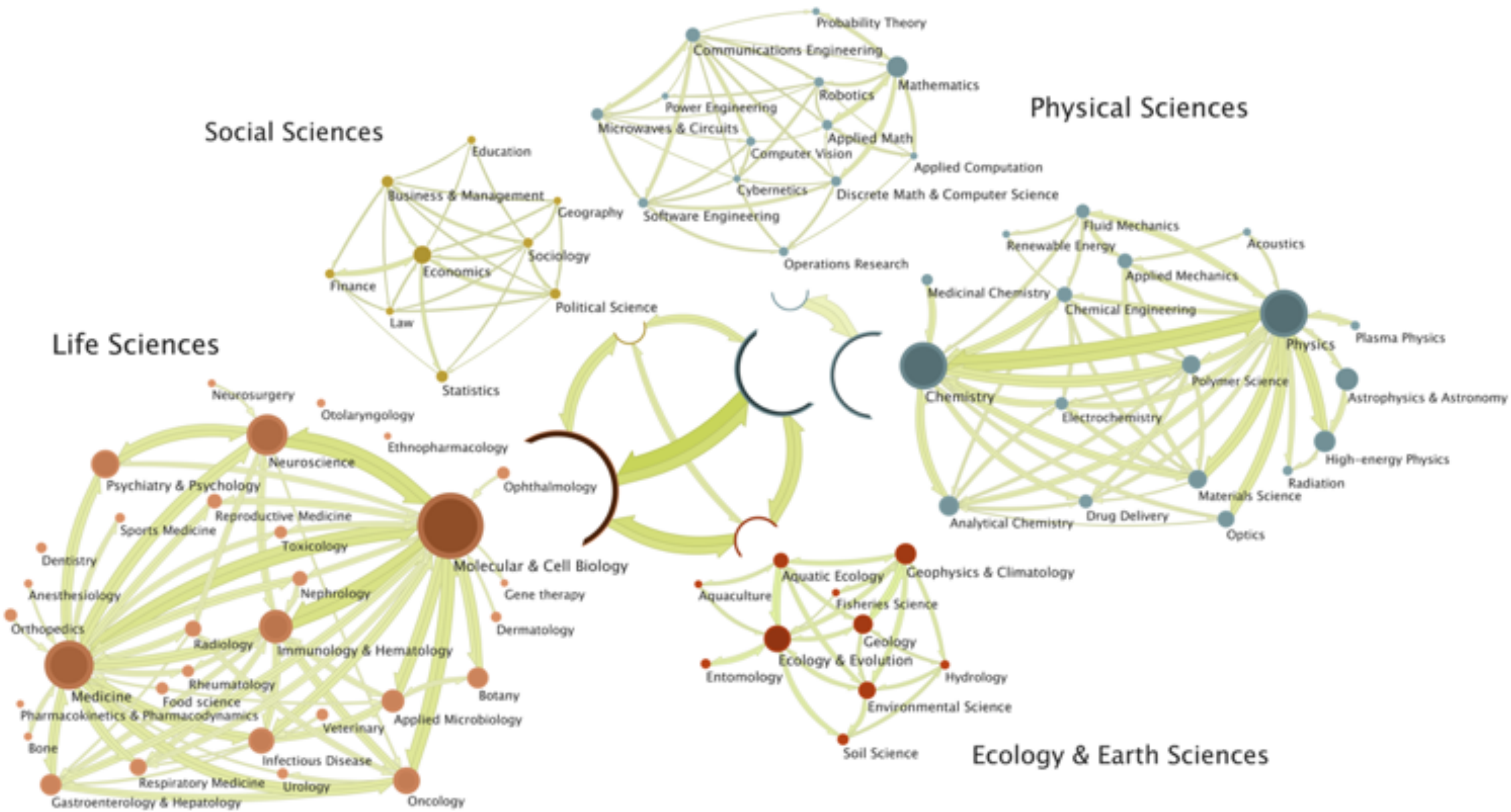
M. Rosvall, C. T. Bergstrom, PLoS One (2011)

Social Networks	Social circles, communities
Biological networks	Protein complexes, functional modules
Citation networks	Disciplines, scientific communities
...	**...**

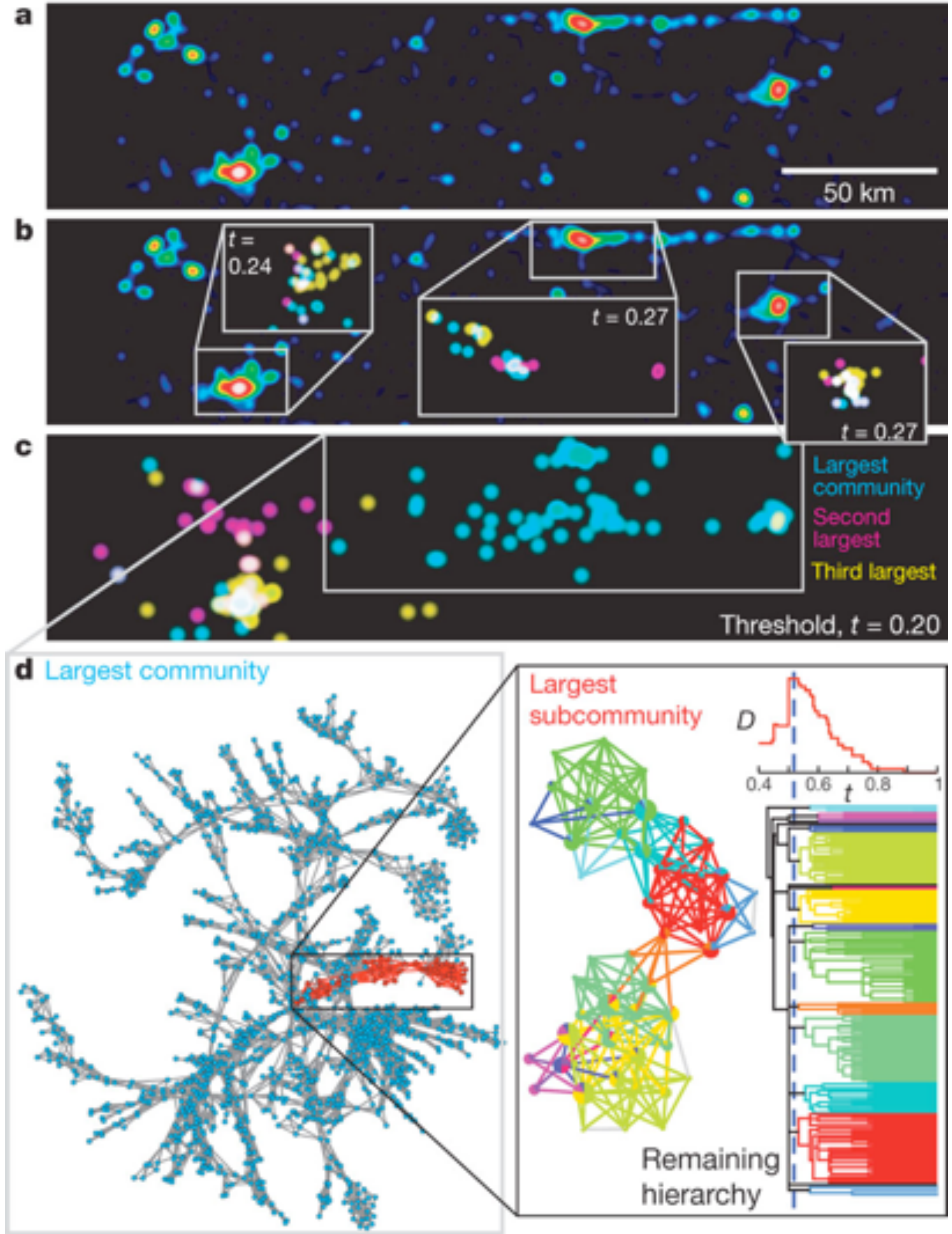
Finding communities:
A nice way to
overview
the whole system



G. Palla, I. Derenyi, I. Farkas, T. Vicsek, *Nature* (2005).



M. Rosvall, C. T. Bergstrom, *PLoS One* (2011)



Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* (2010)

Network community **analysis**

How to
define
communities?

Cohesiveness, Separation

or both

A nice review:
J. Yang and J. Leskovec, Defining and Evaluating Network Communities based on Ground-truth, ICDM 2012

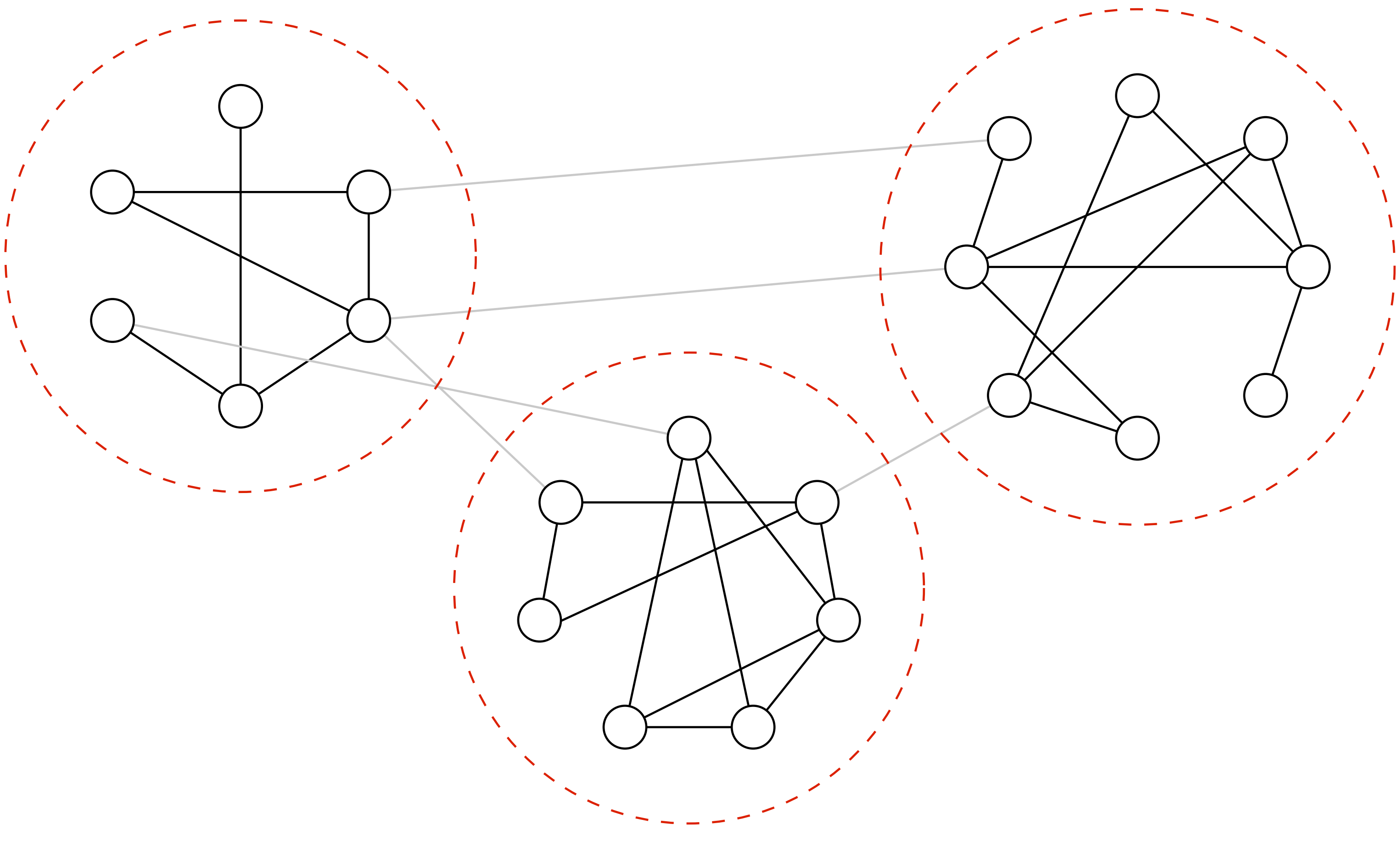
“Cohesiveness”

Clique percolation
Link communities

“Separation”

Girvan-Newman algorithm
Graph cuts
Spectral clustering

Cohesiveness + Separation



Modularity

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

# of links **within**

- # of **expected** links

M. Girvan and M. E. J. Newman, *PNAS* (2002)

How to detect communities?

We should be able to

1. evaluate a community structure
2. explore possible structures effectively

Wait, can we just check **every possible** configurations?

Bell Number: # of partitions of a set of size n.

Bell Number: # of partitions of a set of size n.

$$B_3 = 5$$

Bell Number: # of partitions of a set of size n.

$$B_3 = 5$$

$$B_{100} = ?$$

16187060274460683058556806
28161135741330684513088812
39998409470008912873079240
70443511081340194490281914
80663320741161870602744606
83058556806281611357413306
84513088812399984094700089
12873079240704435110813401
9449028191480663320741

Impossible to enumerate

Fundamental problem of community detection

1. evaluate a community structure

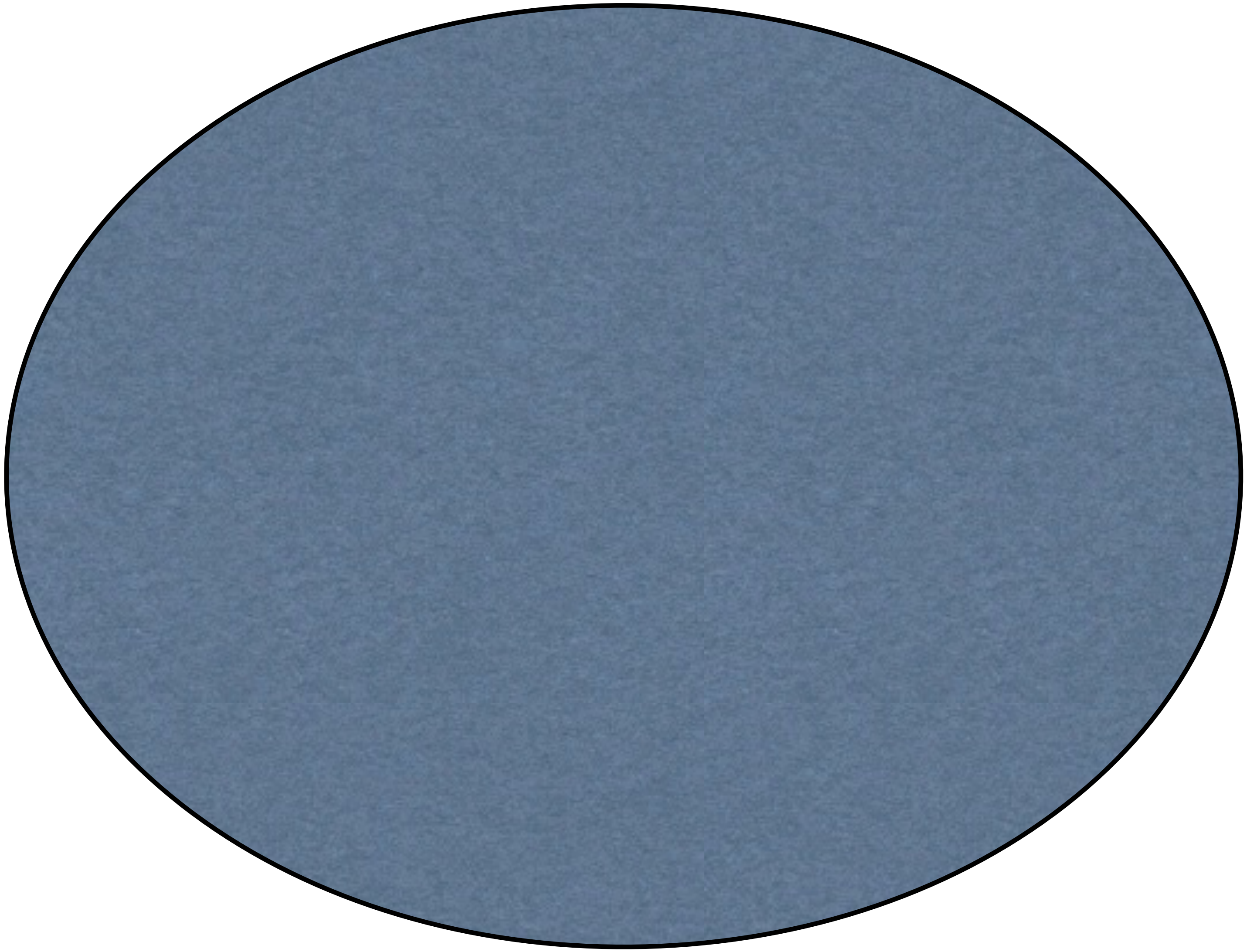
- Modularity, cliques, map equation, partition density, …

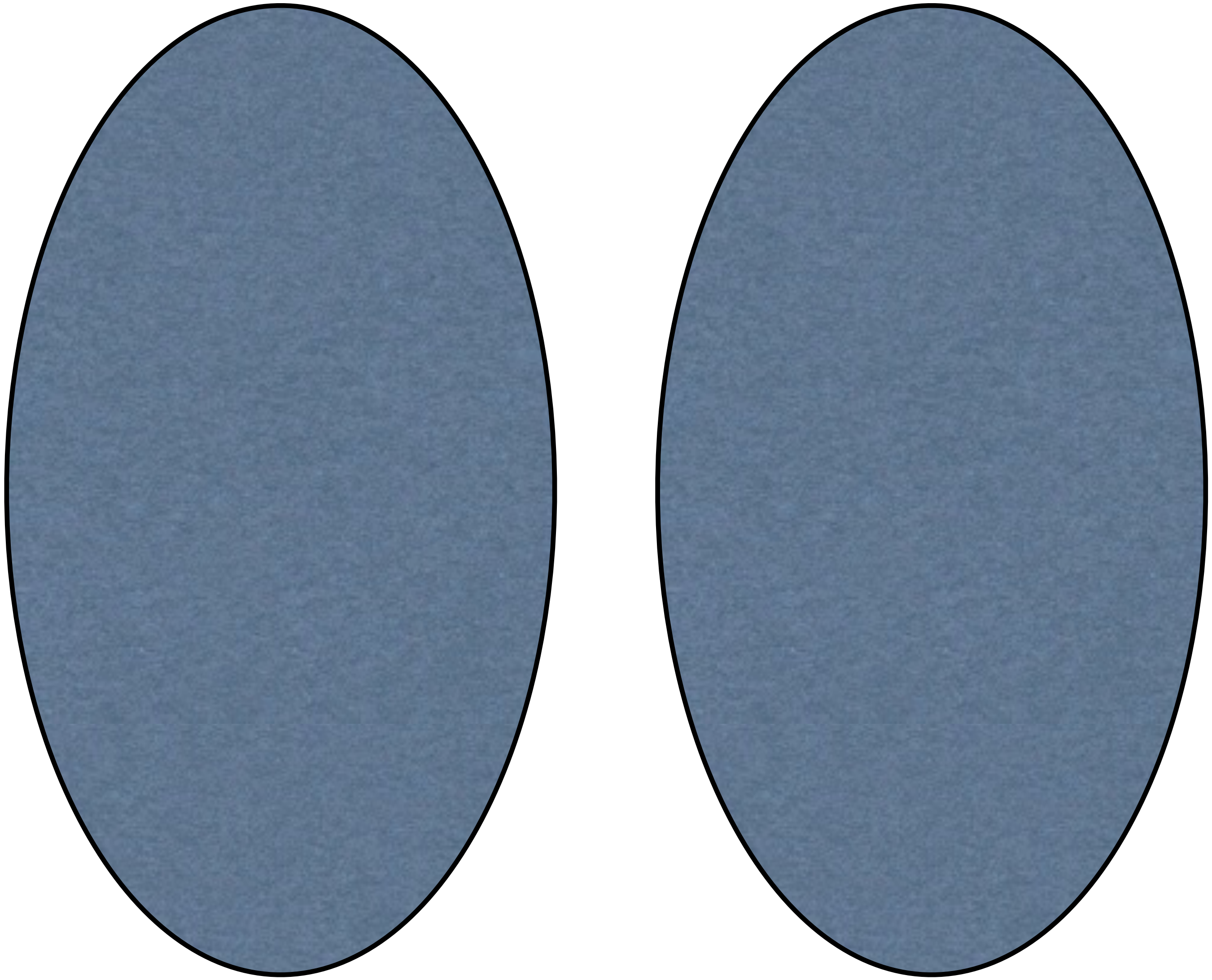
2. explore possible structures effectively

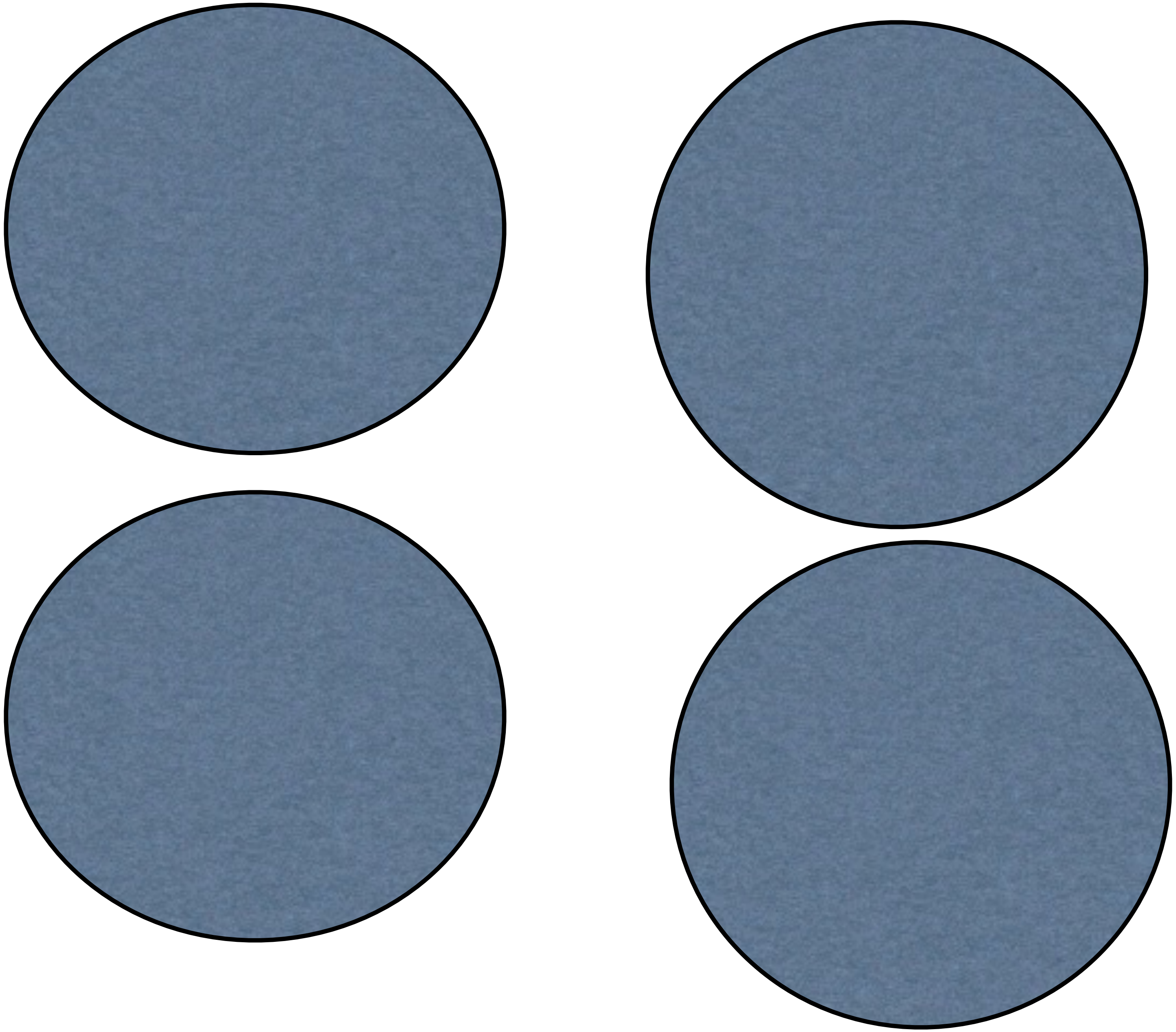
- Many heuristics, Divisive & agglomerative clustering, Monte-carlo, …

Modularity-based methods

Divisive vs. Agglomerative



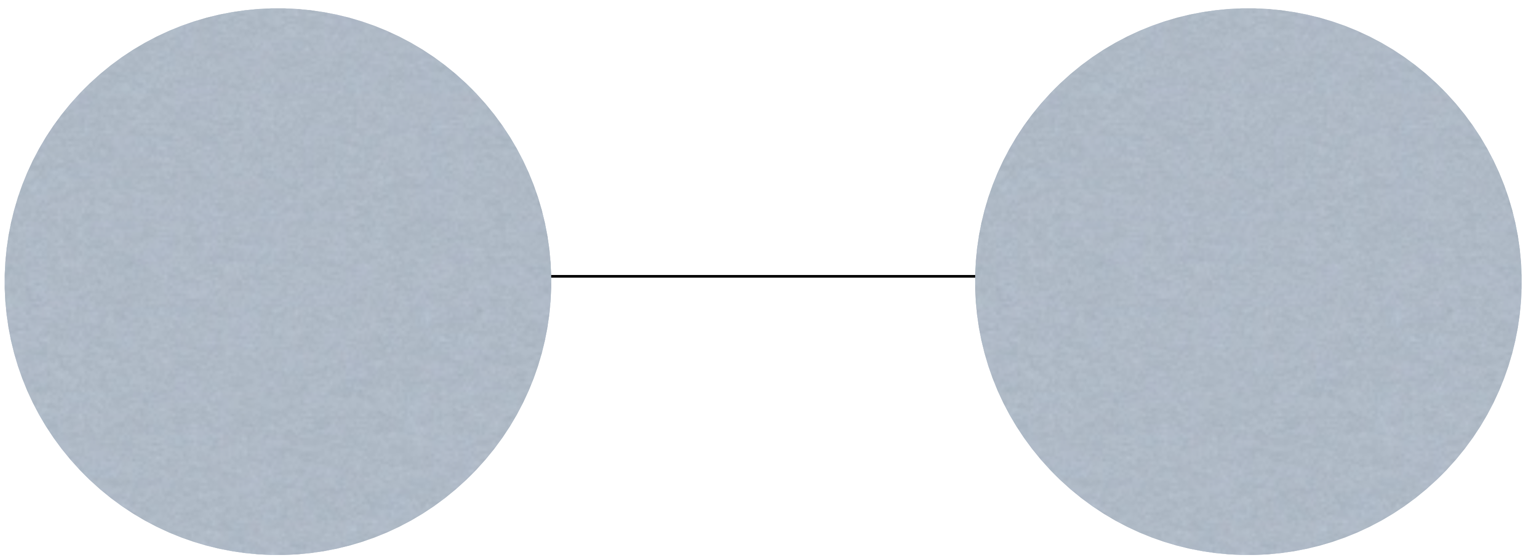


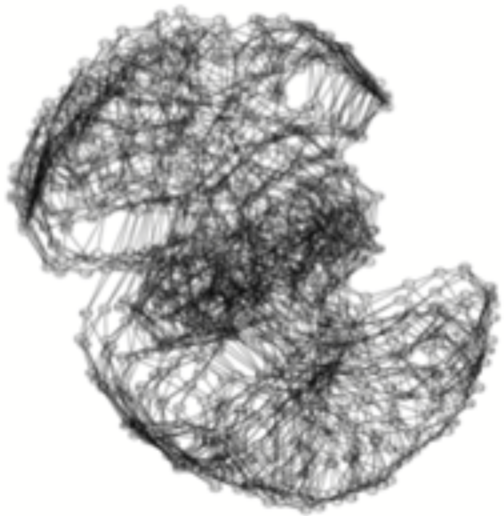


Girvan-Newman algorithm

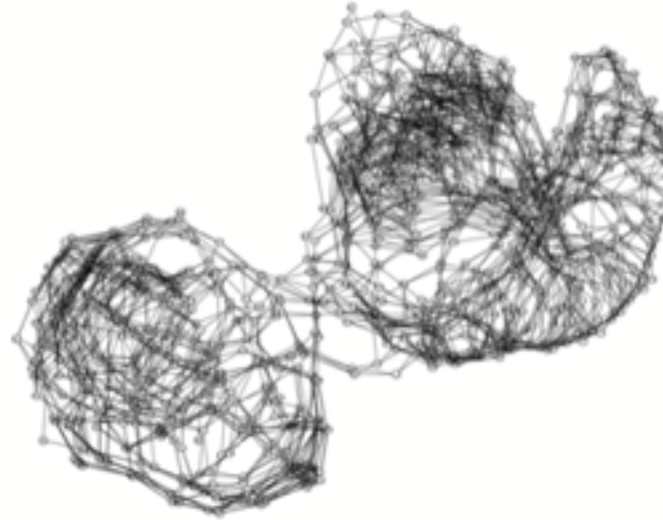
M. Newman, M. Girvan, PNAS (2002)

Idea

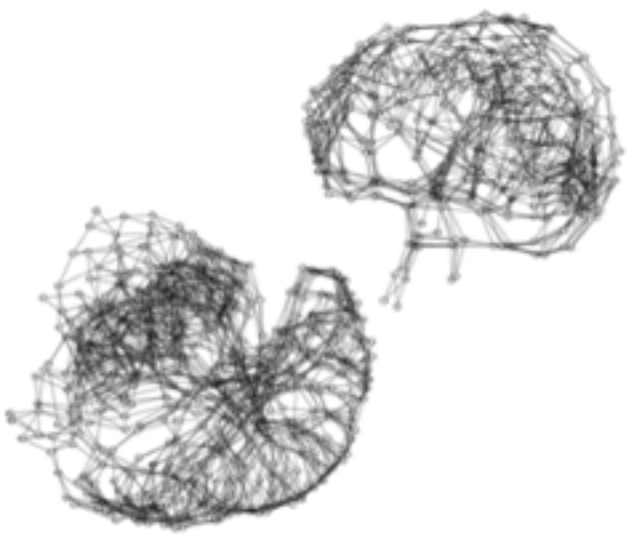




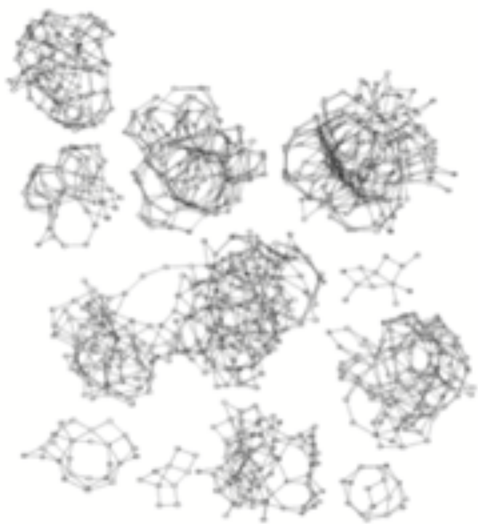
0 cuts



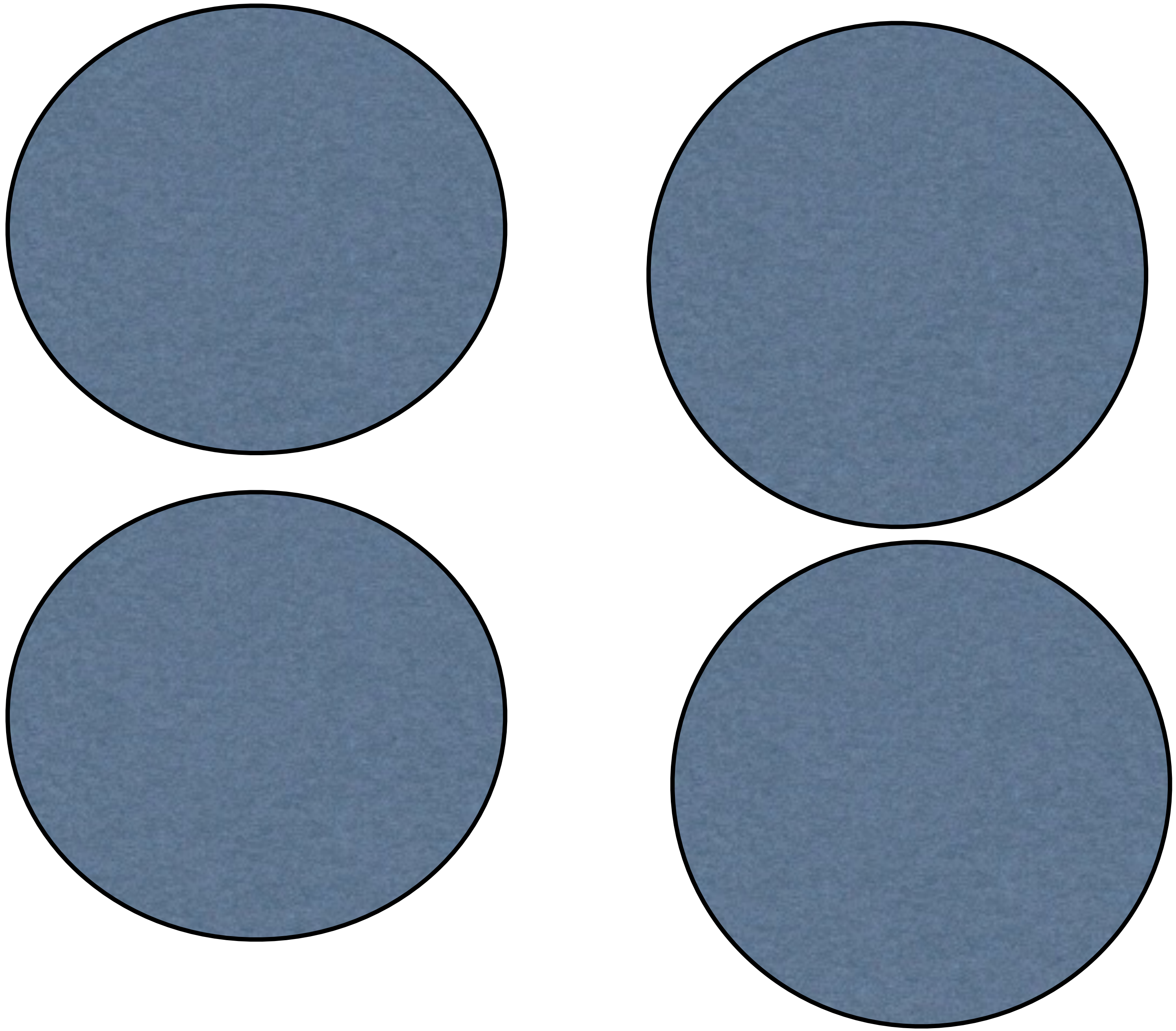
100 cuts

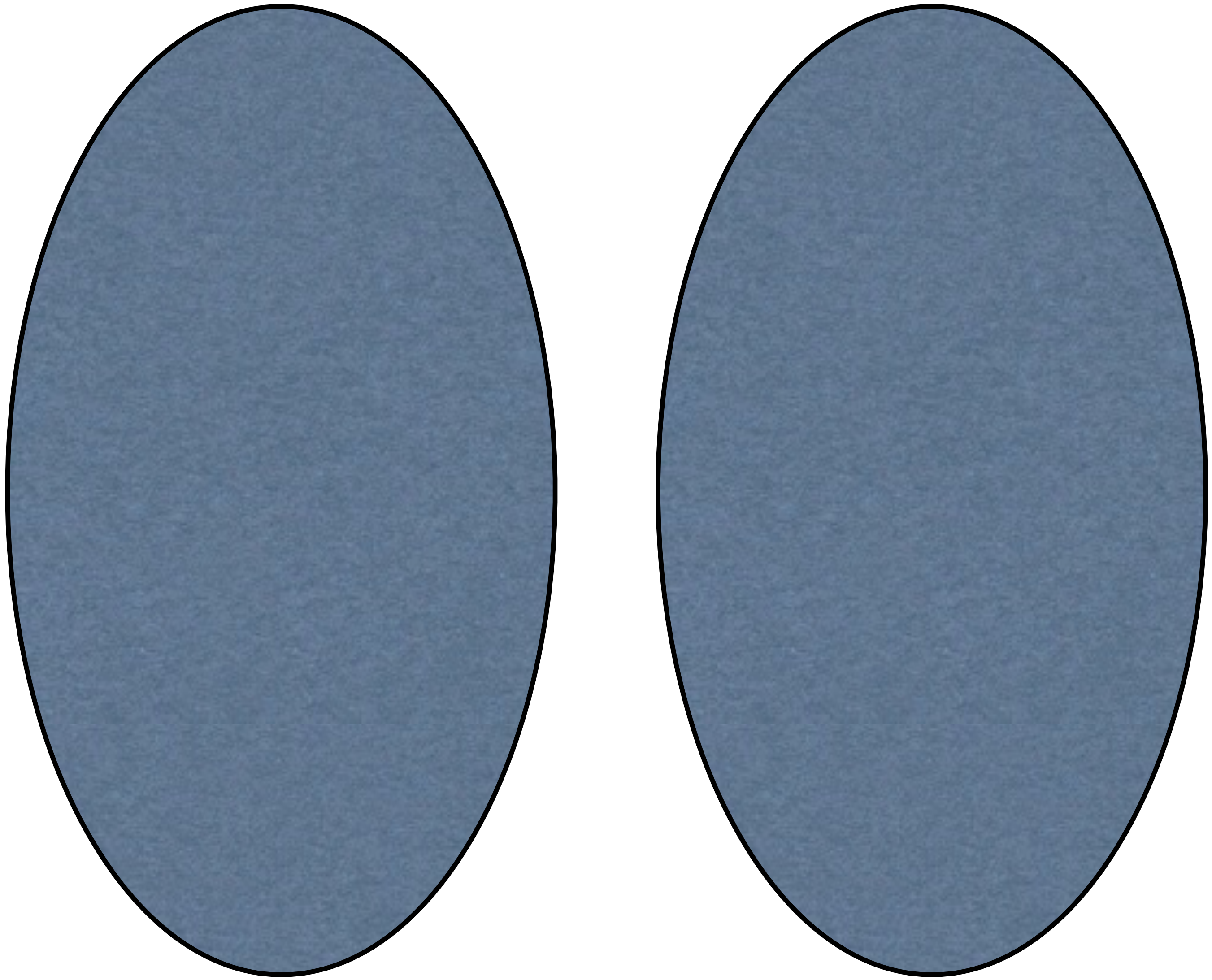


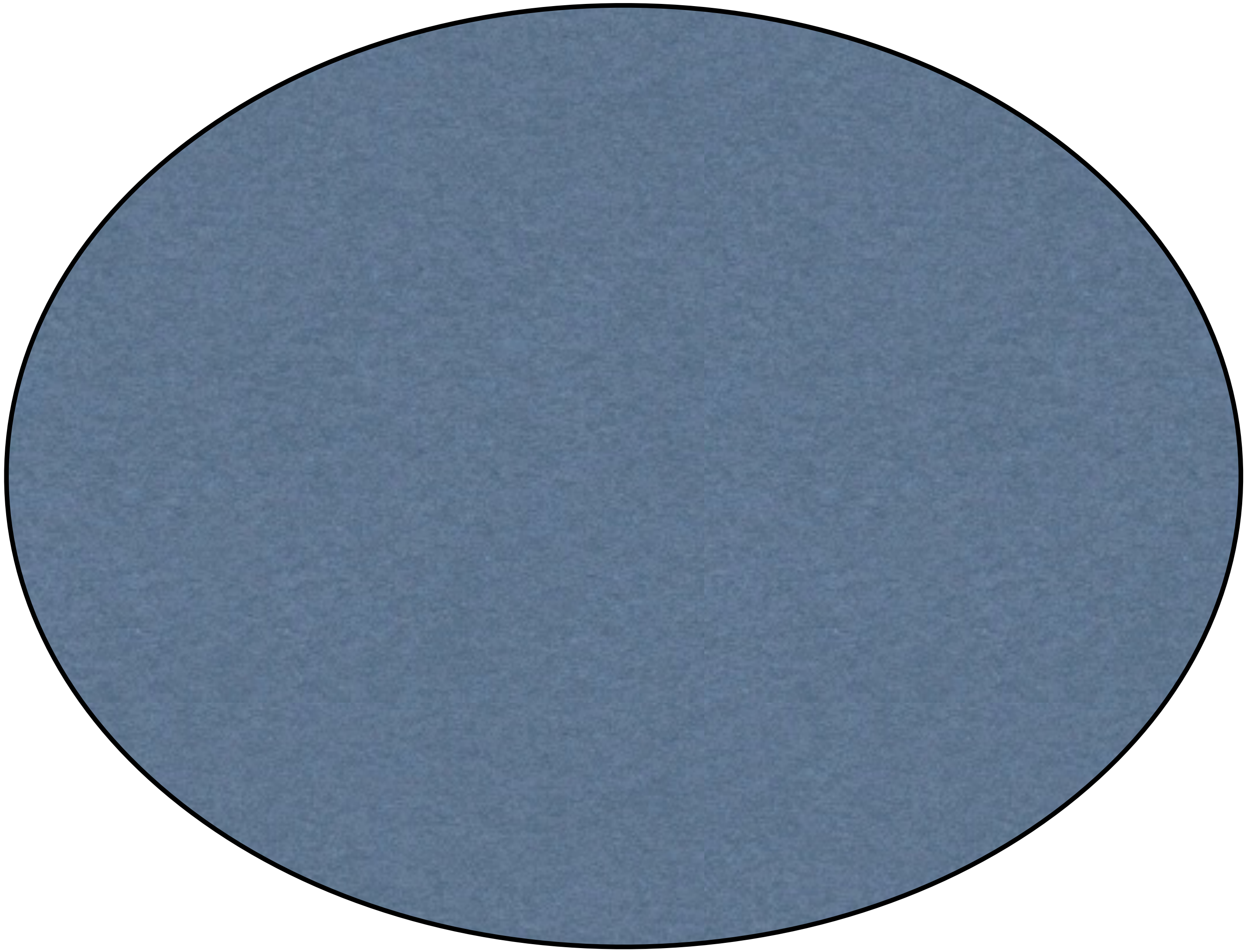
120 cuts



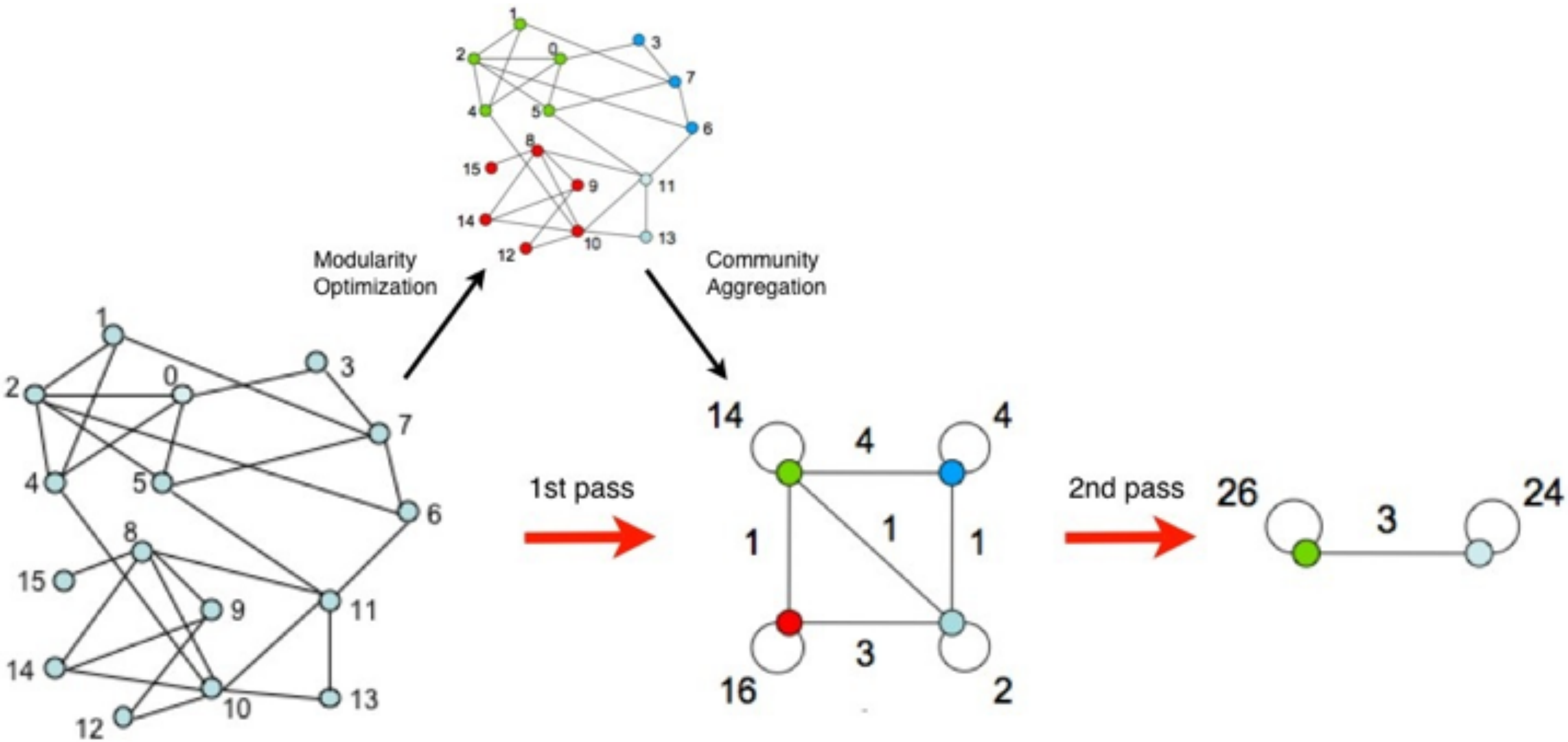
500 cuts







Louvain method



V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre JSTAT (2008).

Various optimization techniques

- A. Clauset, M. Newman, C. Moore: Greedy optimization
- R. Guimera, L. A. N. Amaral: Extremal optimization
- V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre: Hierarchical aggregation
- **Any** optimization technique can be used.

“Cliques”

What is a ‘perfect community’?

A clique!

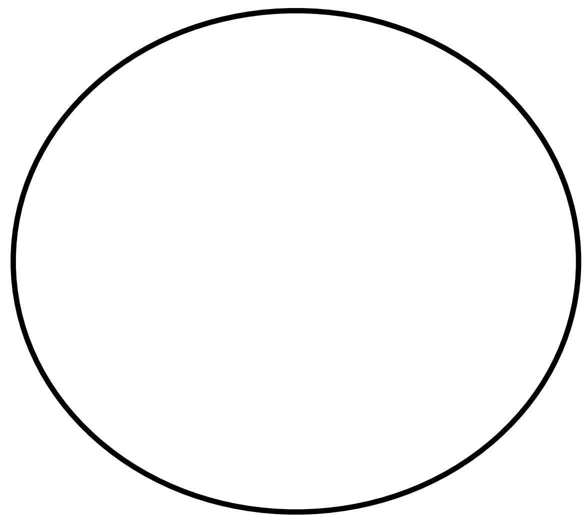
Then, how about finding **quasi-cliques**?

Clique Percolation Method

- “Rolling” a clique to find a quasi-clique.
- Quasi-cliques are communities.

G. Palla, I. Derenyi, I. Farkas, T. Vicsek, *Nature* (2005).

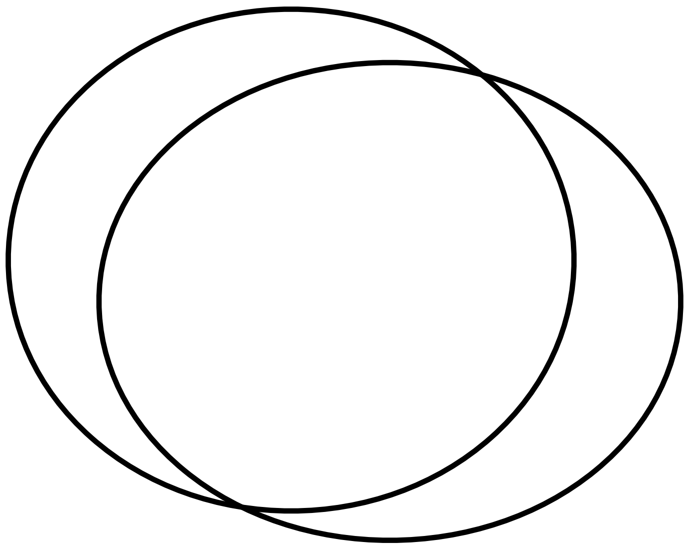
Clique Percolation Method



- “Rolling” a clique to find a quasi-clique.
- Quasi-cliques are communities.

G. Palla, I. Derenyi, I. Farkas, T. Vicsek, *Nature* (2005).

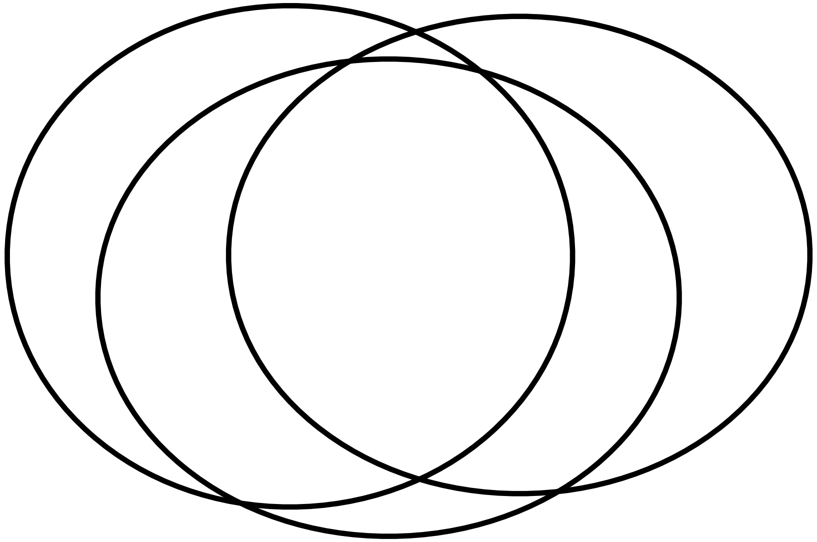
Clique Percolation Method



- “Rolling” a clique to find a quasi-clique.
- Quasi-cliques are communities.

G. Palla, I. Derenyi, I. Farkas, T. Vicsek, *Nature* (2005).

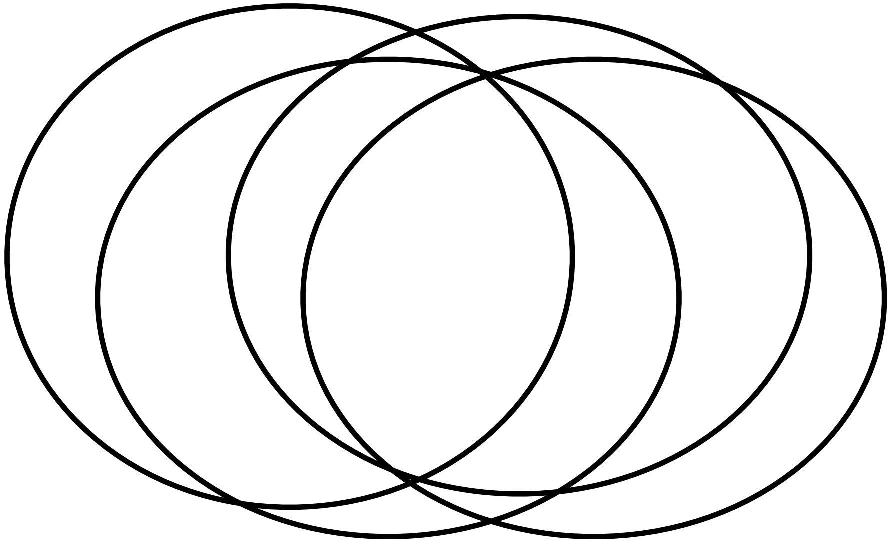
Clique Percolation Method



- “Rolling” a clique to find a quasi-clique.
- Quasi-cliques are communities.

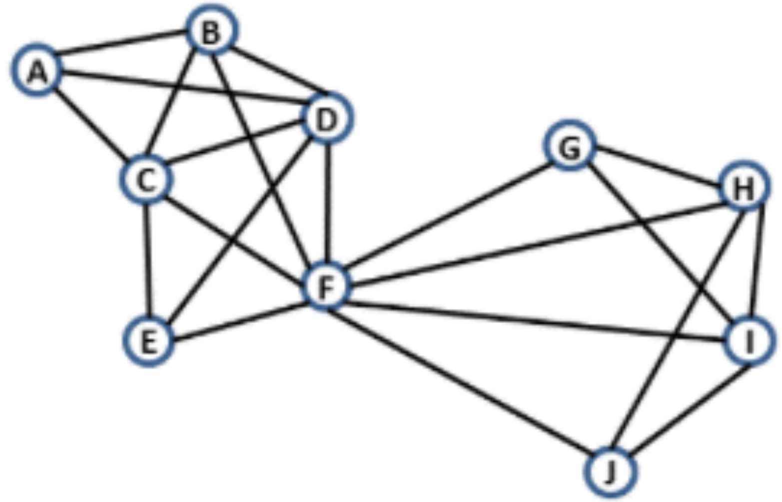
G. Palla, I. Derenyi, I. Farkas, T. Vicsek, *Nature* (2005).

Clique Percolation Method



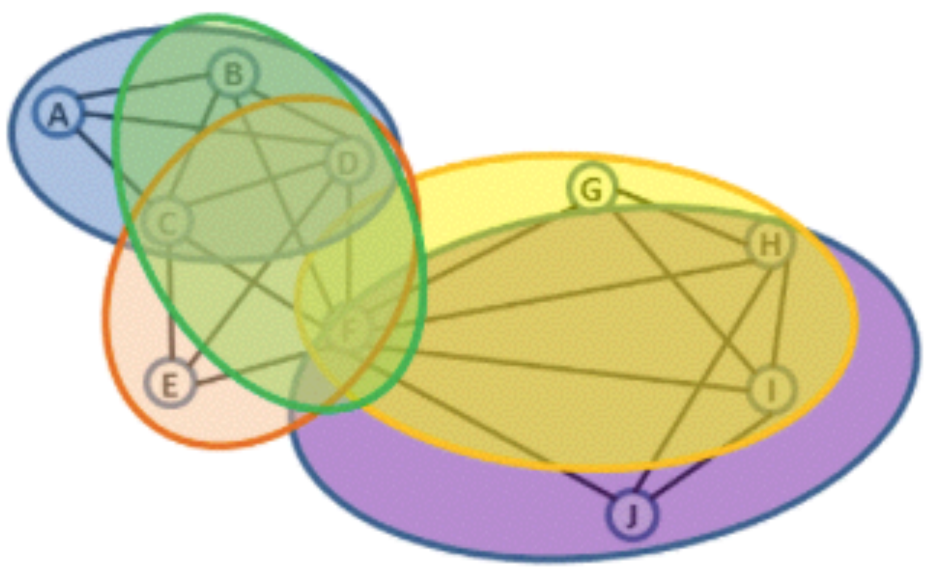
- “Rolling” a clique to find a quasi-clique.
- Quasi-cliques are communities.

G. Palla, I. Derenyi, I. Farkas, T. Vicsek, *Nature* (2005).



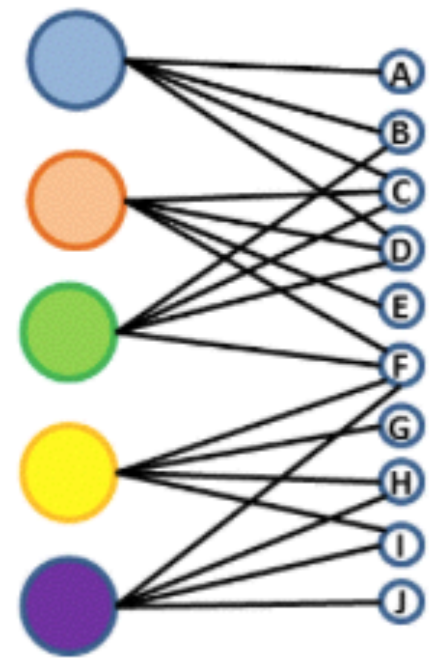
Original Graph

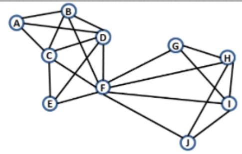
Step 1: Find all K-Cliques (K = 4)



Bi-partite graph

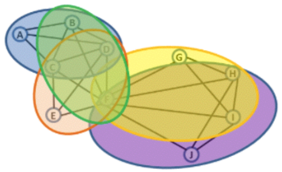
Clique Node





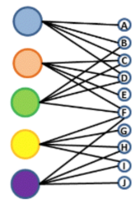
Original Graph

Step 1: Find all K-Cliques (K = 4)

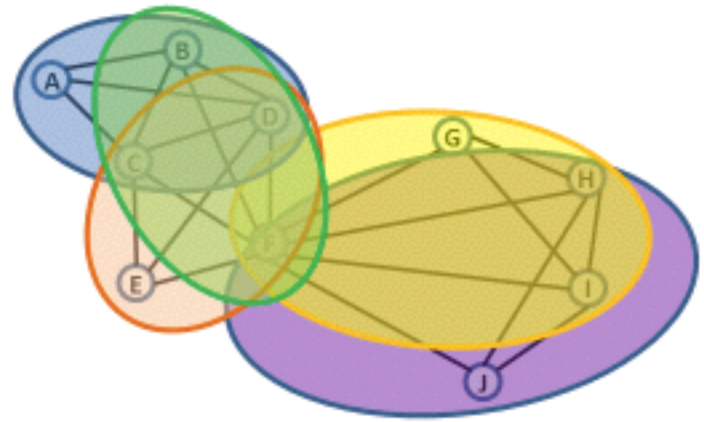


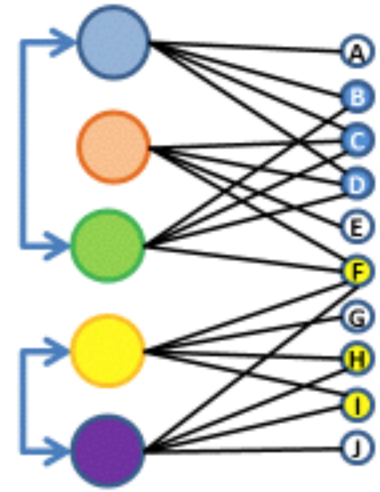
Bi-partite graph

Clique Node

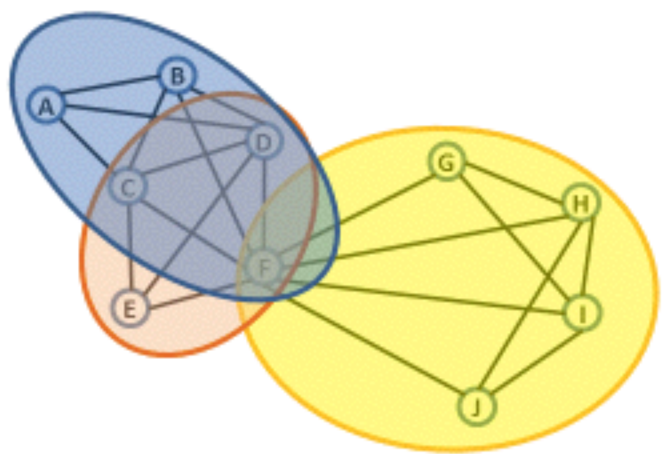


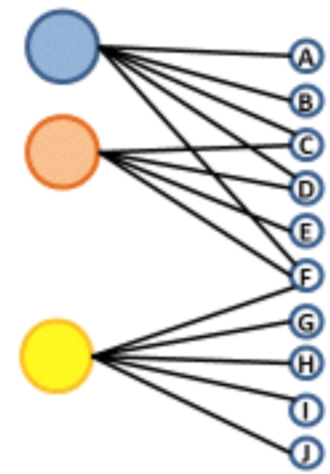
Step 2: Combine adjacent cliques (with K-1 = 3 shared nodes)



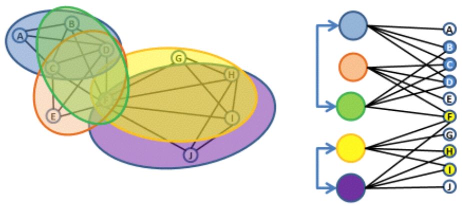


After merging adjacent cliques

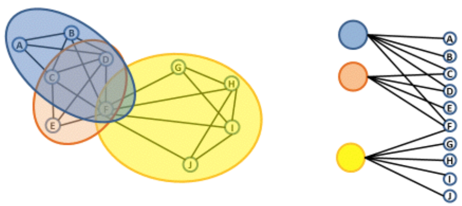




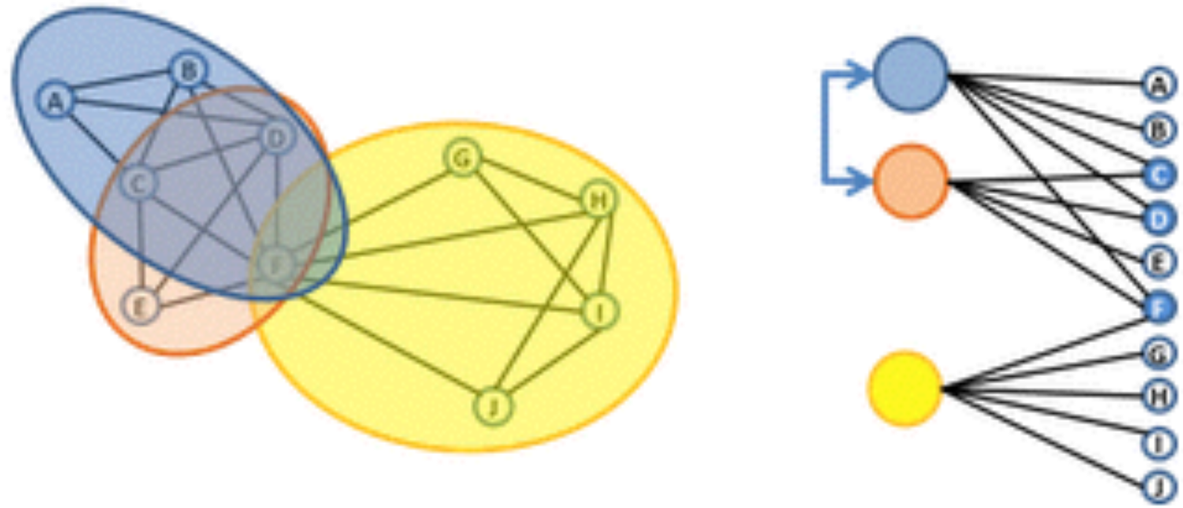
Step 2: Combine adjacent cliques (with K-1 = 3 shared nodes)



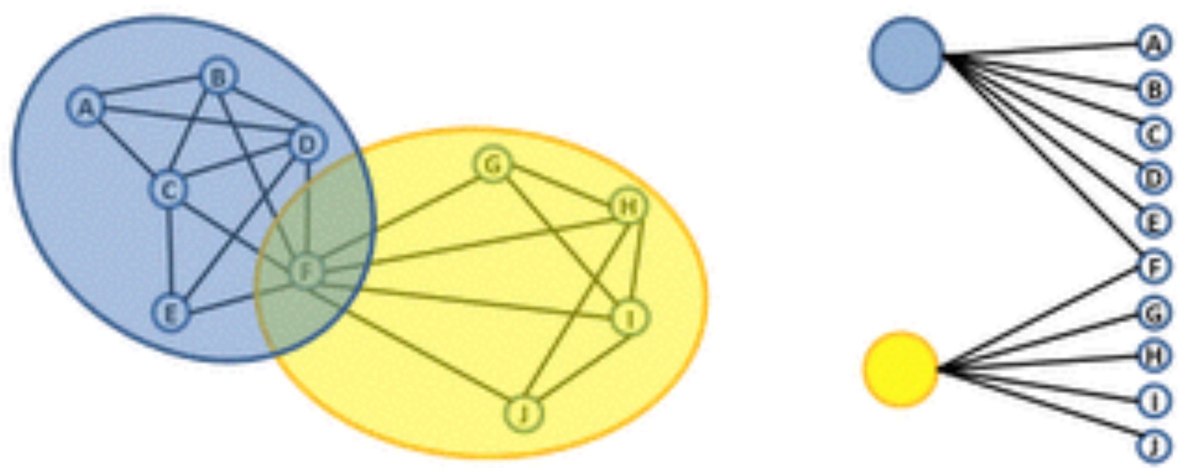
After merging adjacent cliques

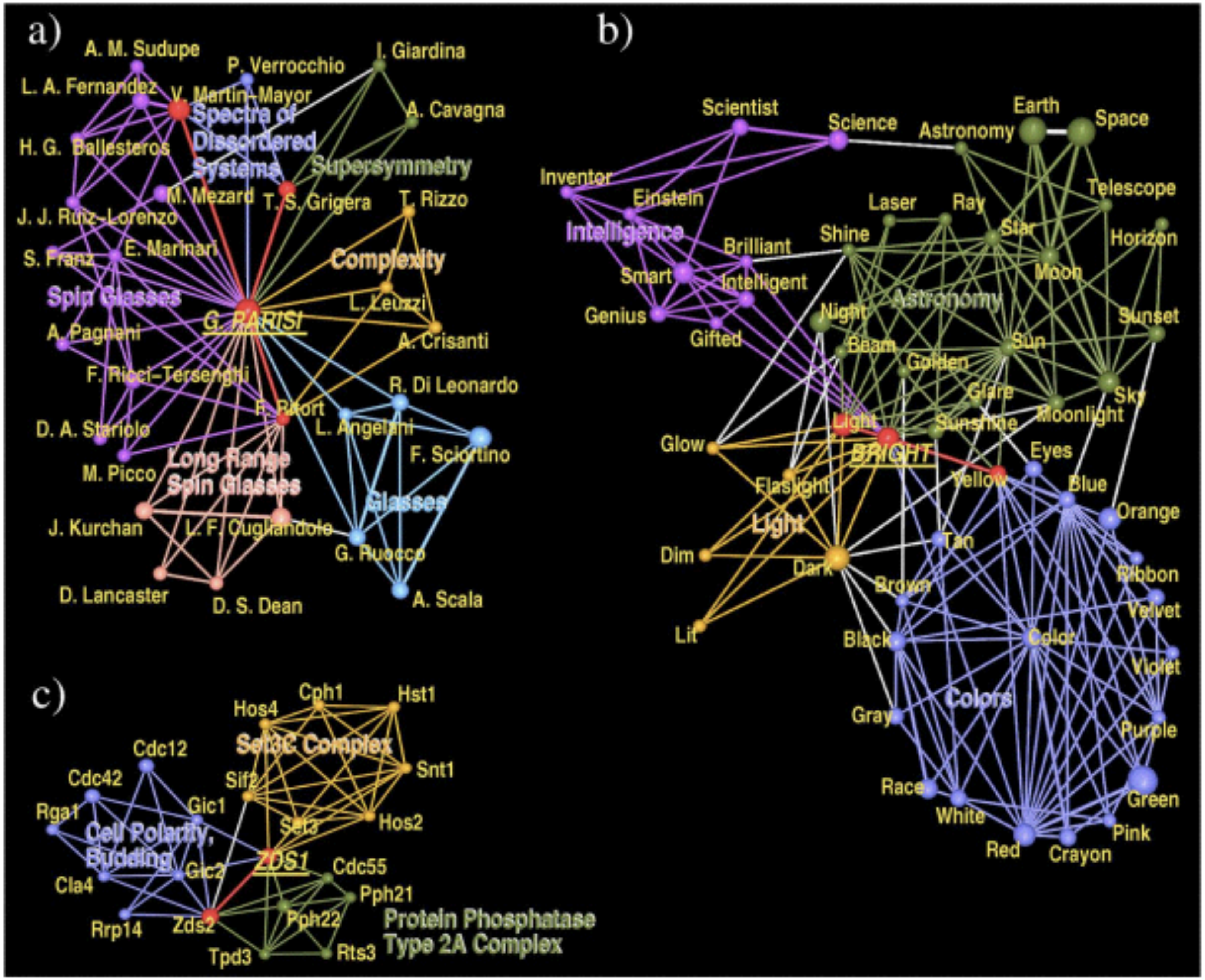


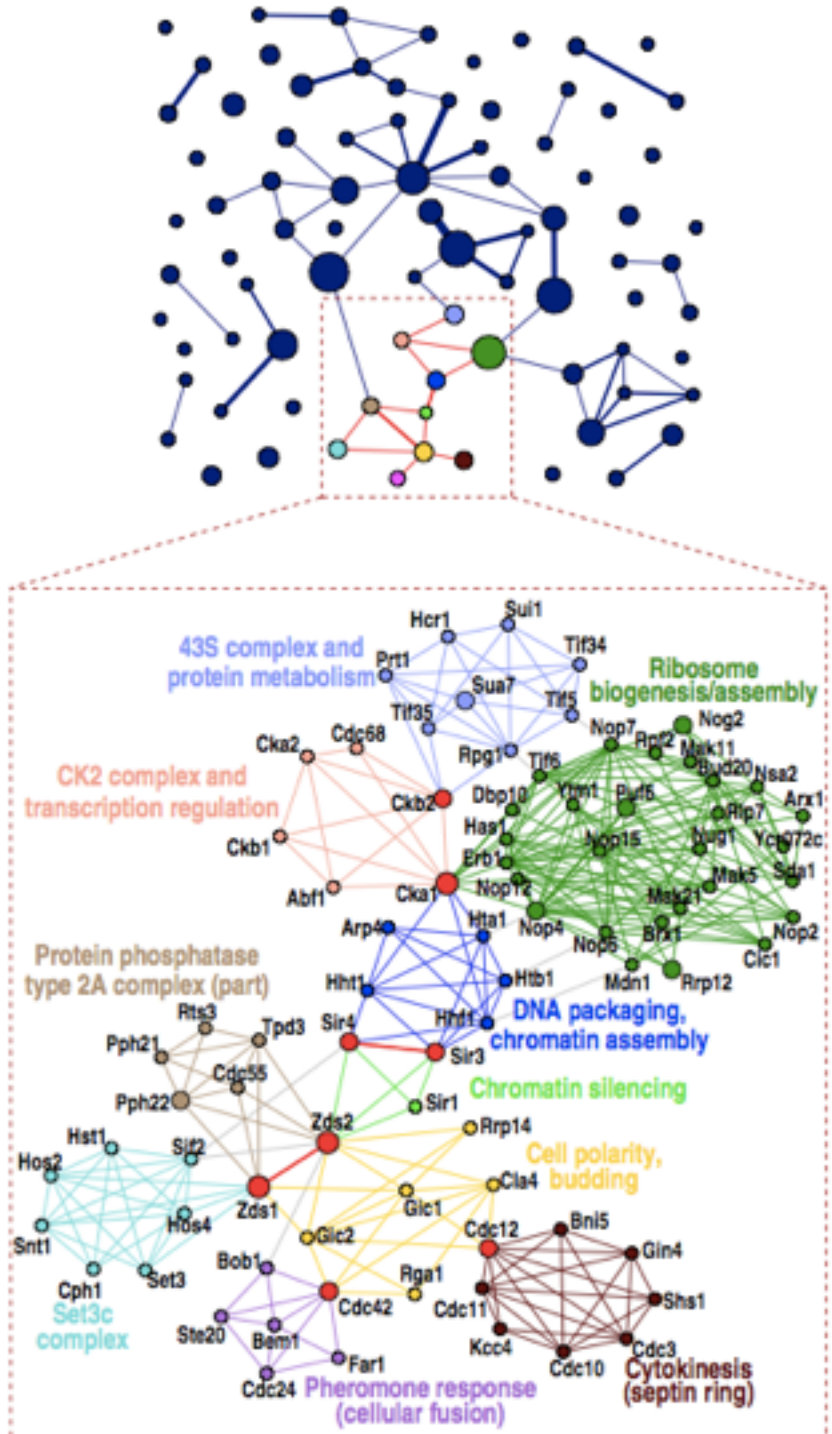
Step 3: Combine adjacent cliques (with K-1 = 3 shared nodes)



After merging adjacent cliques, there are 2 overlapping communities




a)
A. M. Sudupe
P. Verrocchio
I. Giardina
L. A. Fernandez
V. Martin-Mayor
Spectra of Disordered Systems
A. Cavagna
H. G. Ballesteros
Supersymmetry
M. Mezard
T. S. Grigera
T. Rizzo
J. J. Ruiz-Lorenzo
S. Franz
E. Marinari
Complexity
Spin Glasses
L. Leuzzi
G. PARISI
A. Pagnani
A. Crisanti
F. Ricci-Tersenghi
R. Di Leonardo
F. Ritort
L. Angelani
D. A. Stariolo
F. Sciortino
Long Range Spin Glasses
M. Picco
Glasses
J. Kurchan
L. F. Cugliandolo
G. Ruocco
D. Lancaster
D. S. Dean
A. Scala
b)
Scientist
Science
Astronomy
Earth
Space
Inventor
Einstein
Telescope
Laser
Ray
Intelligence
Shine
Star
Horizon
Brilliant
Moon
Smart
Intelligent
Astronomy
Night
Sunset
Genius
Gifted
Beam
Sun
Golden
Glare
Sky
Moonlight
Light
Sunshine
Glow
BRIGHT
Eyes
Yellow
Blue
Flashlight
Orange
Light
Tan
Dim
Dark
Brown
Ribbon
Velvet
Lit
Black
Color
Violet
Colors
Gray
Purple
Race
White
Green
Pink
Red
Crayon
c)
Cph1
Hst1
Hos4
Set3C Complex
Cdc12
Snt1
Cdc42
Sif2
Gic1
Rga1
Hos2
Set3
Cell Polarity, Budding
ZDS1
Gic2
Cdc55
Cla4
Pph21
Pph22
Protein Phosphatase Type 2A Complex
Rrp14
Zds2
Tpd3
Rts3



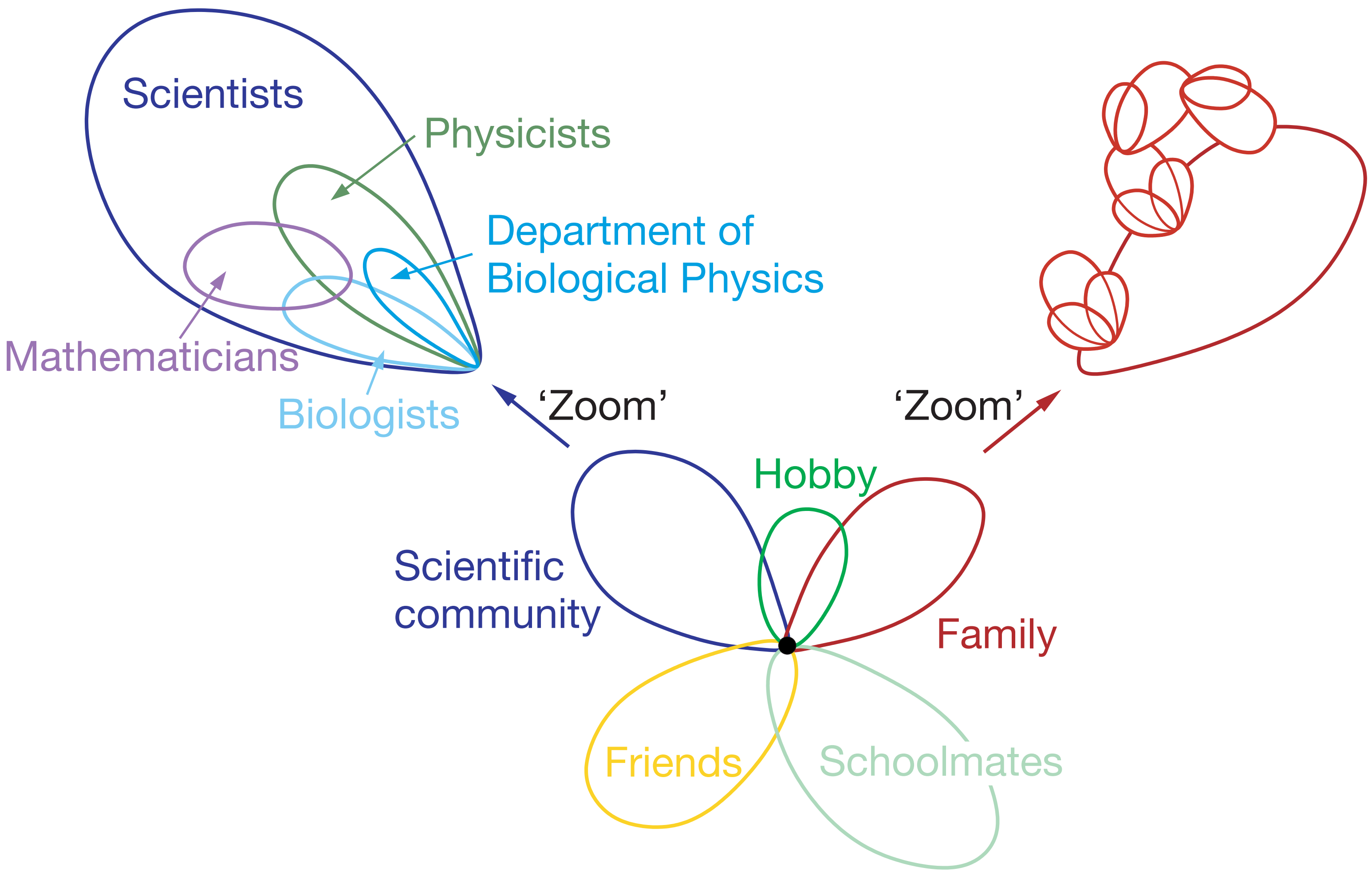
43S complex and protein metabolism
Hcr1
Sui1
Tif34
Prt1
Sua7
Tif5
Tif35
Rpg1
Ribosome biogenesis/assembly
Nop7
Rpf2
Nog2
Mak11
Tif6
Bud20
Nsa2
Ytm1
Puf6
Rlp7
Arx1
Dbp10
Has1
Nop15
Nug1
Ycr072c
Erb1
Sda1
Nop12
Mak5
Mak21
Nop4
Brx1
Nop2
Nop6
Clc1
Mdn1
Rrp12
CK2 complex and transcription regulation
Cka2
Cdc68
Ckb2
Ckb1
Abf1
Cka1
Arp4
Hta1
Htb1
Hht1
Hhf1
DNA packaging, chromatin assembly
Protein phosphatase type 2A complex (part)
Rts3
Tpd3
Pph21
Sir4
Sir3
Cdc55
Pph22
Chromatin silencing
Sir1
Zds2
Rrp14
Cell polarity, budding
Hst1
Sif2
Hos2
Cla4
Zds1
Glc1
Hos4
Snt1
Gic2
Cdc12
Bni5
Gin4
Bob1
Rga1
Shs1
Cph1
Set3
Set3c complex
Cdc42
Cdc11
Ste20
Bem1
Kcc4
Cdc3
Far1
Cdc10
Cytokinesis (septin ring)
Cdc24
Pheromone response (cellular fusion)

“Information”

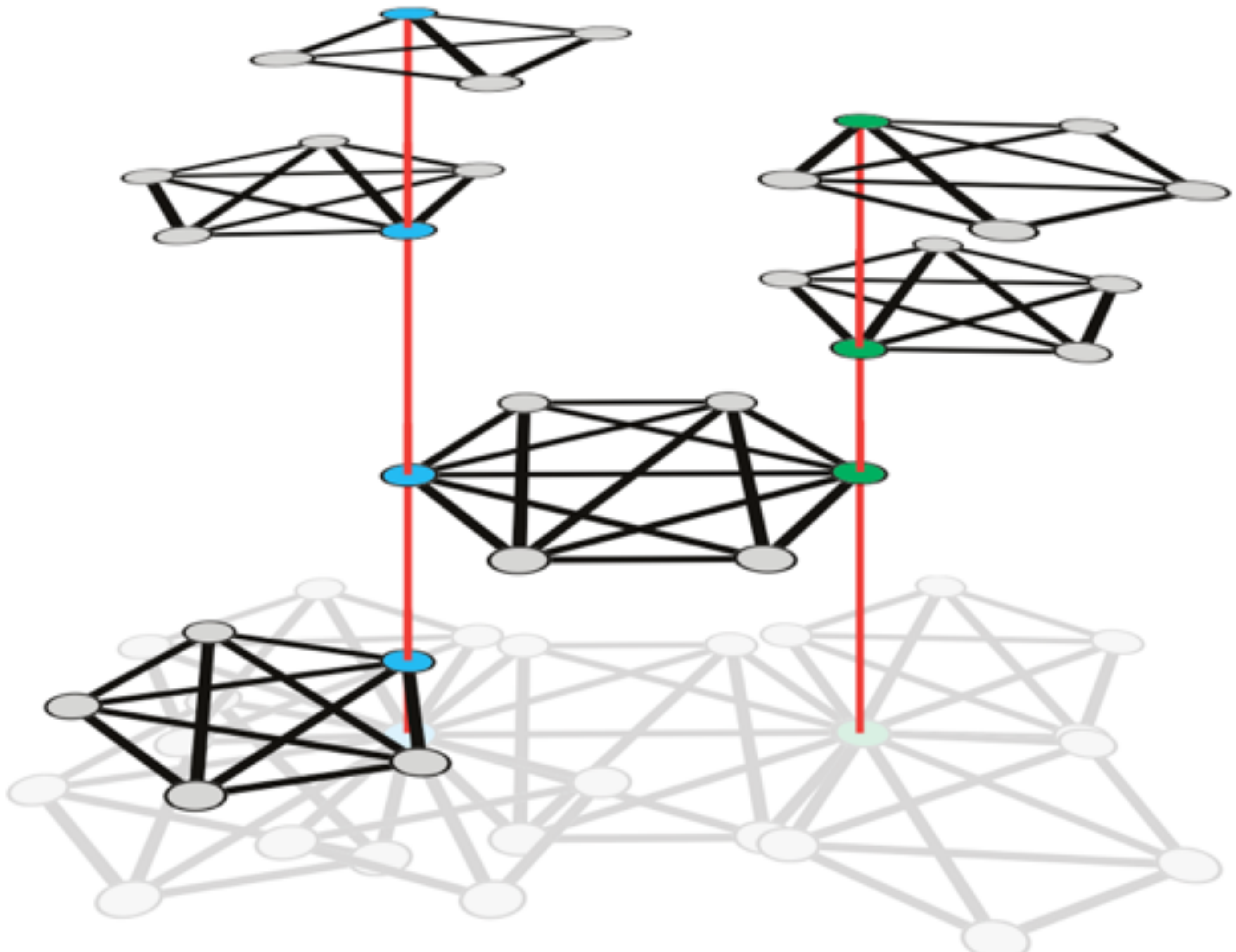
If there is a **random walker** on the network, it will be **trapped** inside each community.

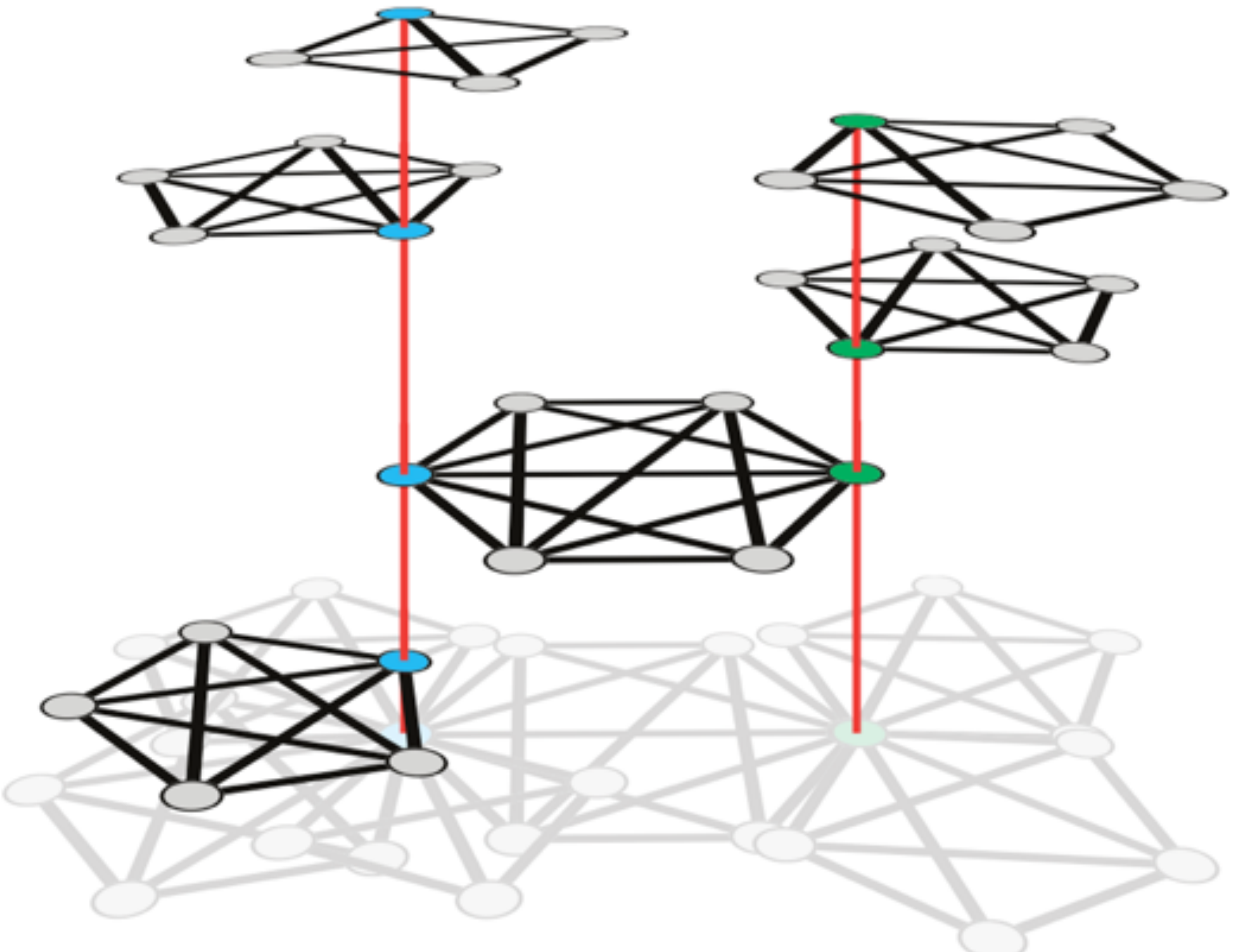
Demo

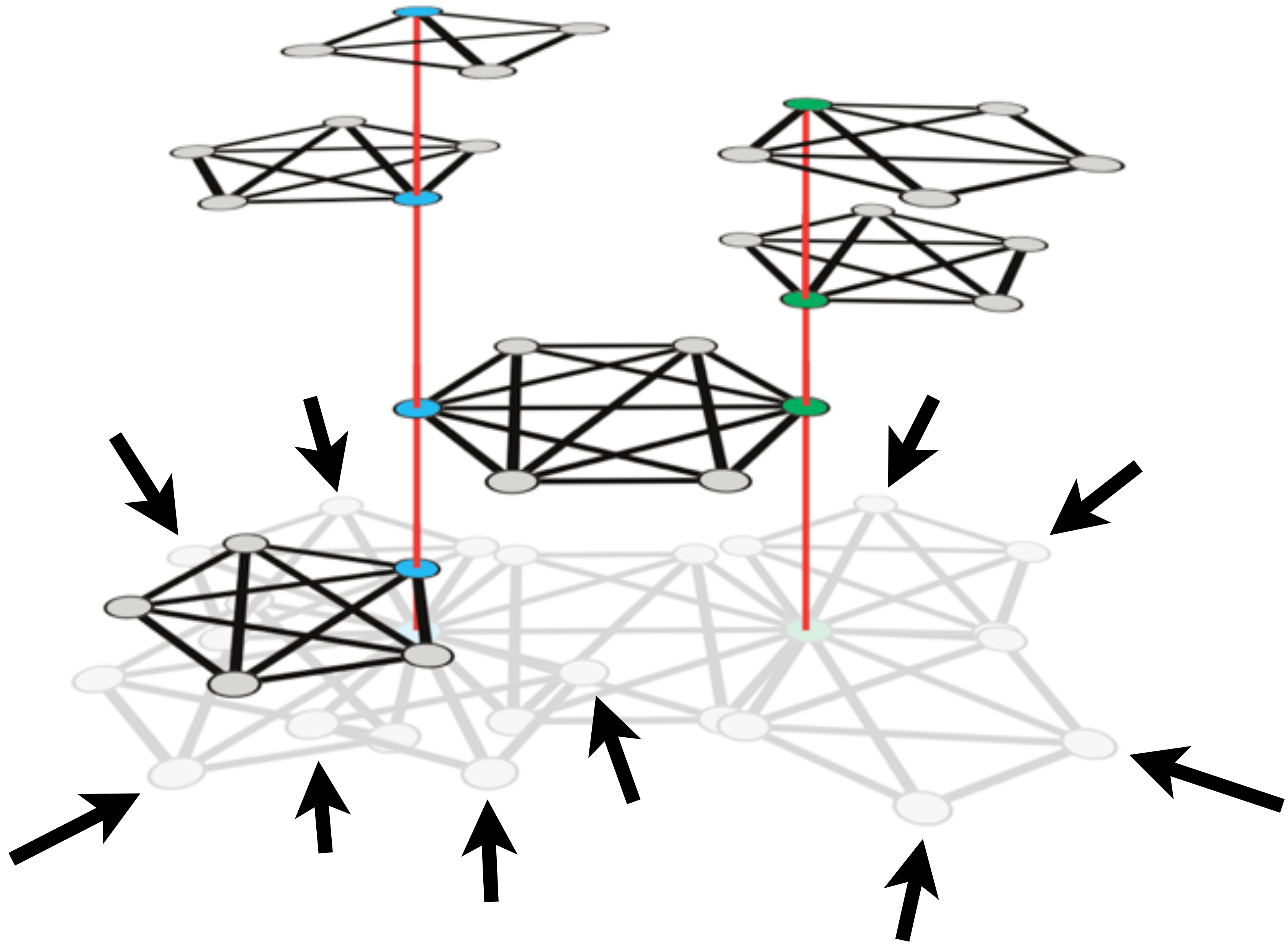
“Overlap”

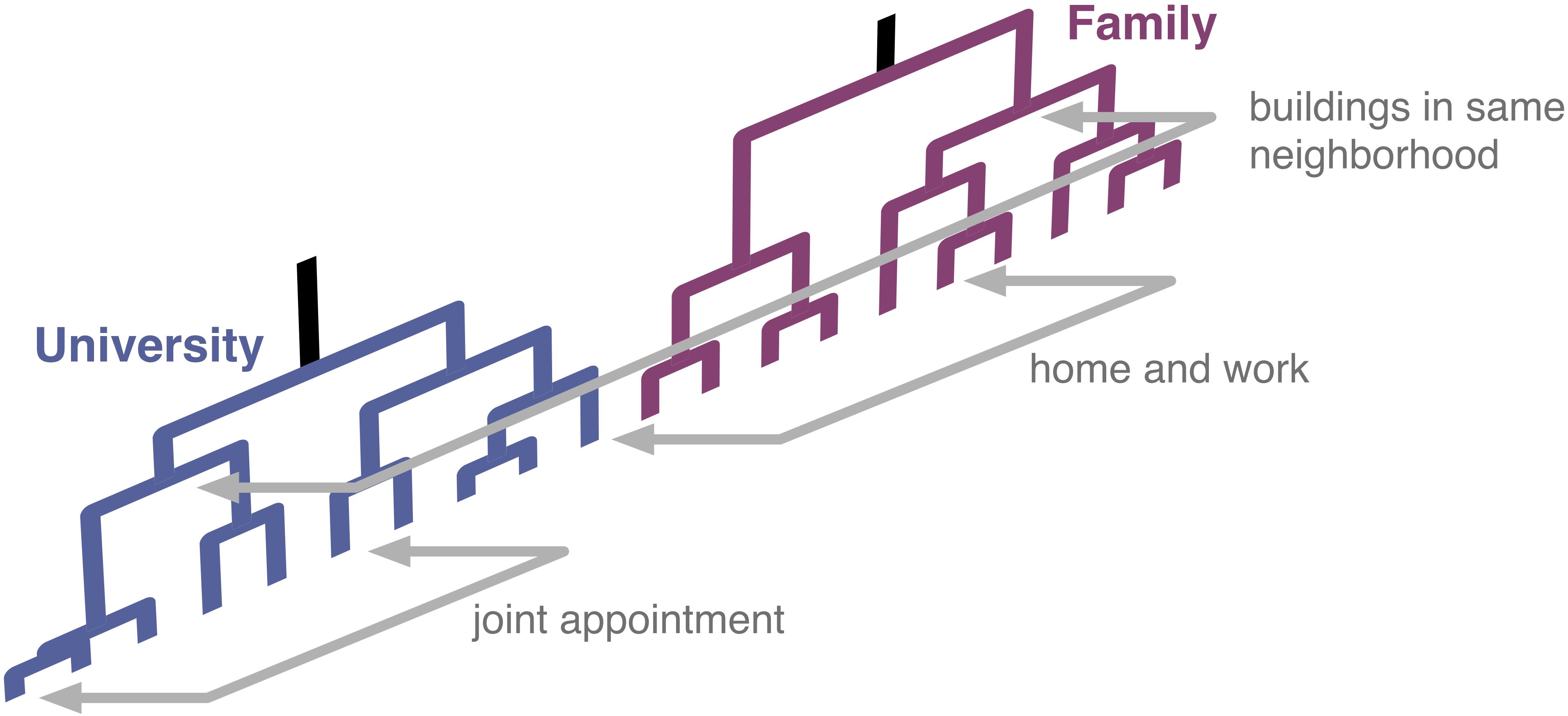


G. Palla, I. Derényi, I. Farkas & T. Vicsek, *Nature* (2005)



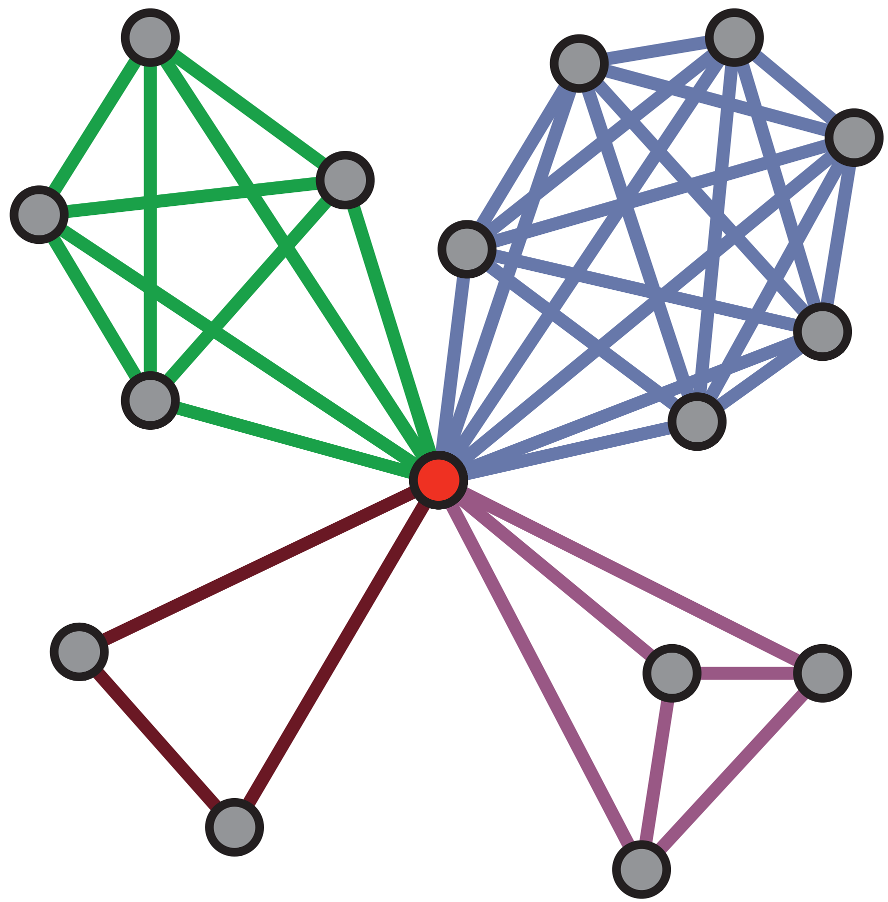




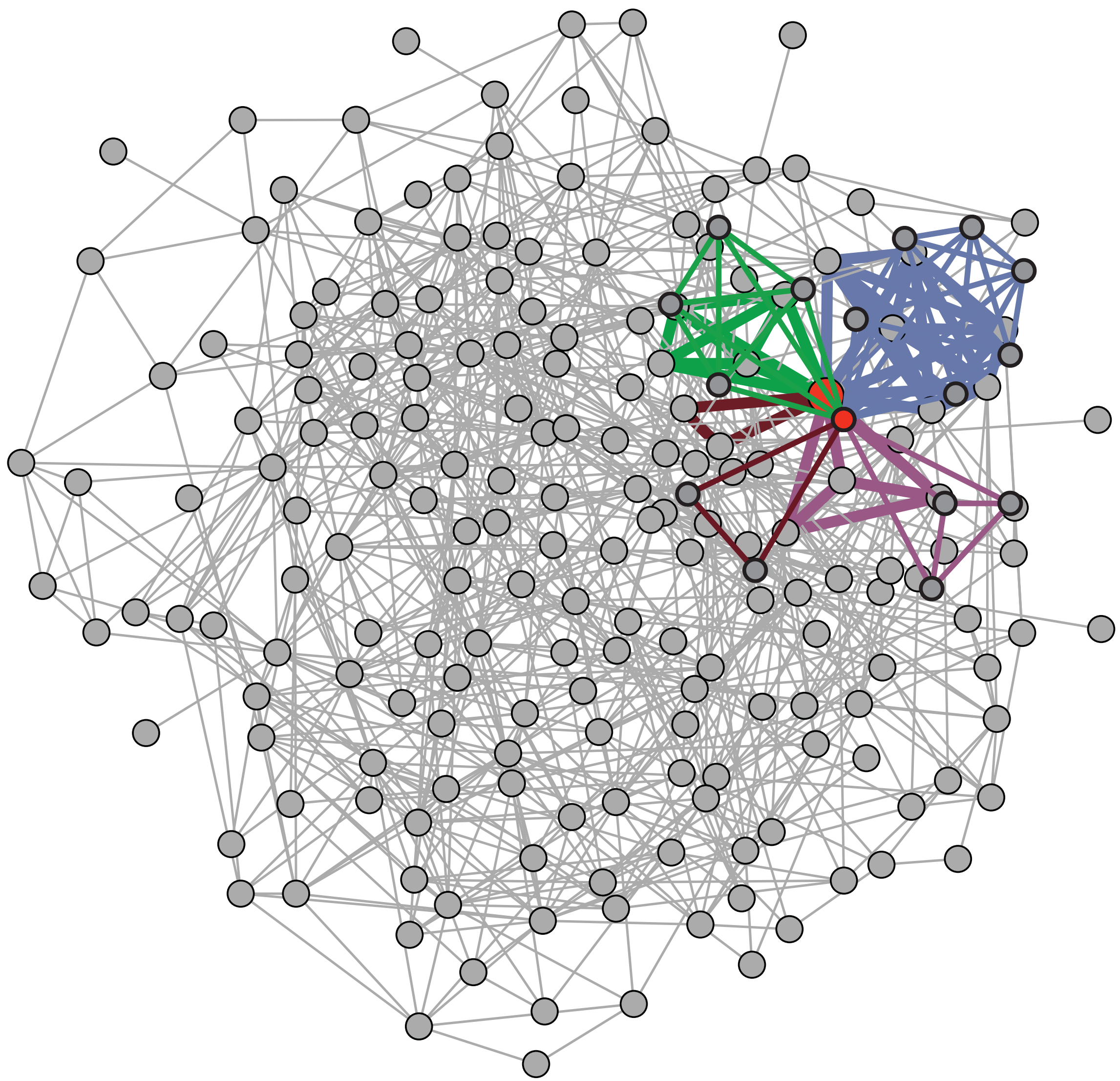


It is **impossible** to obtain a single dendrogram.

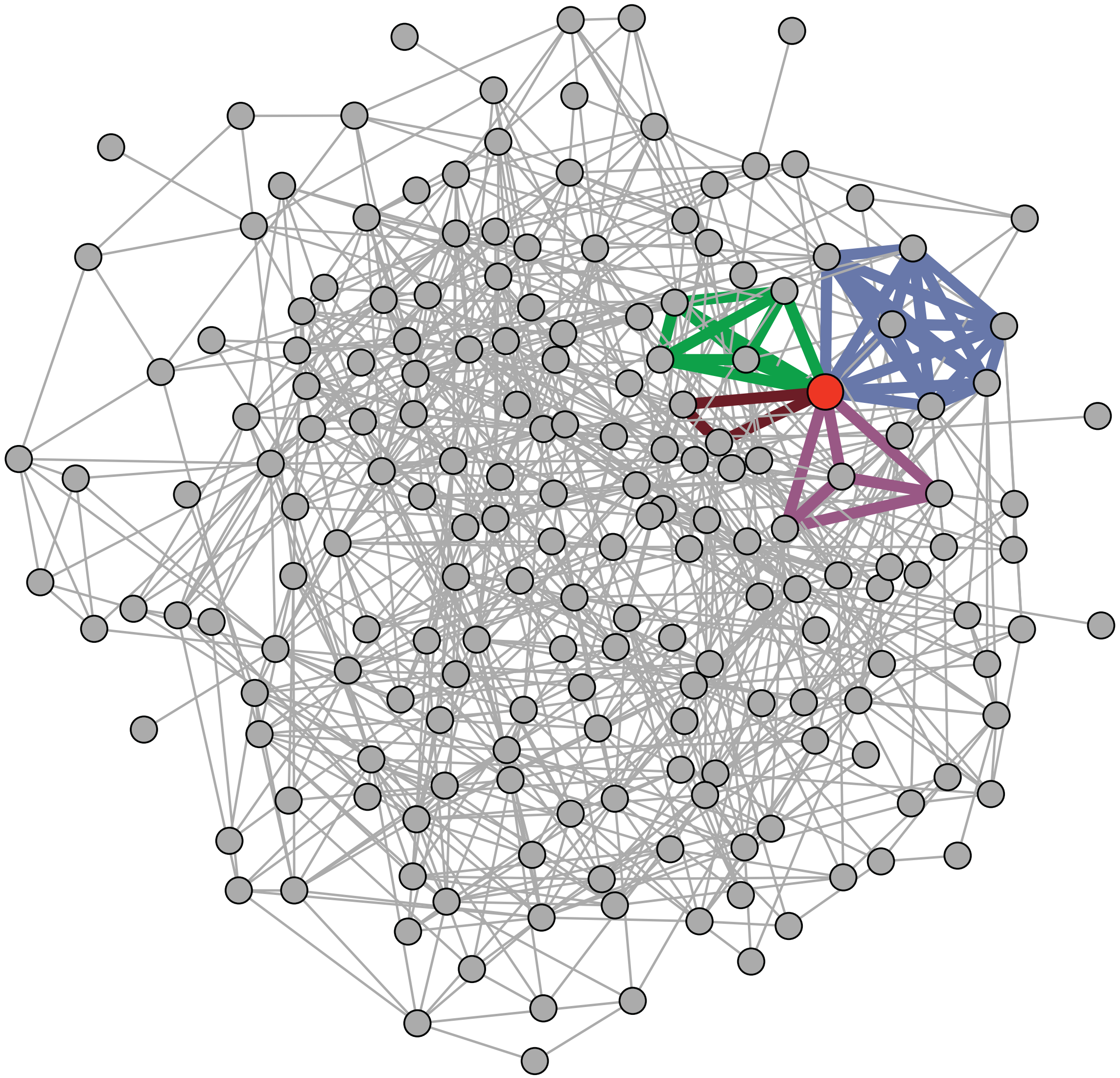
Simple local structure

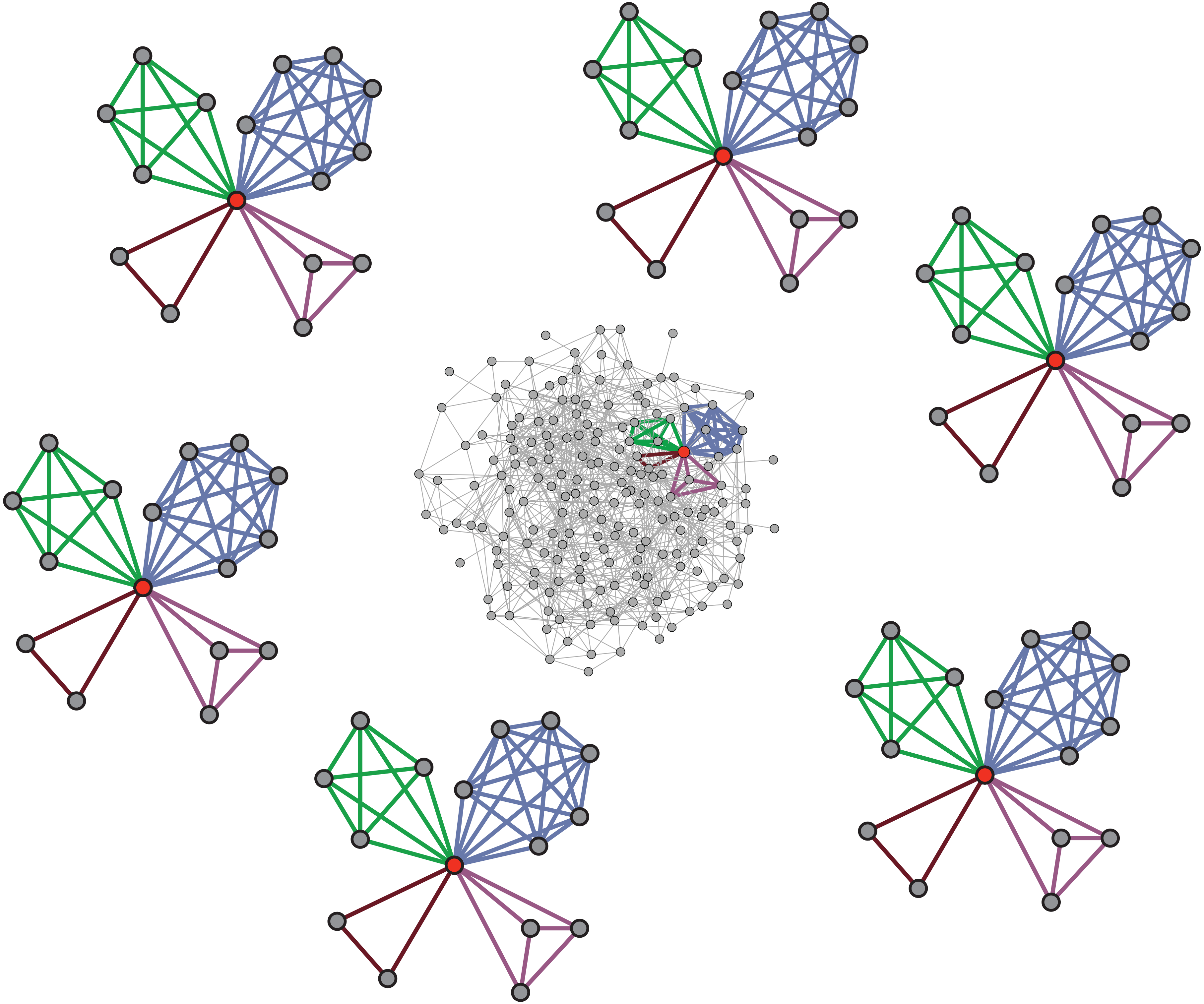


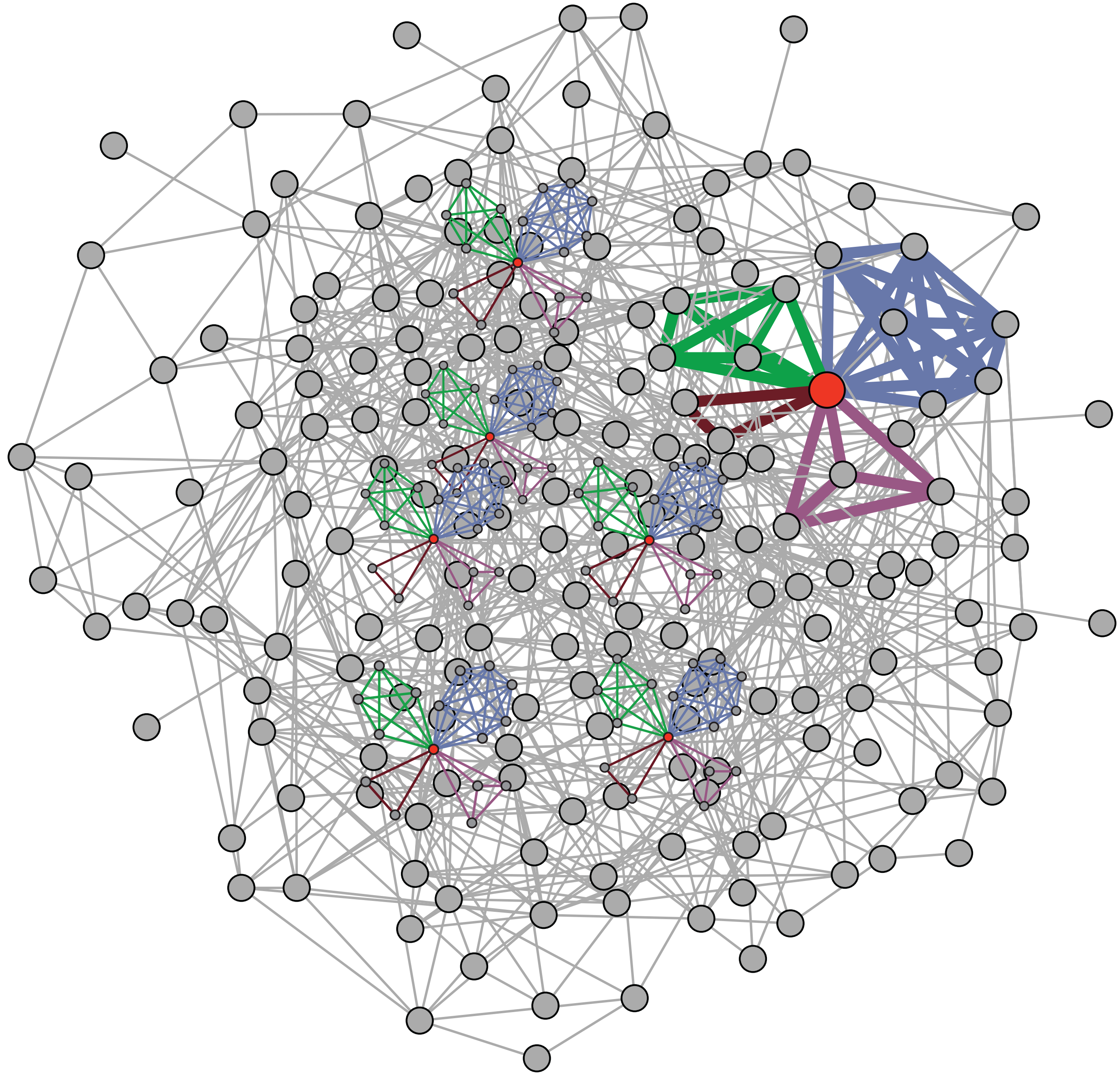
Complex global structure

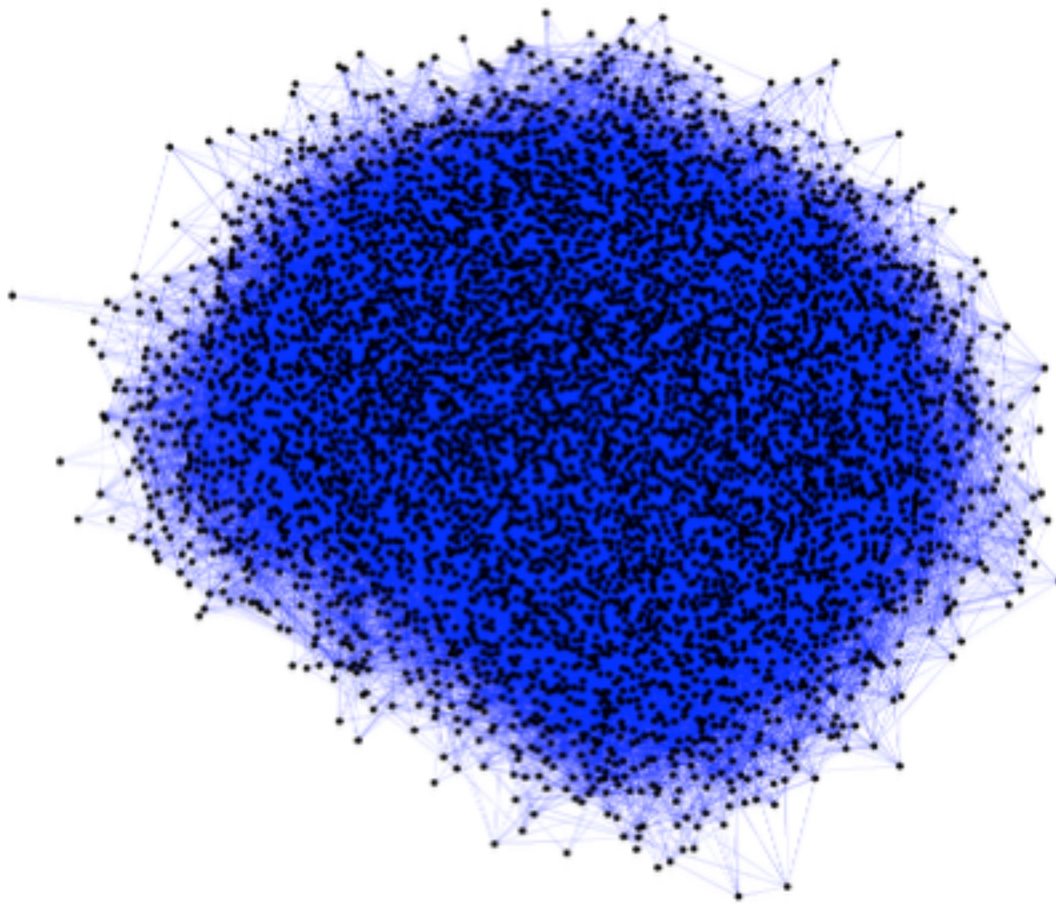


Complex global structure



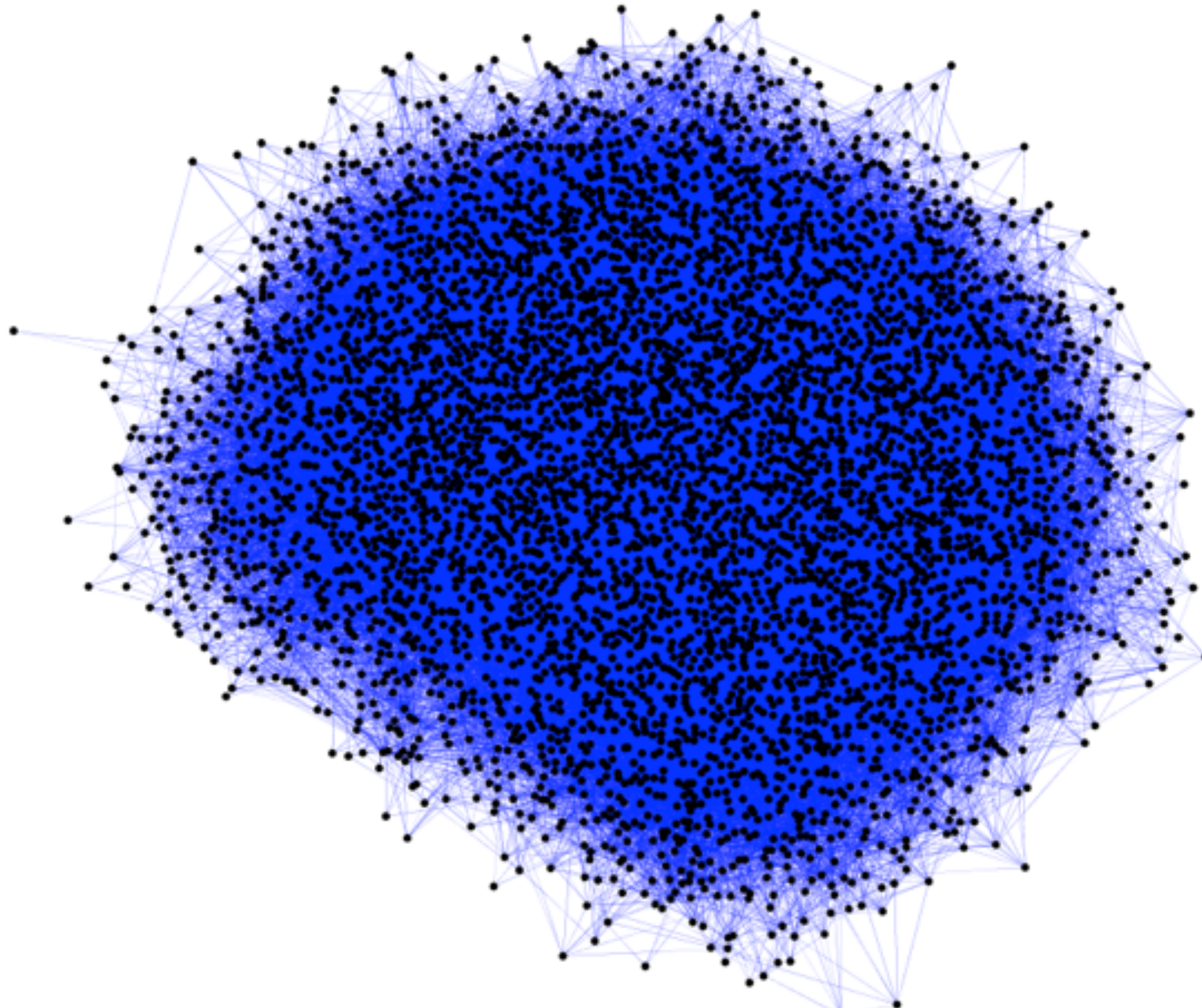




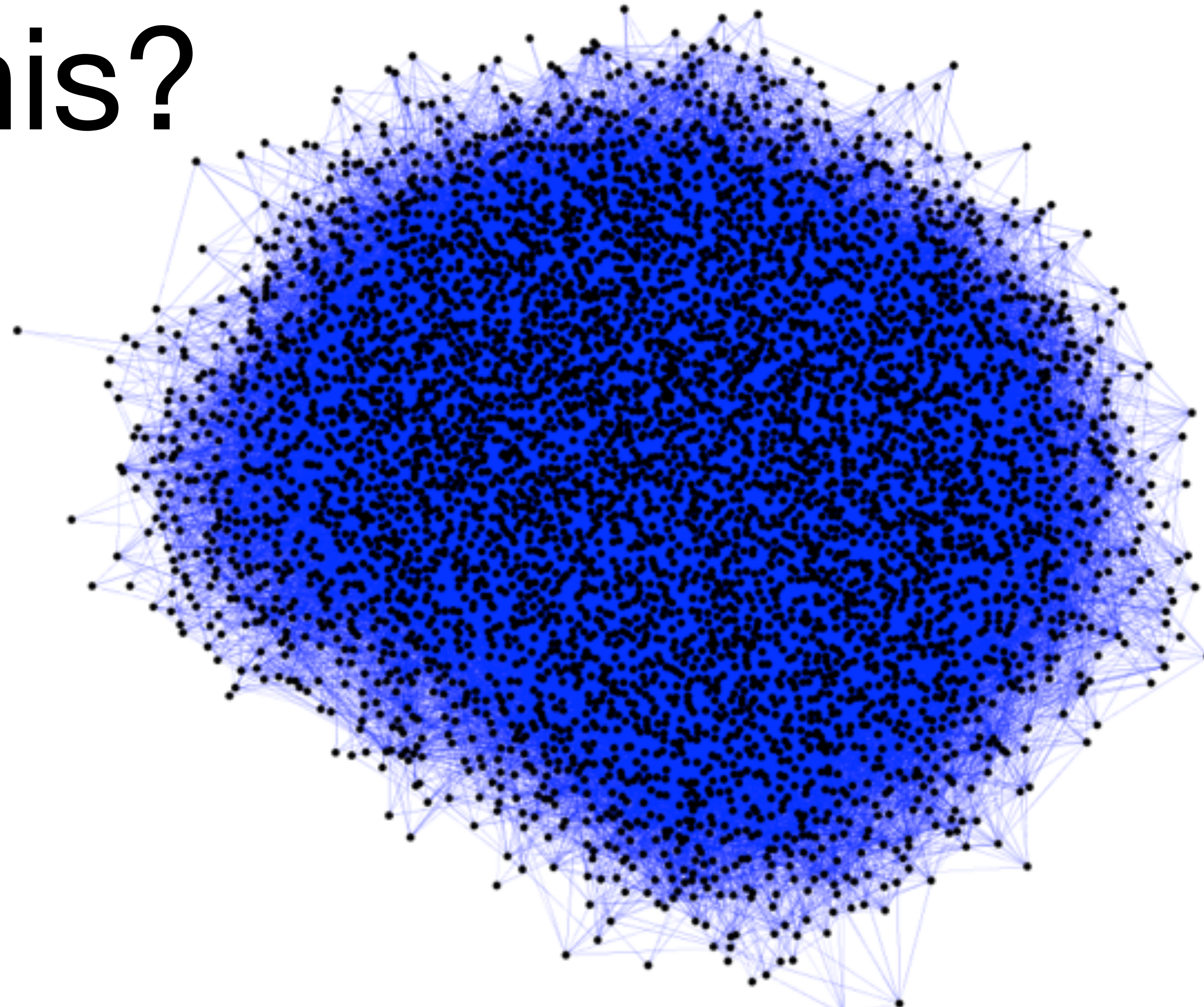


This is a modular network.

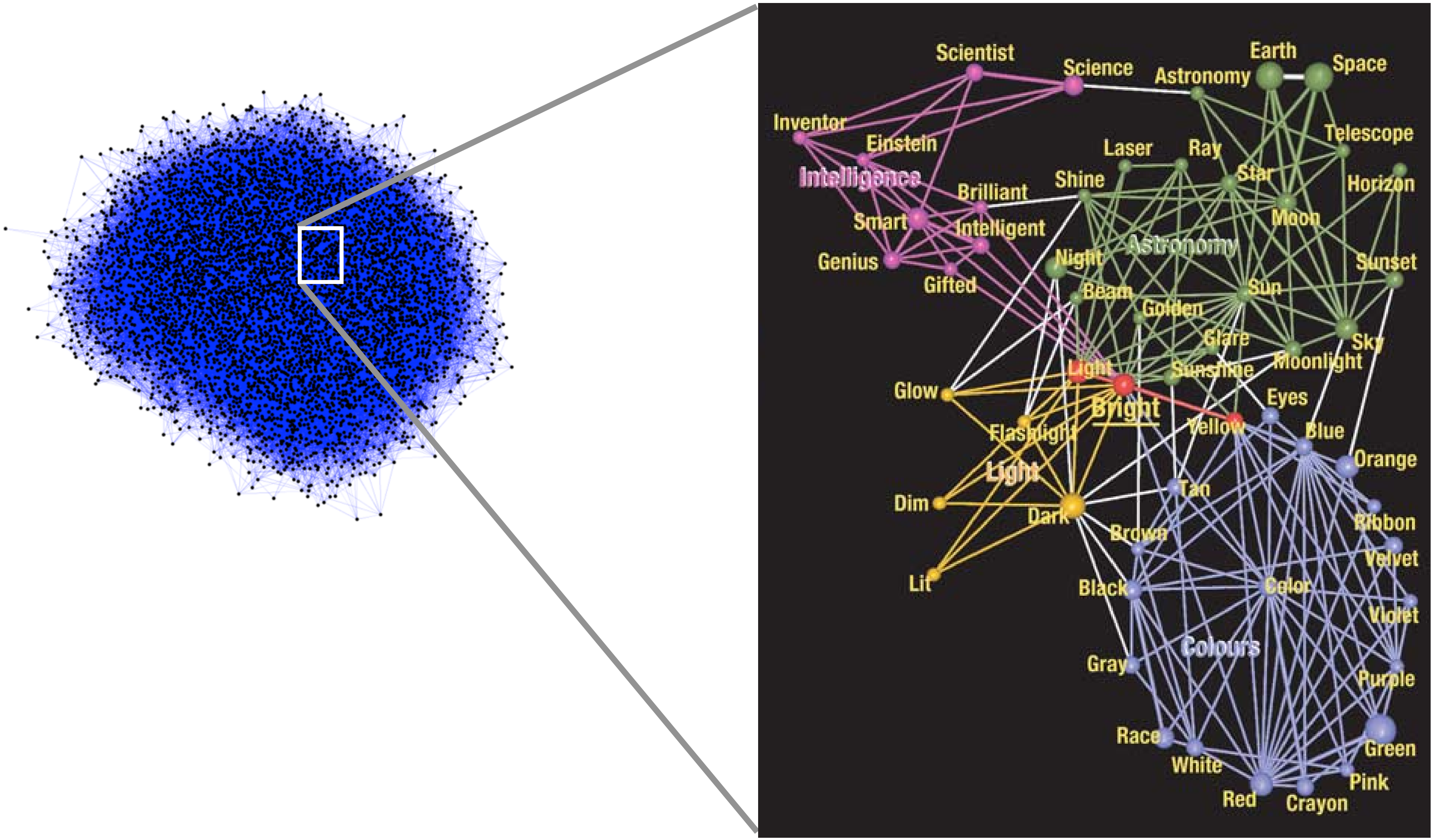
What is this?



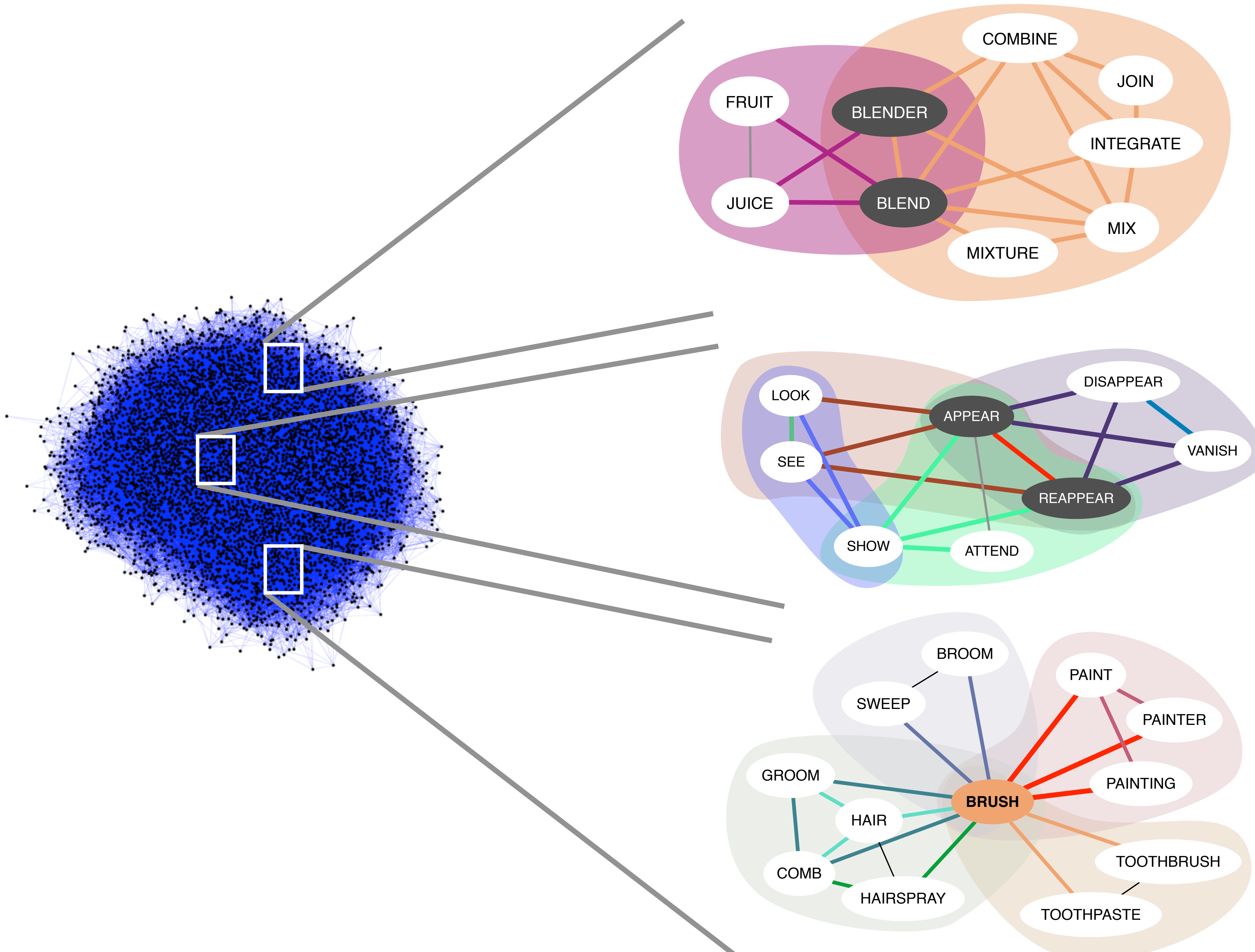
What the xxxx is this?



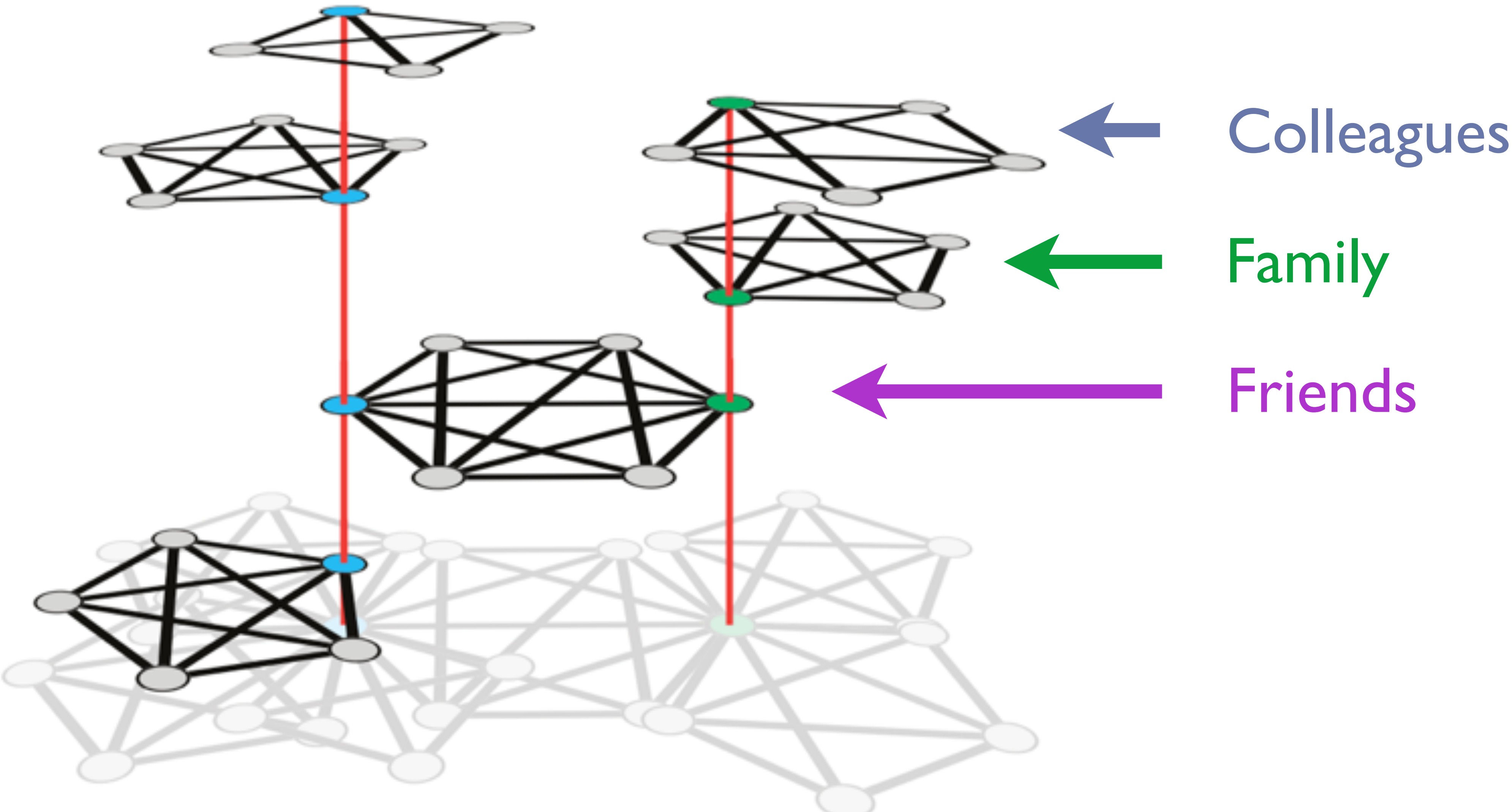
Word association network: Network of "commonly associated English words"

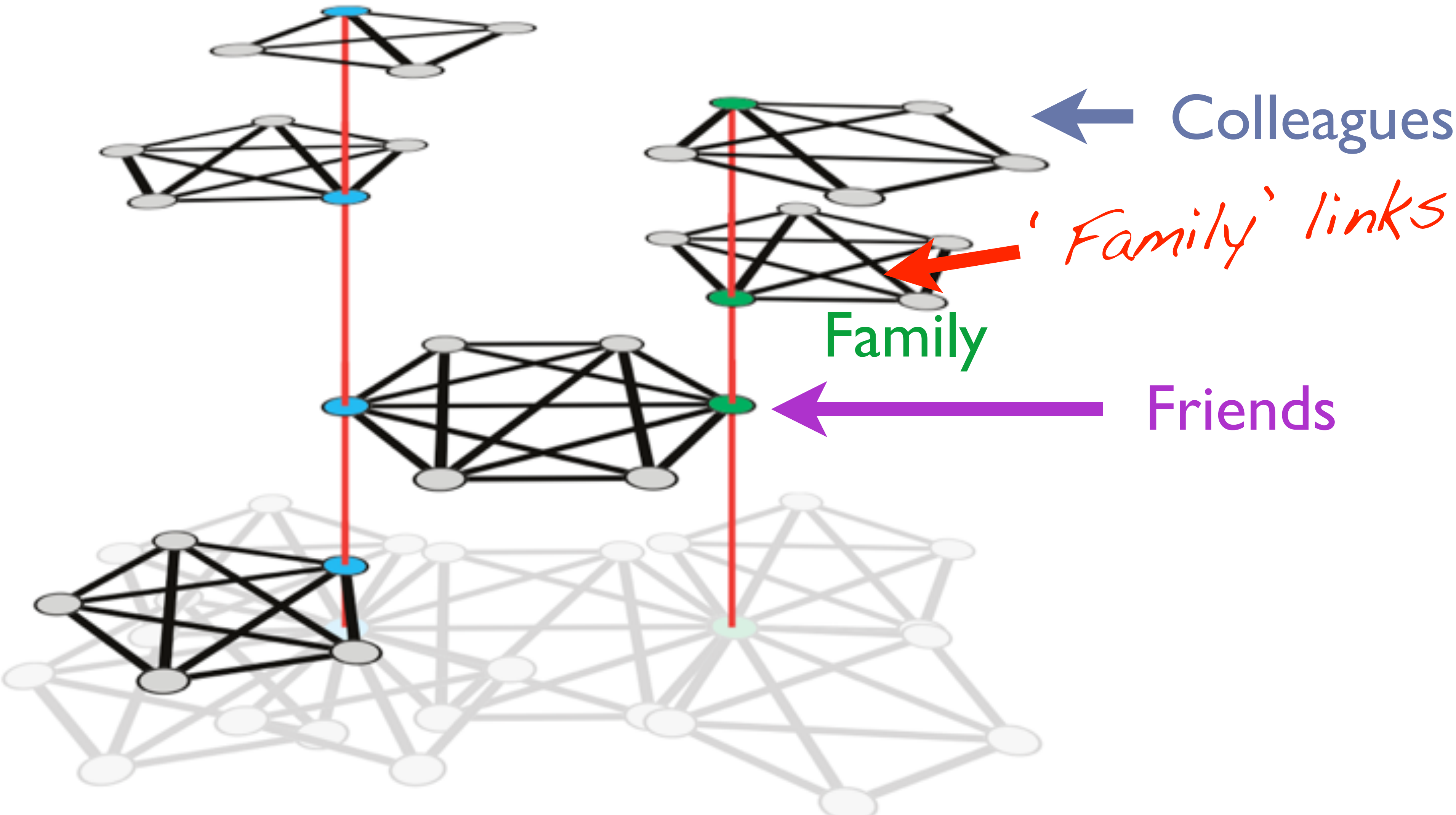


G. Palla, I. Derényi, I. Farkas & T. Vicsek, *Nature*, 2005

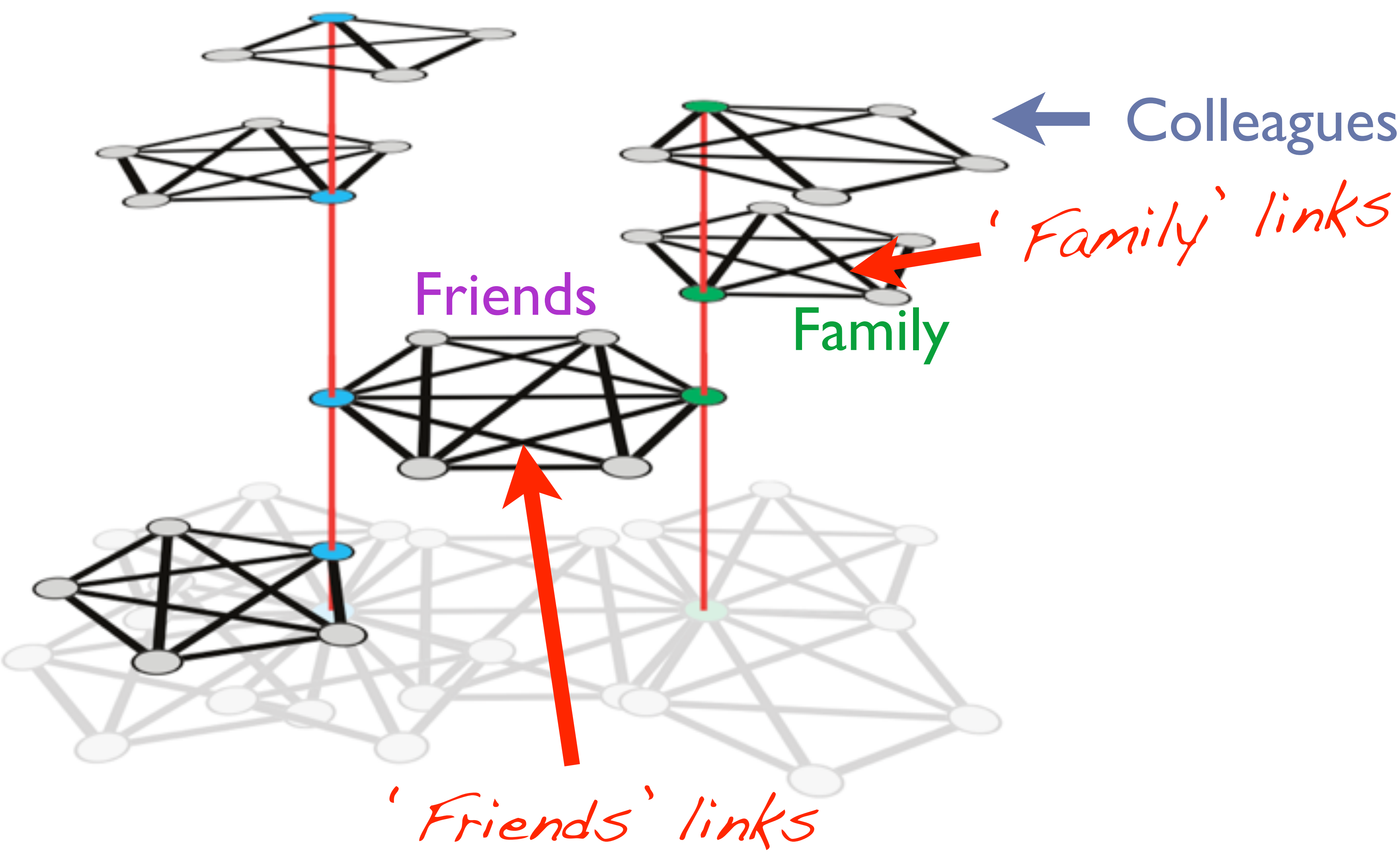

COMBINE
JOIN
FRUIT
BLENDER
INTEGRATE
JUICE
BLEND
MIX
MIXTURE
LOOK
DISAPPEAR
APPEAR
SEE
VANISH
REAPPEAR
SHOW
ATTEND
BROOM
PAINT
SWEEP
PAINTER
GROOM
PAINTING
BRUSH
HAIR
TOOTHBRUSH
COMB
HAIRSPRAY
TOOTHPASTE

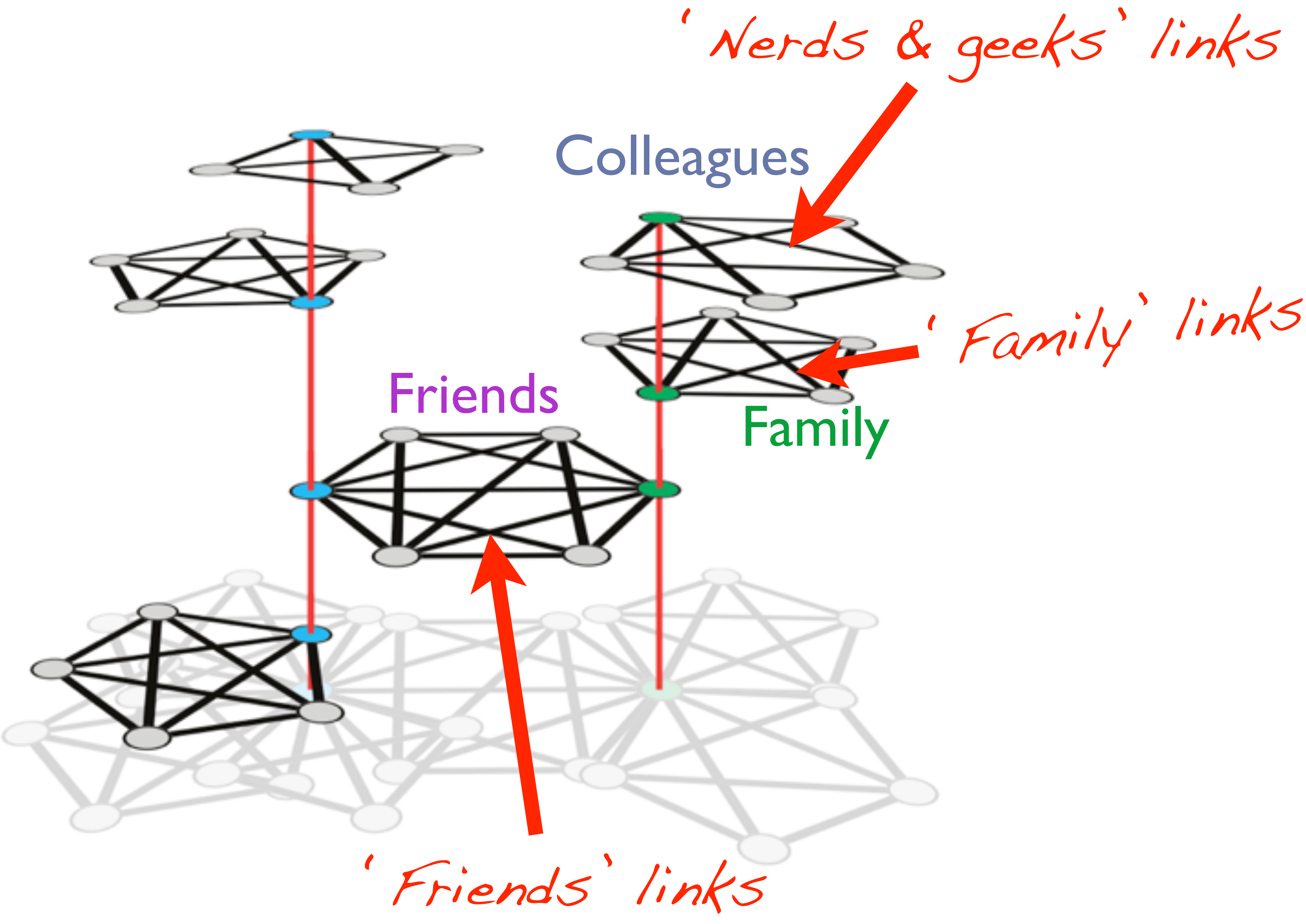
Link communities


Colleagues
Family
Friends



Colleagues
'Family' links
Family
Friends


Colleagues
'Family' links
Friends
Family
'Friends' links



'Nerds & geeks' links
Colleagues
'Family' links
Friends
Family
'Friends' links

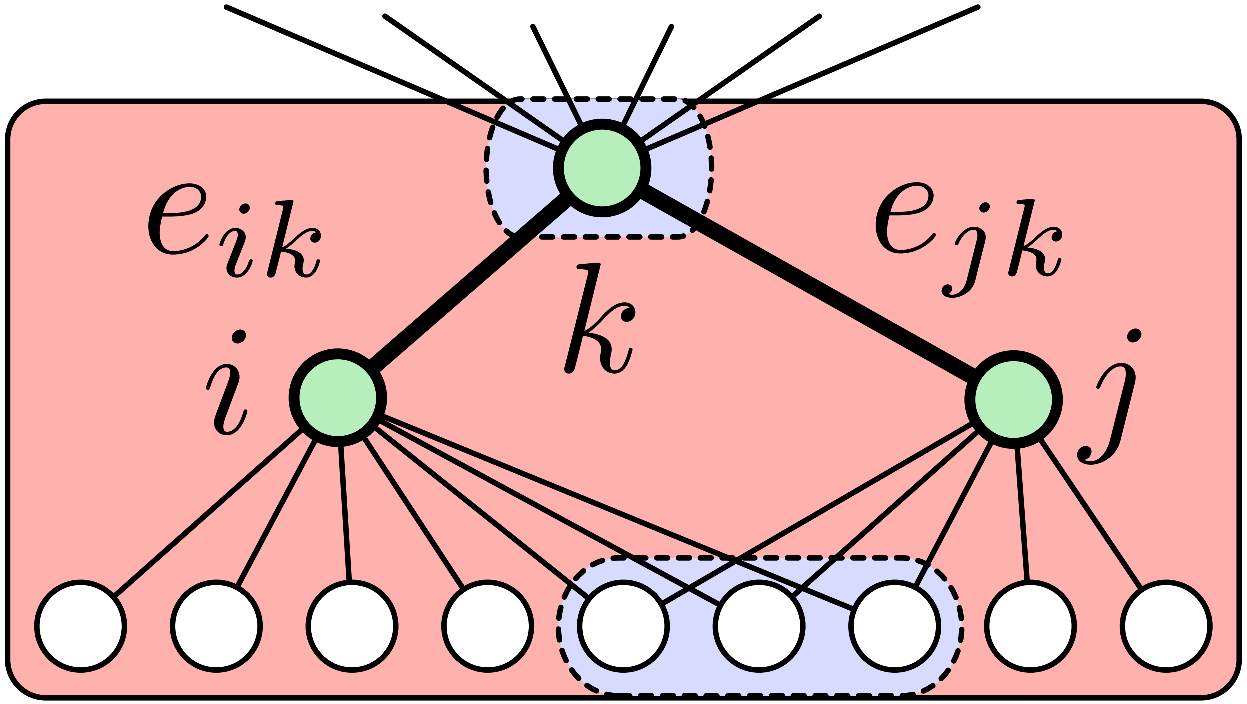
Nodes: multiple membership

Links: unique membership

Similarity between links

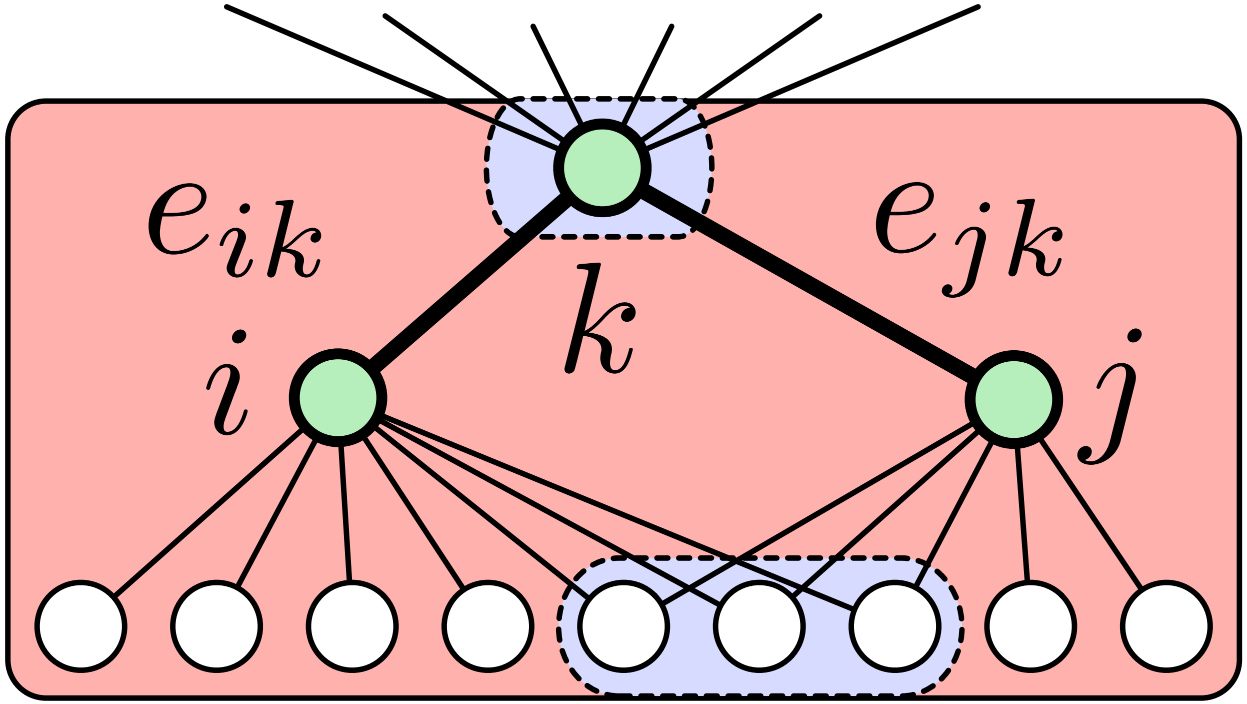
↓

Hierarchical Clustering



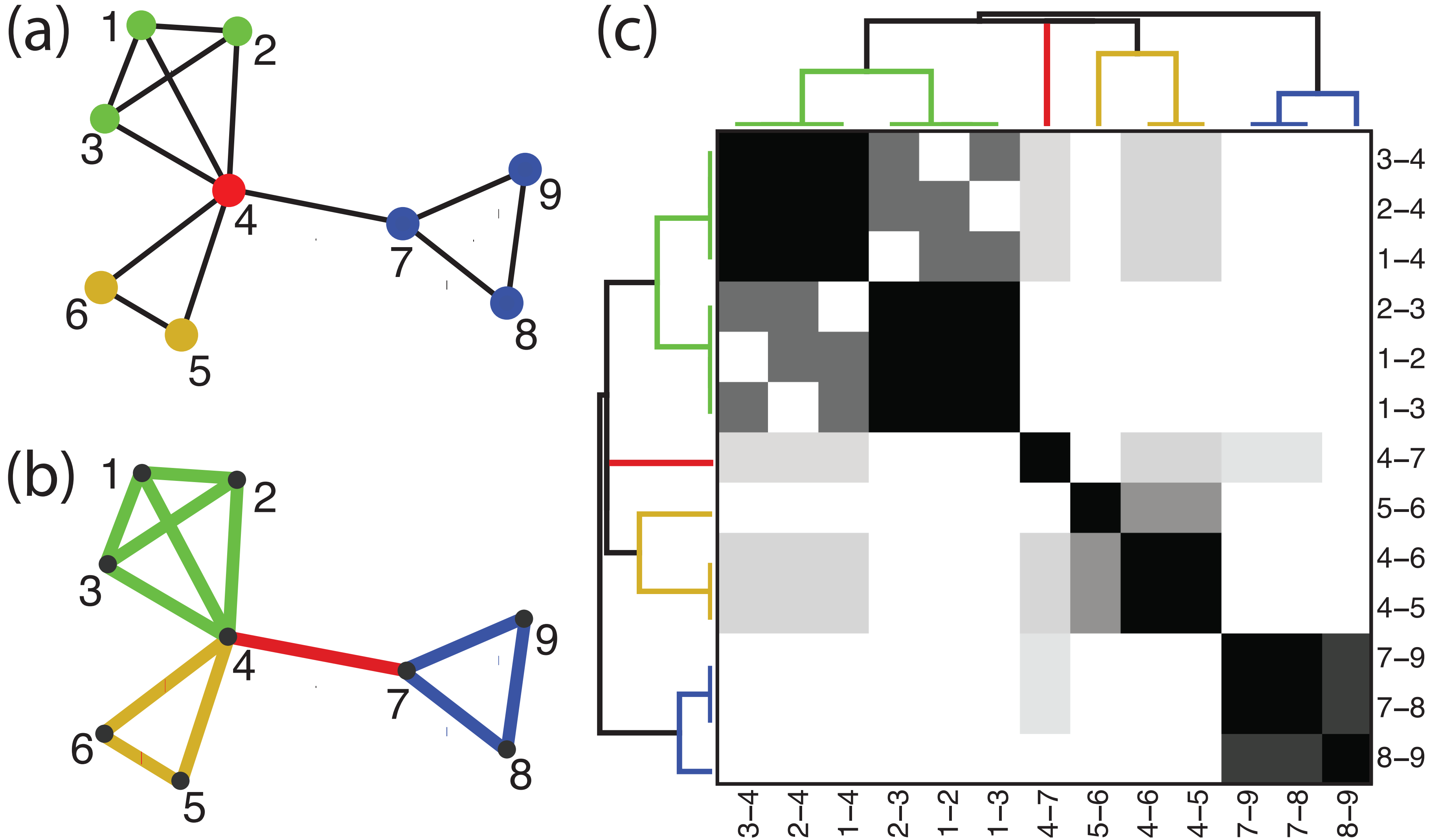
$$n_+(i) \equiv \{x \mid d(i,x) \leq 1\}$$

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

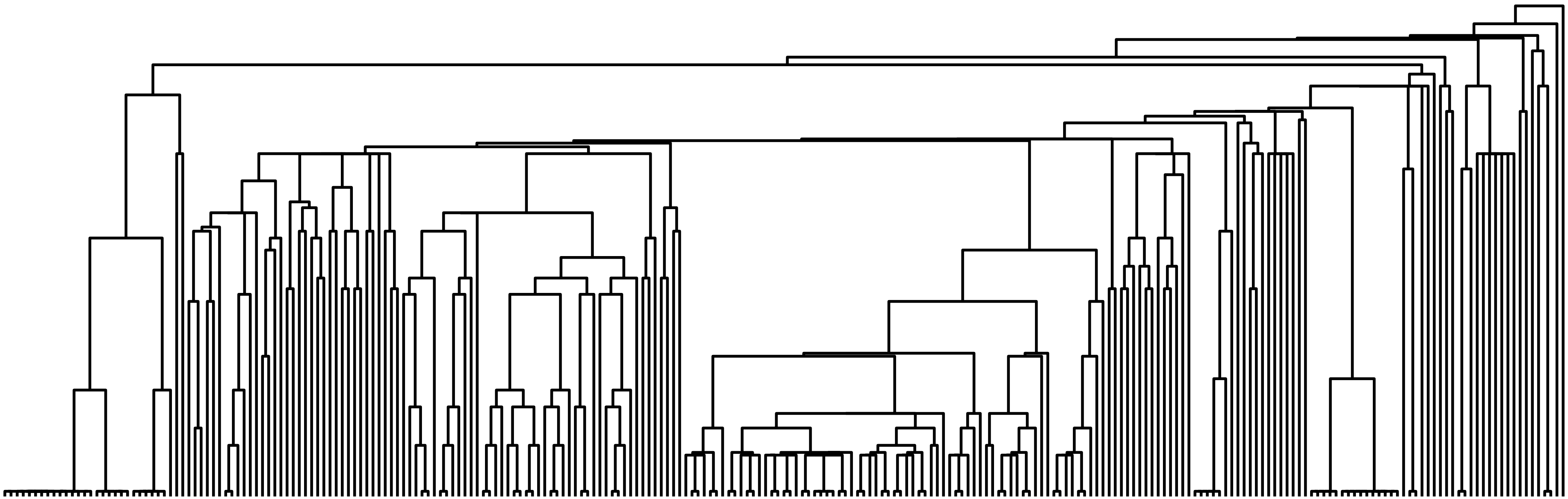


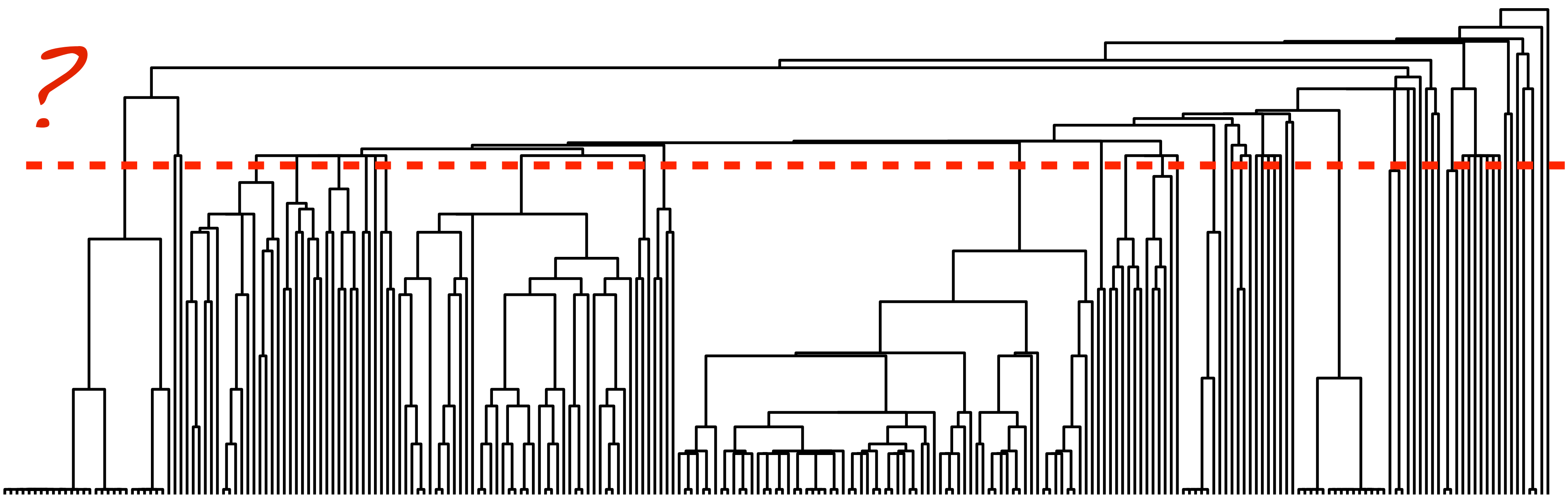
$$n_+(i) \equiv \{x \mid d(i,x) \leq 1\}$$

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \qquad \frac{4}{12}$$



(a)
1
2
3
4
5
6
7
8
9
(b)
1
2
3
4
5
6
7
8
9
(c)
3–4
2–4
1–4
2–3
1–2
1–3
4–7
5–6
4–6
4–5
7–9
7–8
8–9



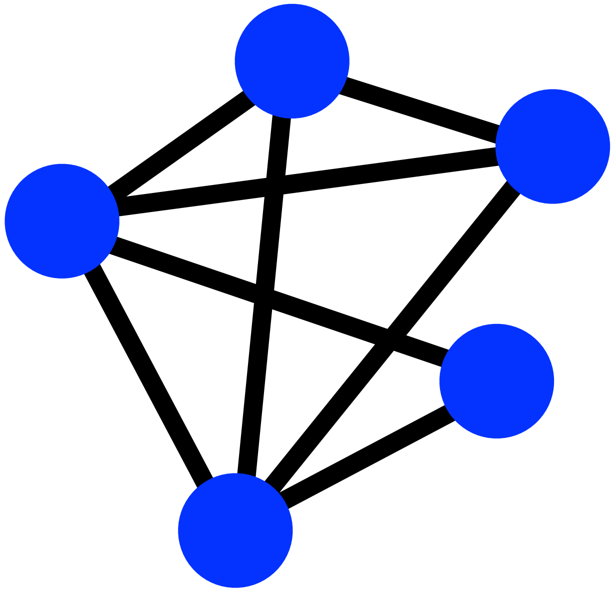


Partition Density

Community c has m_c edges and n_c induced nodes

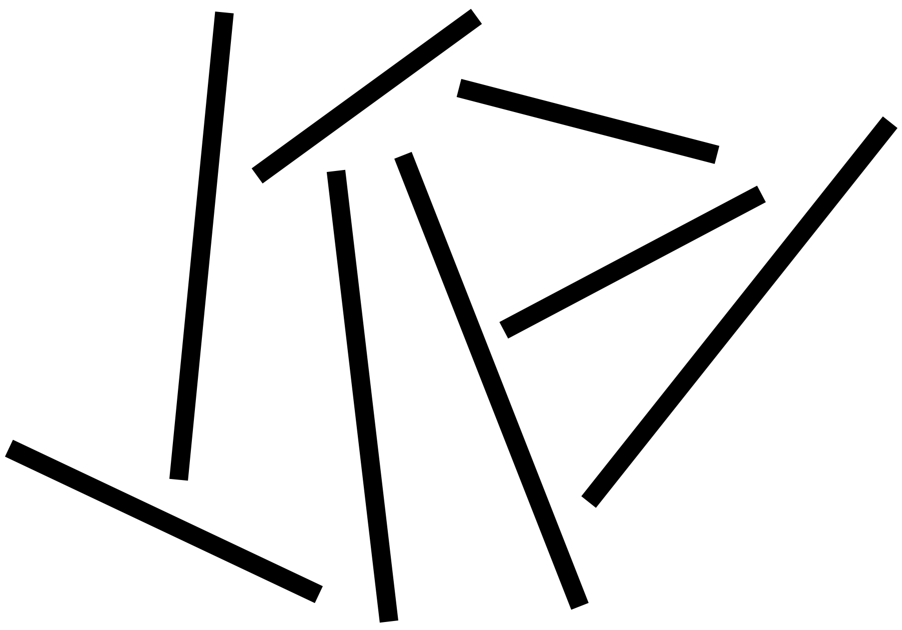
Partition Density

Community c has m_c edges and n_c induced nodes

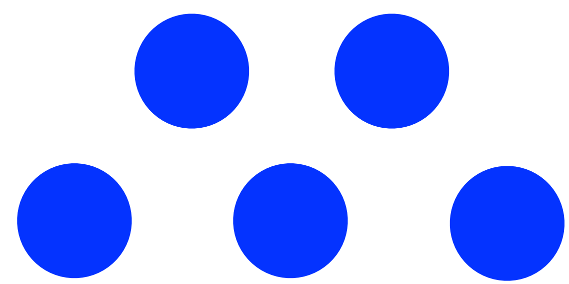


Partition Density

Community c has m_c edges and n_c induced nodes



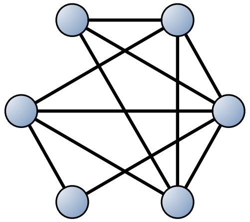
$$m_c = 8$$



$$n_c = 5$$

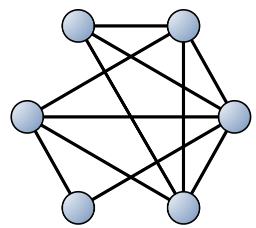
Partition Density

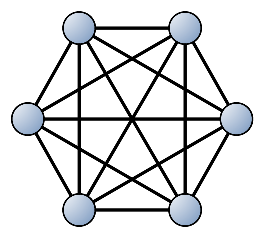
Community c has m_c edges and n_c induced nodes

 $= m_c$

Partition Density

Community c has m_c edges and n_c induced nodes

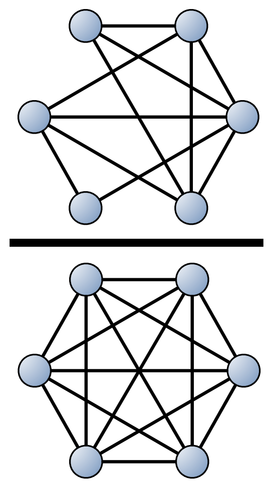




$$= \frac{m_c}{\frac{n_c(n_c-1)}{2}}$$

Partition Density

Community c has m_c edges and n_c induced nodes

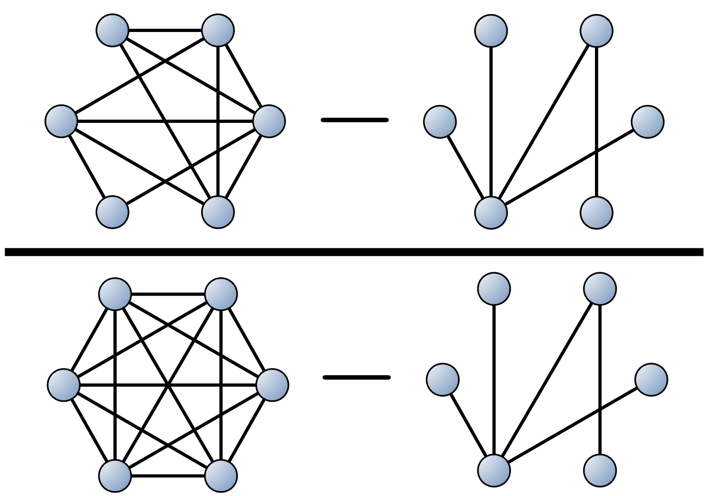


$$= \frac{m_c}{\frac{n_c(n_c-1)}{2}}$$

A **single** link is maximally dense

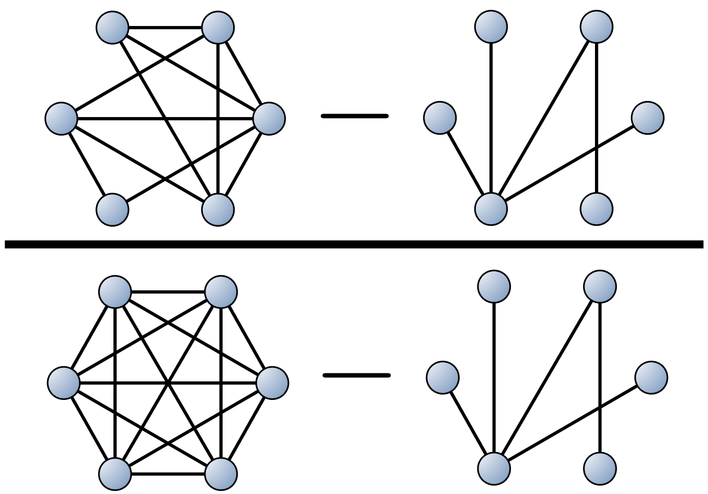
Partition Density

Community c has m_c edges and n_c induced nodes



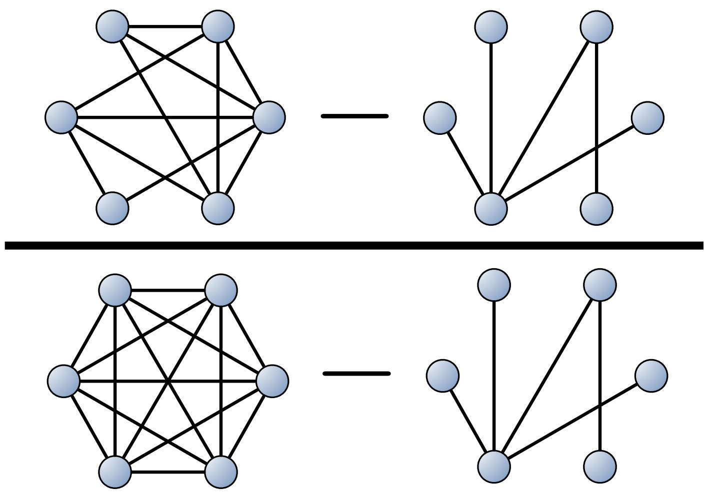
$$= \frac{m_c - (n_c - 1)}{\frac{n_c(n_c-1)}{2} - (n_c - 1)}$$

Partition Density



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

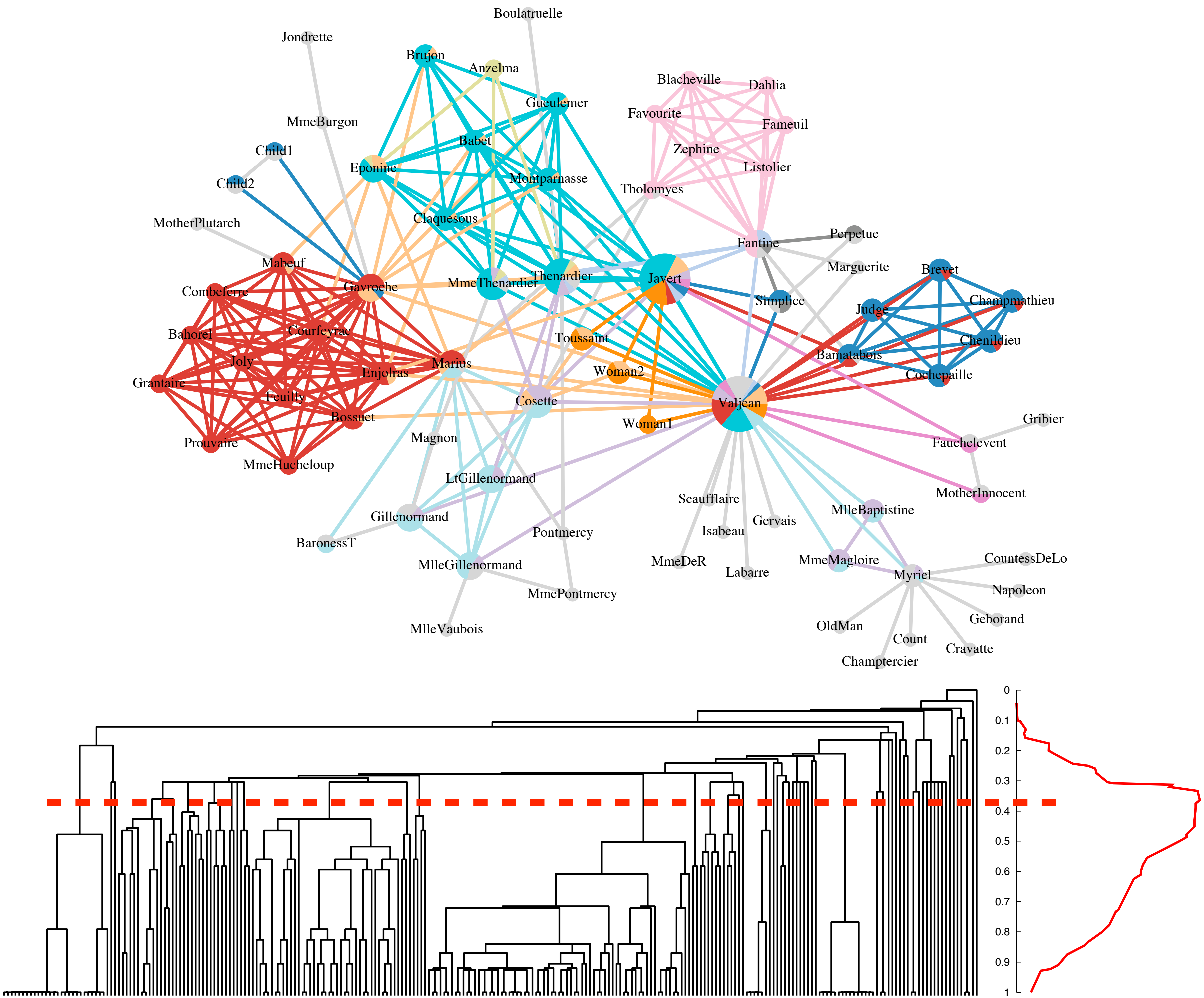
Partition Density



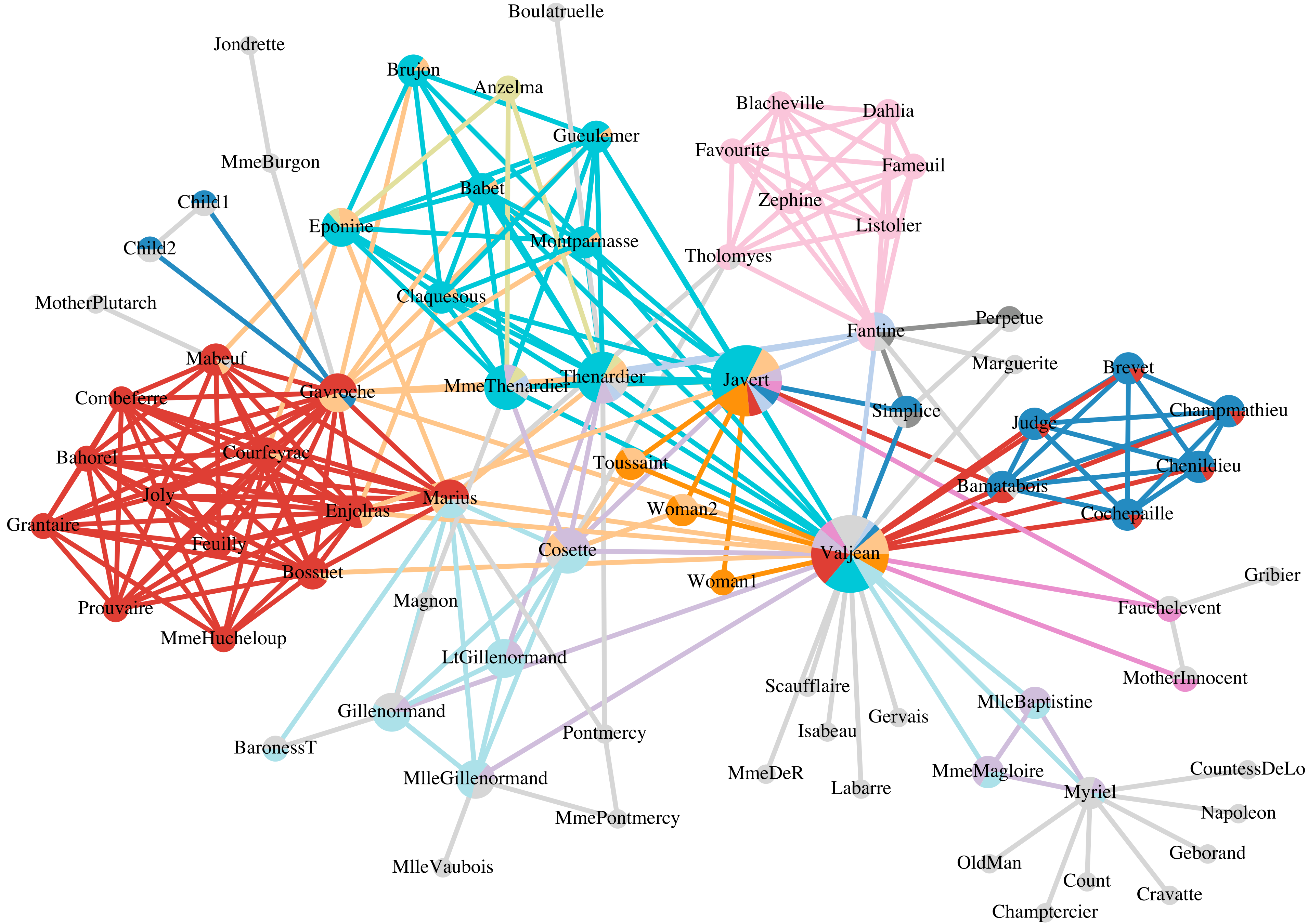
$$= \frac{m_c - (n_c - 1)}{\frac{n_c(n_c-1)}{2} - (n_c - 1)}$$

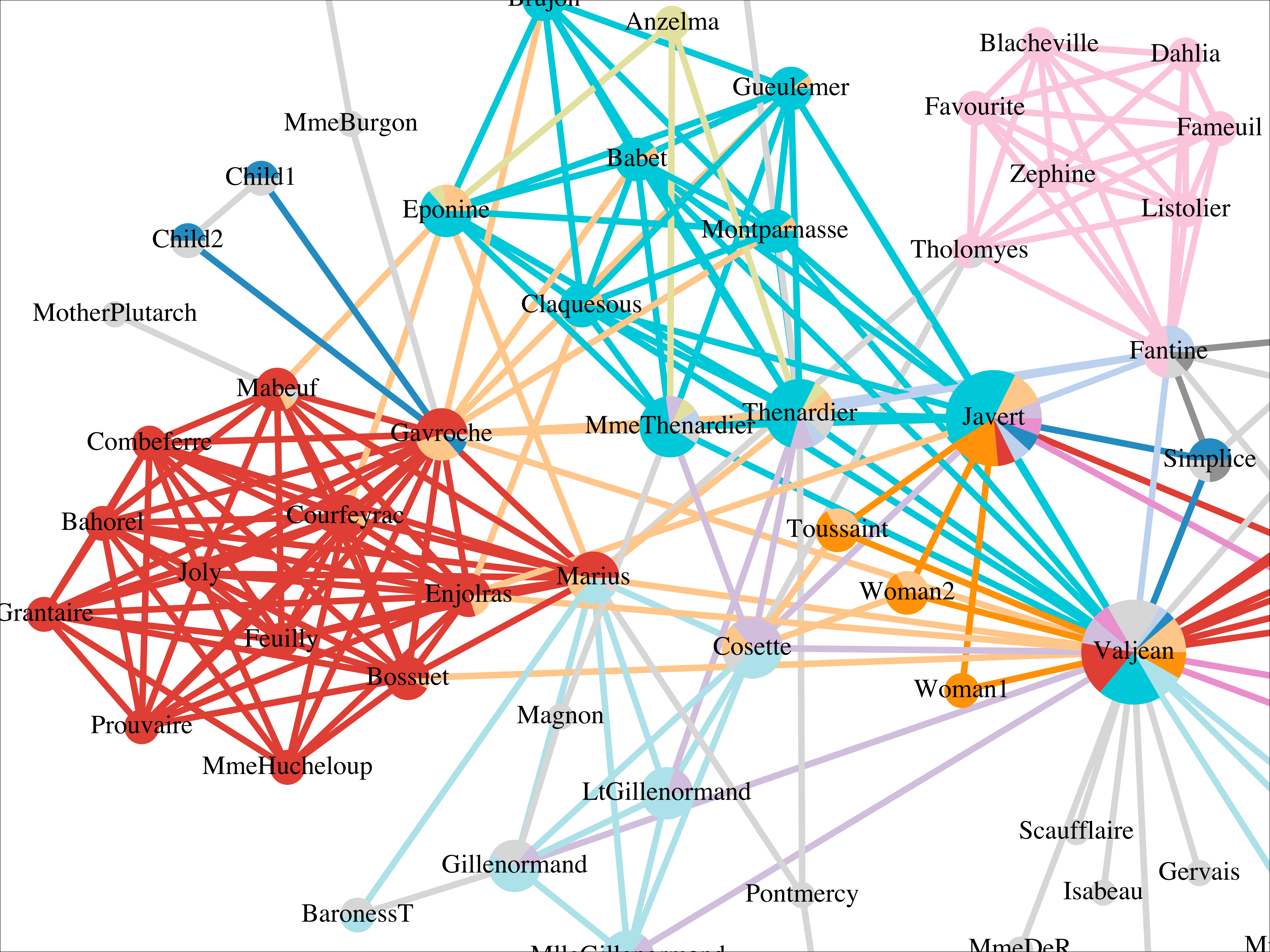
$$= 2\frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}$$

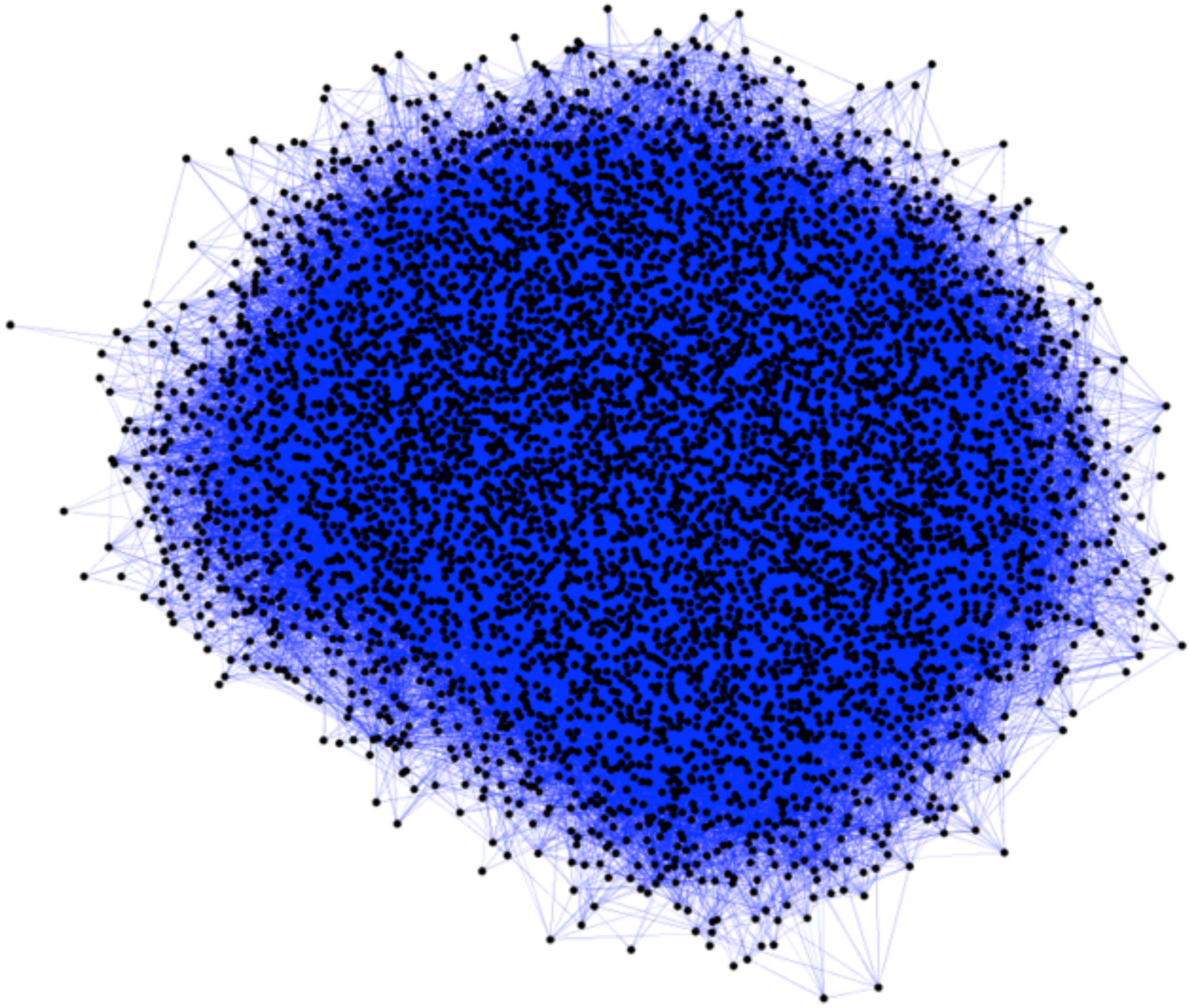
$$D \equiv \frac{2}{M}\sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}$$

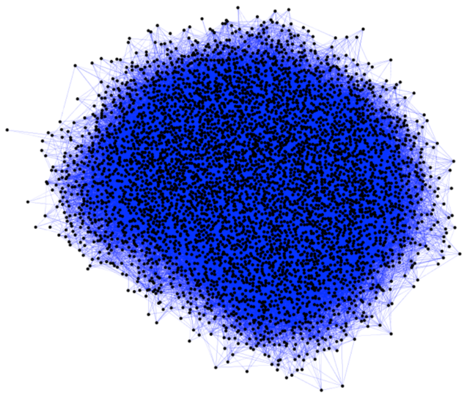


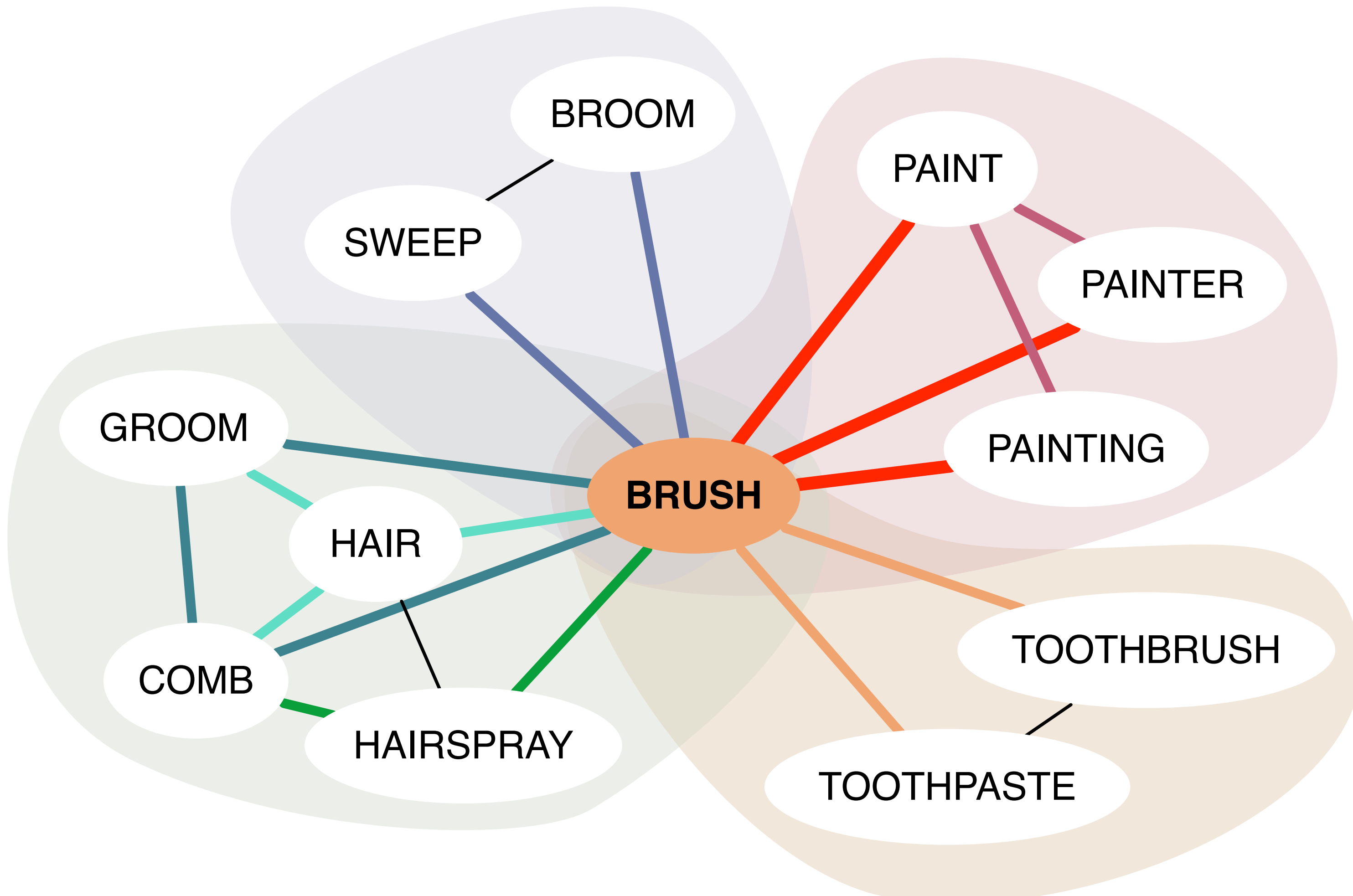
Boulatruelle
Jondrette
Brujon
Anzelma
Gueulemer
Blacheville
Dahlia
Favourite
Fameuil
Zephine
Listolier
MmeBurgon
Child1
Babet
Eponine
Montparnasse
Tholomyes
Child2
Claquesous
MotherPlutarch
Fantine
Perpetue
Mabeuf
Marguerite
Thenardier
Javert
Brevet
Combeferre
Gavroche
MmeThenardier
Simplice
Champmathieu
Judge
Bahorel
Courfeyrac
Toussaint
Chenildieu
Joly
Bamatabois
Marius
Enjolras
Woman2
Cochepaille
Grantaire
Feuilly
Cosette
Valjean
Bossuet
Woman1
Gribier
Magnon
Fauchelevent
Prouvaire
MmeHucheloup
LtGillenormand
MotherInnocent
Scaufflaire
Gervais
MlleBaptistine
Gillenormand
Isabeau
Pontmercy
BaronessT
MmeDeR
MmeMagloire
CountessDeLo
MlleGillenormand
Labarre
Myriel
Napoleon
MmePontmercy
Geborand
OldMan
Count
MlleVaubois
Cravatte
Champtercier
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

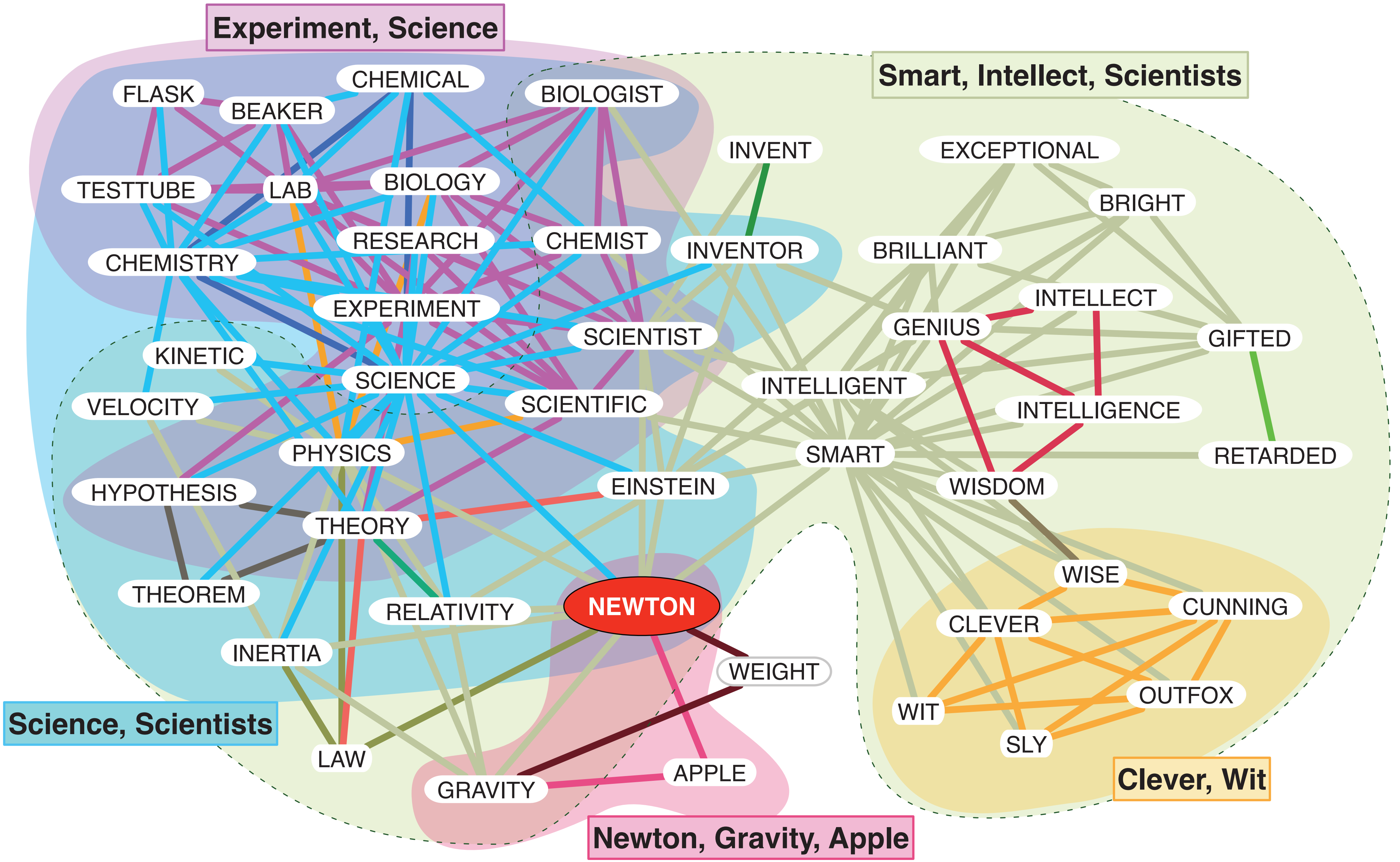

Boulatruelle
Jondrette
Brujon
Anzelma
Blacheville
Dahlia
Gueulemer
Favourite
Fameuil
MmeBurgon
Babet
Zephine
Child1
Listolier
Eponine
Montparnasse
Child2
Tholomyes
Claquesous
MotherPlutarch
Perpetue
Fantine
Mabeuf
Marguerite
Brevet
Thenardier
Javert
Gavroche
MmeThenardier
Combeferre
Simplice
Champmathieu
Judge
Courfeyrac
Toussaint
Chenildieu
Bahorel
Bamatabois
Joly
Marius
Woman2
Cochepaille
Grantaire
Enjolras
Feuilly
Cosette
Valjean
Gribier
Bossuet
Woman1
Magnon
Fauchelevent
Prouvaire
MmeHucheloup
LtGillenormand
MotherInnocent
Scaufflaire
MlleBaptistine
Gillenormand
Gervais
Isabeau
Pontmercy
BaronessT
CountessDeLo
MmeDeR
MmeMagloire
Labarre
MlleGillenormand
Myriel
Napoleon
MmePontmercy
Geborand
OldMan
Count
Cravatte
MlleVaubois
Champtercier


Anzelma
Blacheville
Dahlia
Gueulemer
Favourite
Fameuil
MmeBurgon
Babet
Zephine
Child1
Listolier
Eponine
Montparnasse
Child2
Tholomyes
Claquesous
MotherPlutarch
Fantine
Mabeuf
Thenardier
Javert
MmeThenardier
Gavroche
Combeferre
Simplice
Courfeyrac
Toussaint
Bahorel
Joly
Marius
Woman2
Enjolras
Grantaire
Feuilly
Cosette
Valjean
Bossuet
Woman1
Magnon
Prouvaire
MmeHucheloup
LtGillenormand
Scaufflaire
Gillenormand
Gervais
Isabeau
Pontmercy
BaronessT
MmeDeR

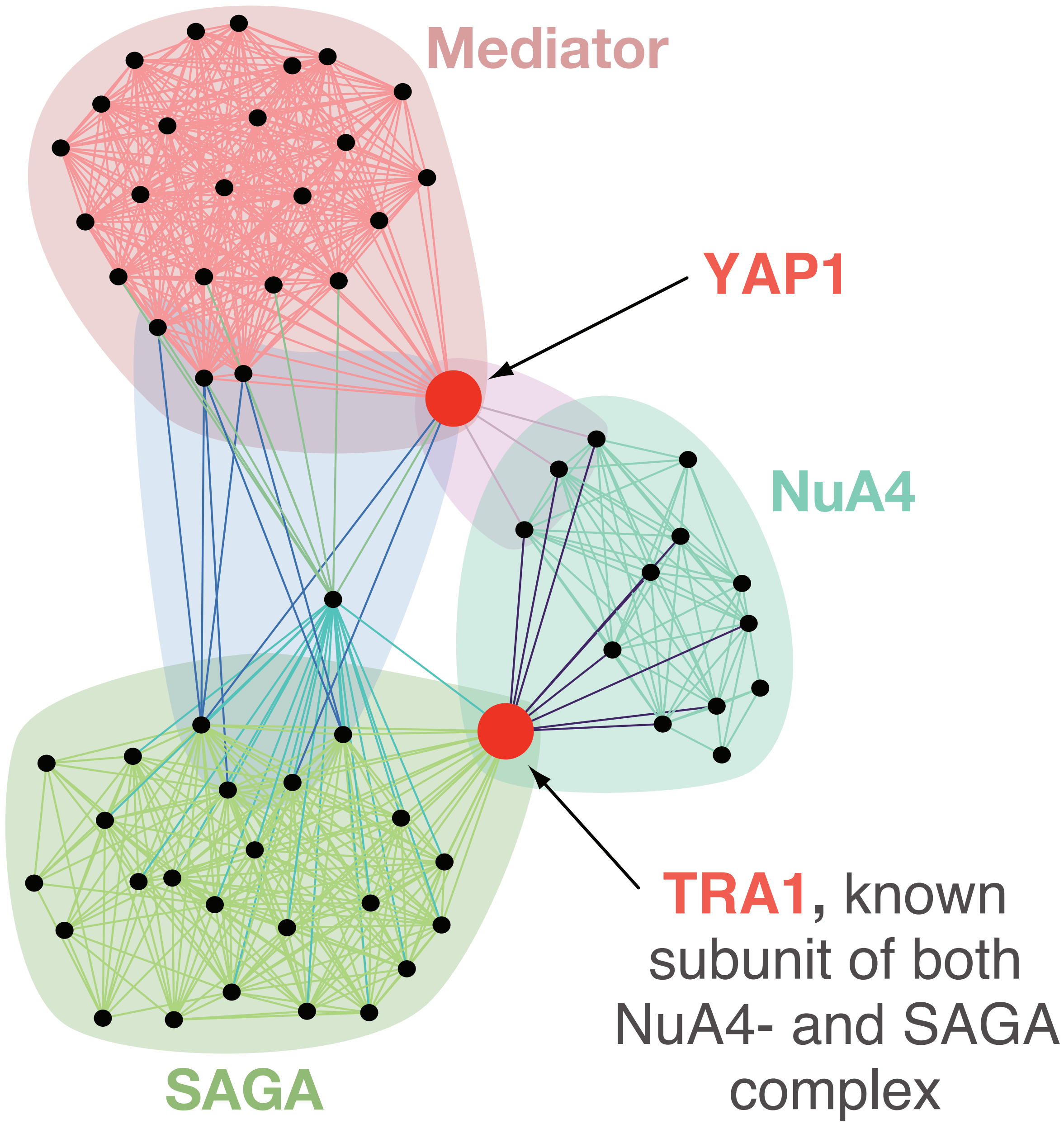




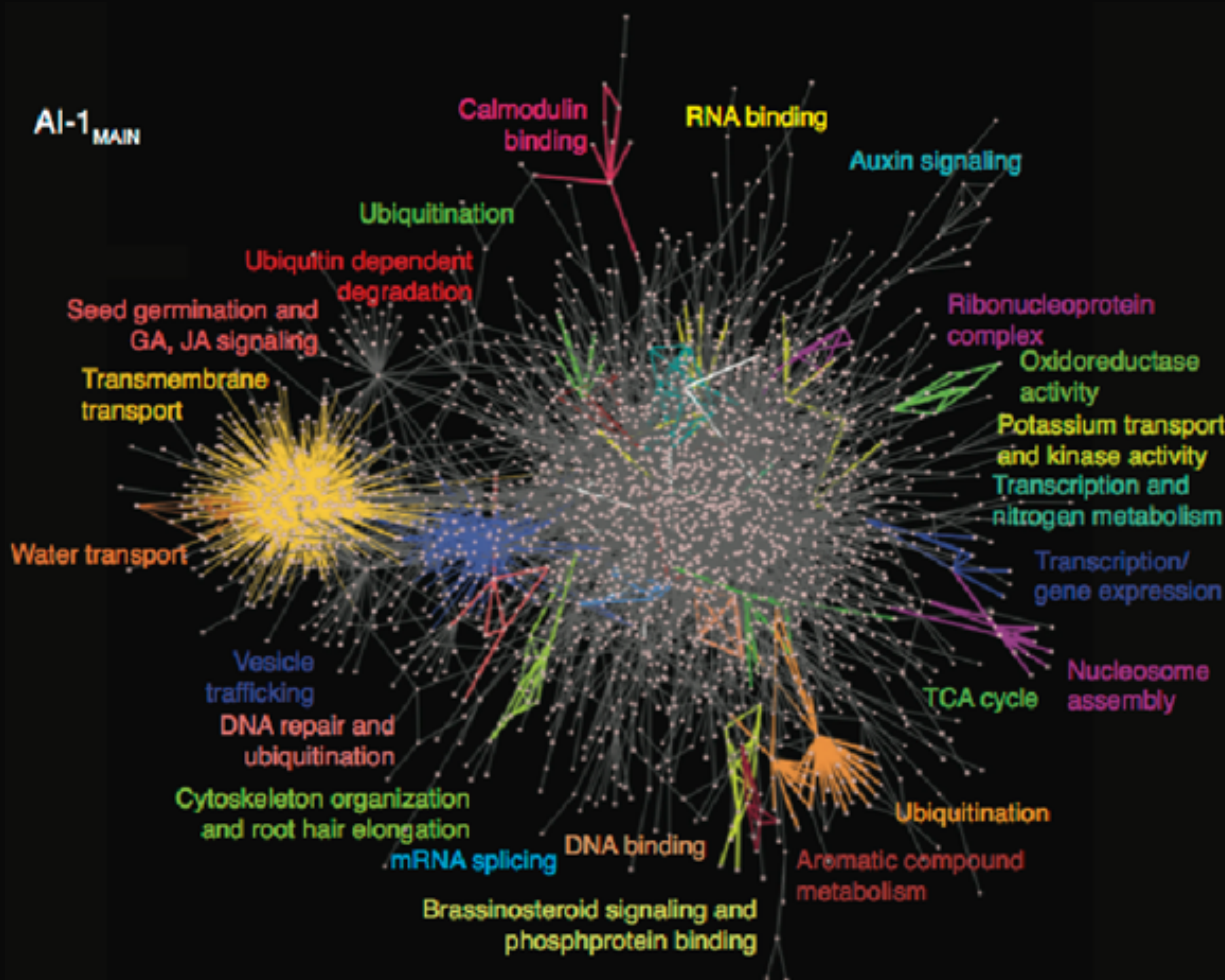

BROOM
SWEEP
PAINT
PAINTER
PAINTING
GROOM
HAIR
BRUSH
COMB
HAIRSPRAY
TOOTHBRUSH
TOOTHPASTE



Experiment, Science
Smart, Intellect, Scientists
Science, Scientists
Clever, Wit
Newton, Gravity, Apple
FLASK
CHEMICAL
BEAKER
BIOLOGIST
TESTTUBE
LAB
BIOLOGY
INVENT
EXCEPTIONAL
BRIGHT
RESEARCH
CHEMIST
INVENTOR
BRILLIANT
CHEMISTRY
EXPERIMENT
INTELLECT
GENIUS
GIFTED
SCIENTIST
KINETIC
SCIENCE
INTELLIGENT
VELOCITY
SCIENTIFIC
INTELLIGENCE
PHYSICS
SMART
RETARDED
HYPOTHESIS
EINSTEIN
WISDOM
THEORY
WISE
THEOREM
CUNNING
RELATIVITY
NEWTON
CLEVER
INERTIA
WEIGHT
OUTFOX
WIT
LAW
SLY
APPLE
GRAVITY

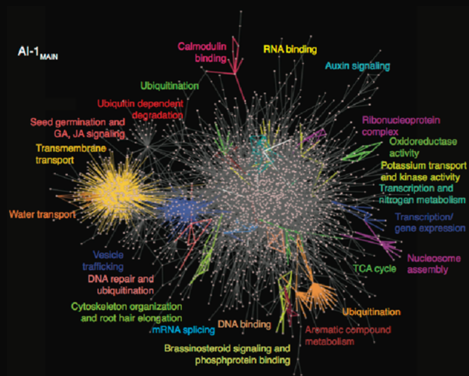

Mediator
YAP1
NuA4
SAGA
TRA1, known subunit of both NuA4- and SAGA complex

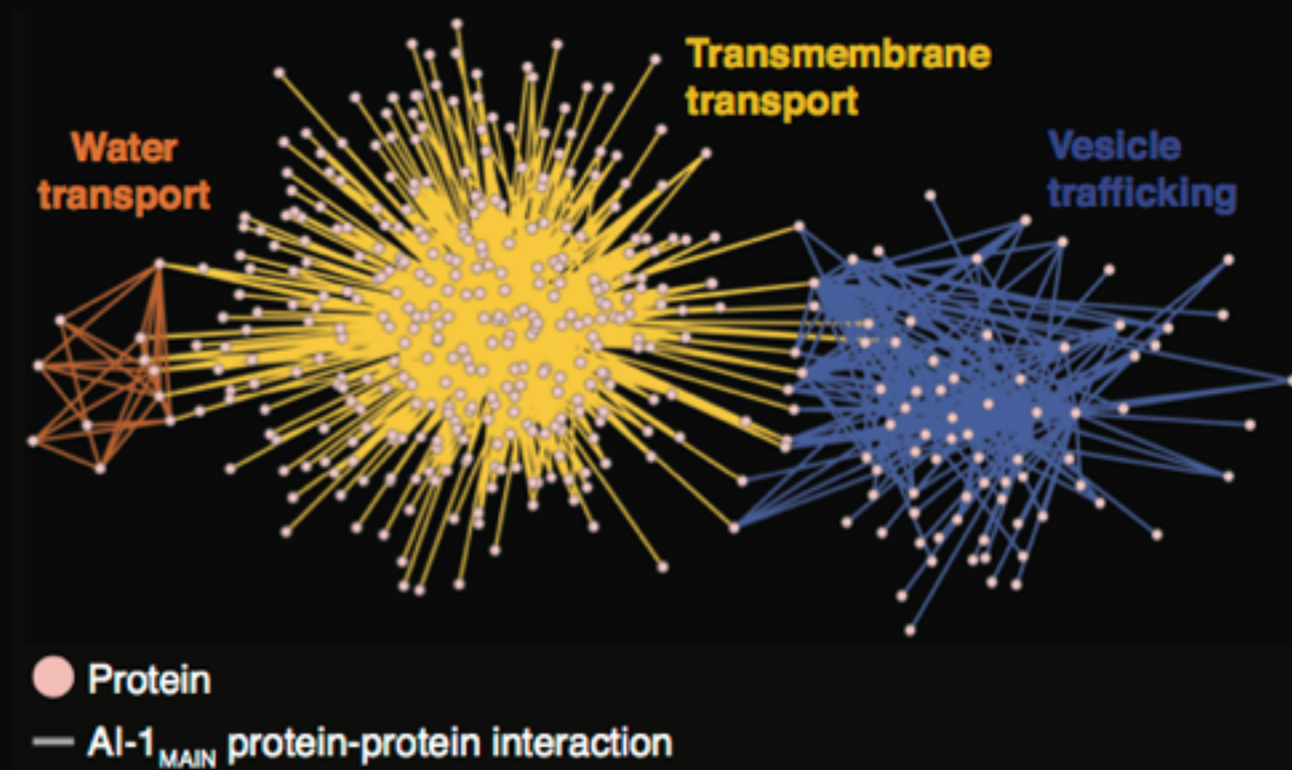
The first plant (genomic scale) interactome



Arabidopsis Interactome Mapping Consortium, *Science, 2011*

The first plant (genomic scale) interactome





Arabidopsis Interactome Mapping Consortium, *Science, 2011*

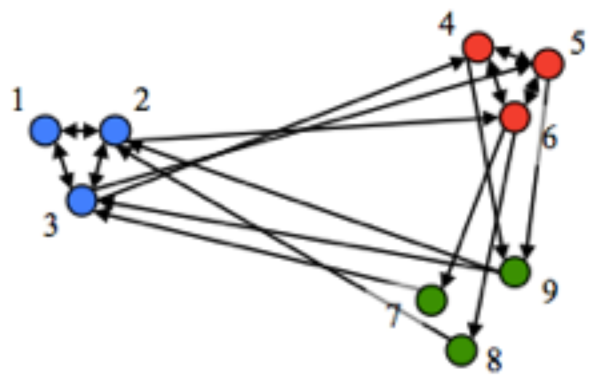
Statistical inference

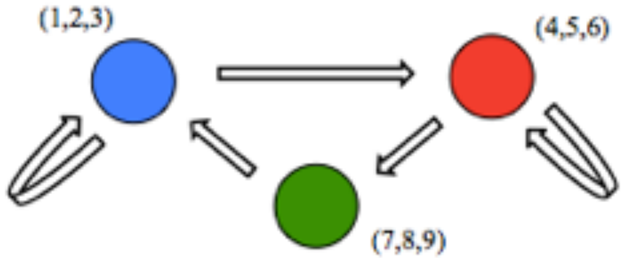
Given a graph G, and a generative model with parameters θ

Likelihood

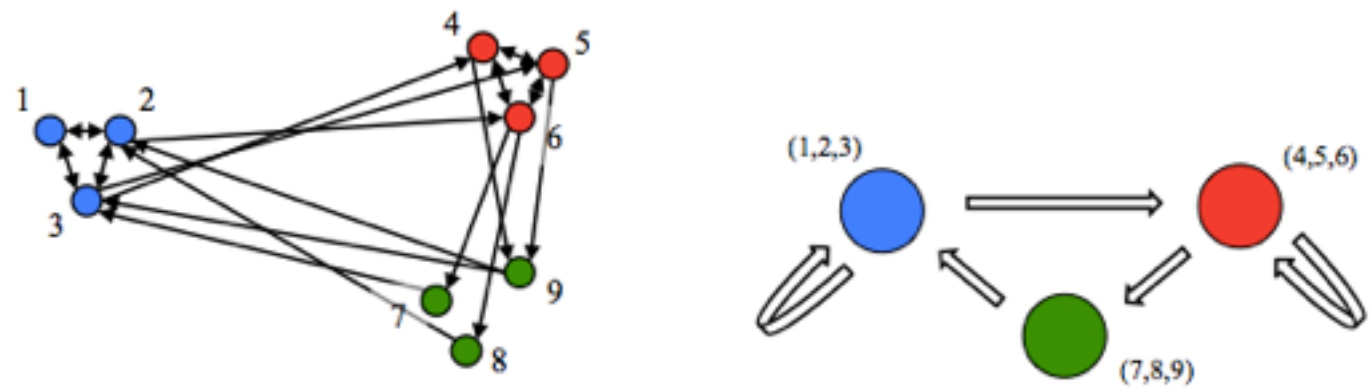
$$P(\theta|G) = \frac{P(G|\theta)P(\theta)}{P(G)}$$

Stochastic Block Model



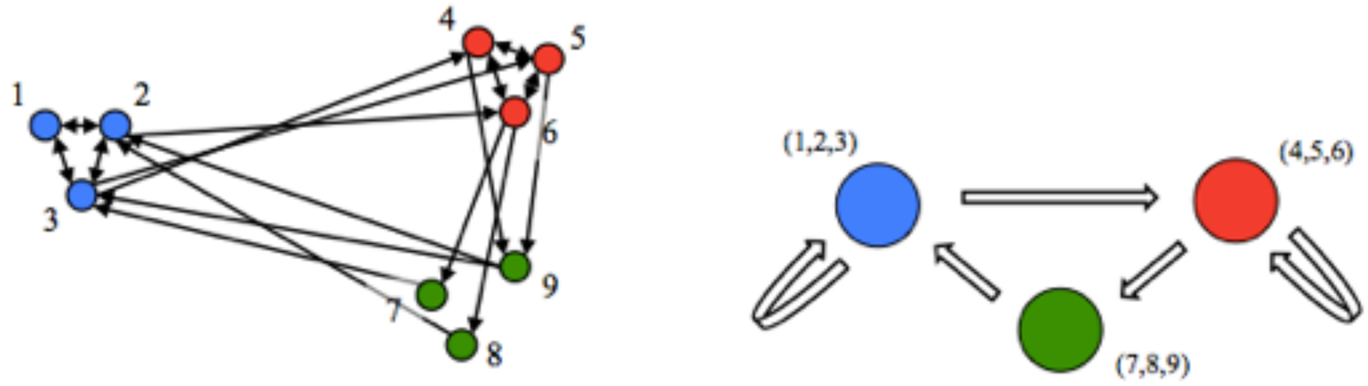


Stochastic Block Model



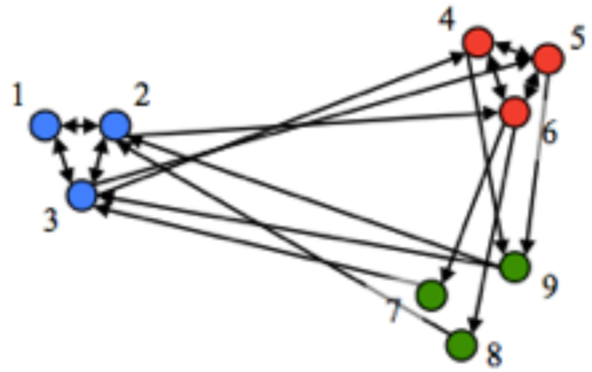
$$c_i$$

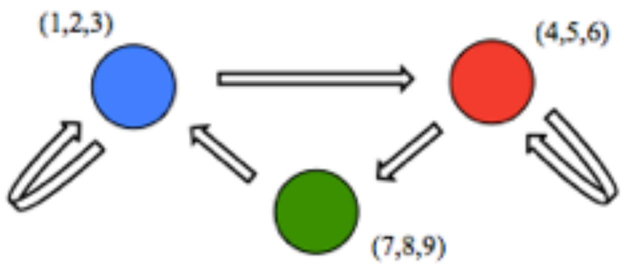
Stochastic Block Model



$$c_i \quad p_{c_i c_j}$$

Stochastic Block Model





$$c_i \quad p_{c_i c_j}$$

$$\prod_{i<j} p_{c_i c_j}^{A_{ij}} (1 - p_{c_i c_j})^{1-A_{ij}}$$

So, what should I use?

1. No silver bullet.

2. Hard to know beforehand.

Accuracy

Accuracy

- It is very hard to compare performance of methods on a fair ground because each method is usually very good at finding what it is looking for.

Accuracy

- It is very hard to compare performance of methods on a fair ground because each method is usually very good at finding what it is looking for.
- Most studies use ‘**benchmark networks**’ to evaluate the performance.

Accuracy

- It is very hard to compare performance of methods on a fair ground because each method is usually very good at finding what it is looking for.
- Most studies use '**benchmark networks**' to evaluate the performance.
- **Infomap** and **Louvain** method are the best in these benchmarks.

Accuracy

- It is very hard to compare performance of methods on a fair ground because each method is usually very good at finding what it is looking for.
- Most studies use ‘**benchmark networks**’ to evaluate the performance.
- **Infomap** and **Louvain** method are the best in these benchmarks.
- However, the performance depends on what kinds of community structure the benchmark networks assume.

Accuracy

- It is very hard to compare performance of methods on a fair ground because each method is usually very good at finding what it is looking for.
- Most studies use '**benchmark networks**' to evaluate the performance.
- **Infomap** and **Louvain** method are the best in these benchmarks.
- However, the performance depends on what kinds of community structure the benchmark networks assume.
- Good performance in the benchmarks *does not guarantee* good performance in real cases.

Computational complexity

Computational complexity

- Some methods are much faster than others.

Computational complexity

- Some methods are much faster than others.
- $O(\exp(n))$ vs. $O(m^2 n)$ vs. $O(n \log n)$

Computational complexity

- Some methods are much faster than others.
- $O(\exp(n))$ vs. $O(m^2 n)$ vs. $O(n \log n)$
- Usually a good choice for huge (> 1m ~ 1b) networks: **Louvain method** (~$O(n \log n)$)

Computational complexity

- Some methods are much faster than others.
- *O(exp(n))* vs. *O(m^2 n)* vs. *O(n log n)*
- Usually a good choice for huge (> 1m ~ 1b) networks: **Louvain method** (~*O(n log n)*)
- Current version of **infomap** also uses louvain-type multilevel optimization and very fast.

Computational complexity

- Some methods are much faster than others.
- *O(exp(n))* vs. *O(m^2 n)* vs. *O(n log n)*
- Usually a good choice for huge (> 1m ~ 1b) networks: **Louvain method** (~*O(n log n)*)
- Current version of **infomap** also uses louvain-type multilevel optimization and very fast.
- **Link clustering** can also handle large graphs (but it becomes slow with large hubs).

Overlap

Overlap

- If you expect pervasive overlap of communities, you should use overlapping community detection methods.

Overlap

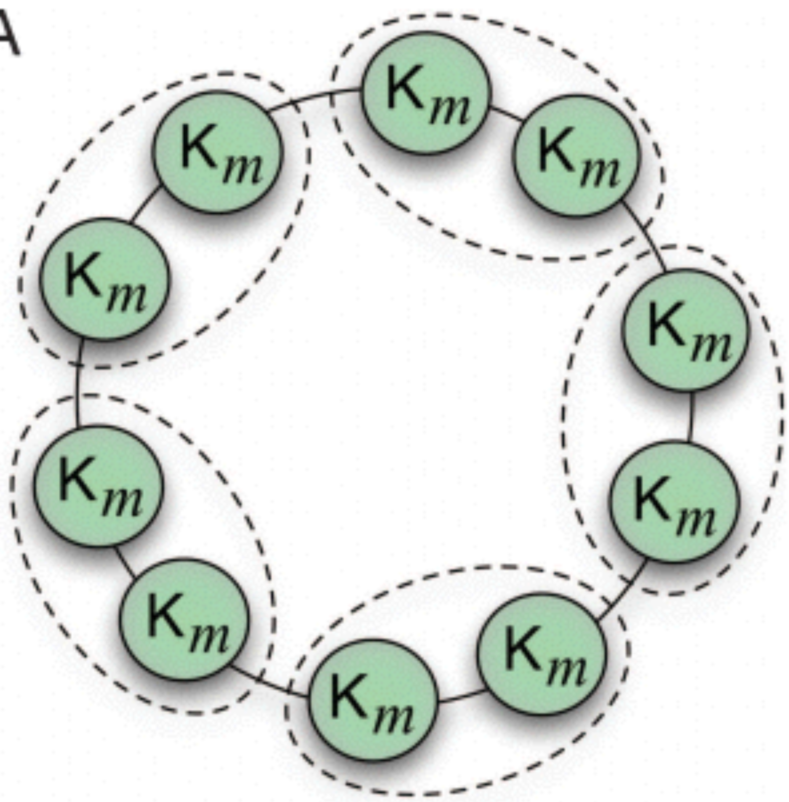
- If you expect pervasive overlap of communities, you should use overlapping community detection methods.
- **Link clustering** and **clique percolation** methods are common choices.

Overlap

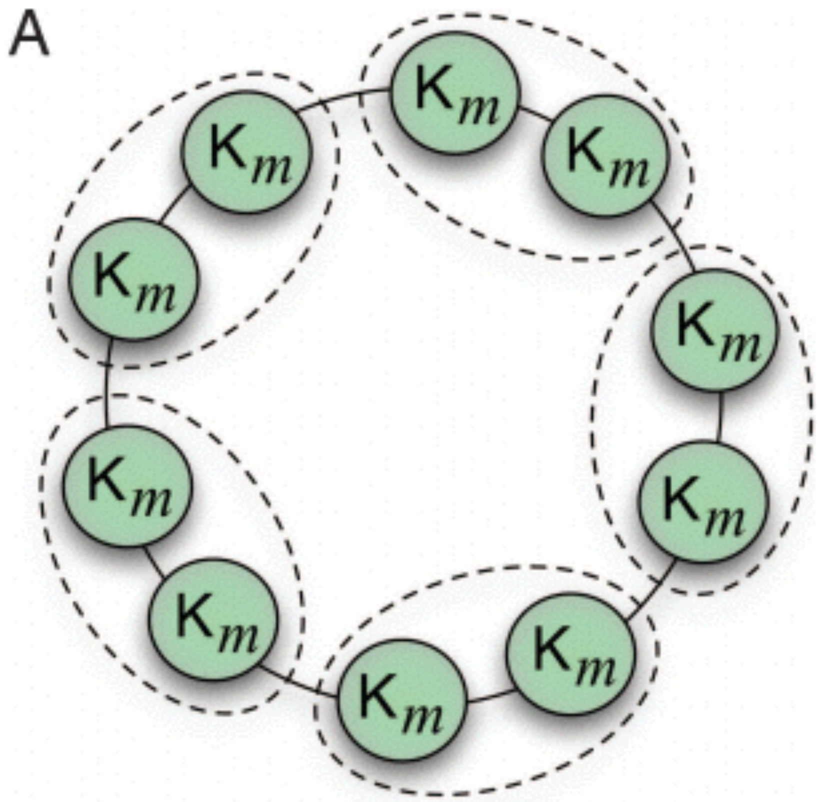
- If you expect pervasive overlap of communities, you should use overlapping community detection methods.
- **Link clustering** and **clique percolation** methods are common choices.
- These methods can detect highly overlapping communities. There are many other methods but most methods only deal with ‘**fuzzy**’ overlaps.

Resolution limit

A

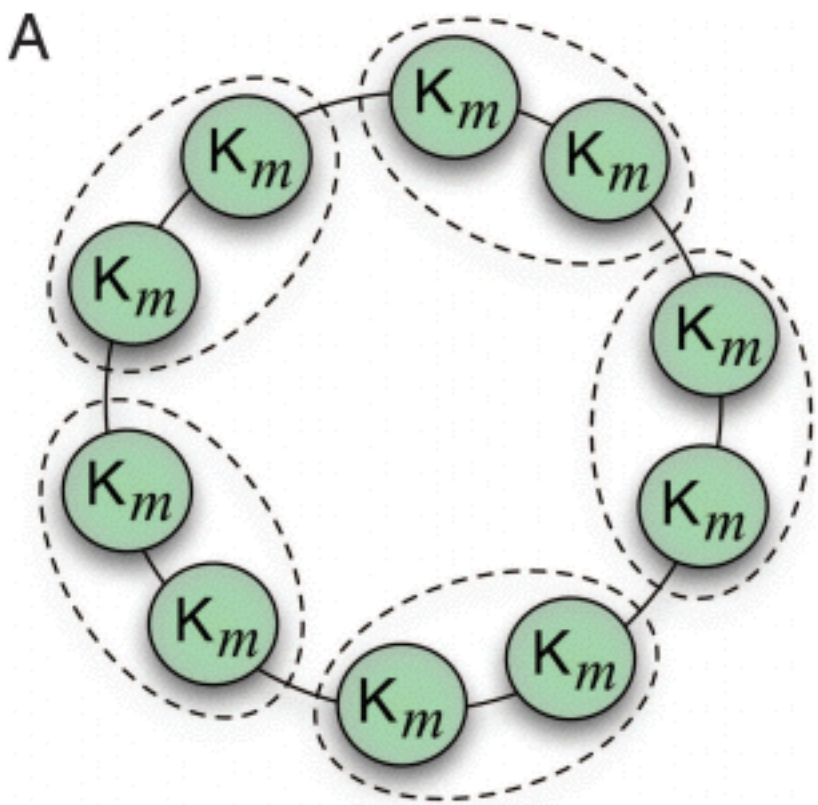


Resolution limit



- Modularity has a resolution limit that depends on the system size.

Resolution limit



- Modularity has a resolution limit that depends on the system size.
- If a community is smaller than this limit, modularity-based optimization cannot find the communities, even though they are cliques.

My heuristic

My heuristic

- I don’t care too much and I just want to get rough clusters in my network — **Infomap** or **Louvain**

My heuristic

- I don’t care too much and I just want to get rough clusters in my network — **Infomap** or **Louvain**
- I expect multiple community membership for many nodes — **Link clustering** (**Clique Percolation**)

My heuristic

- I don’t care too much and I just want to get rough clusters in my network — **Infomap** or **Louvain**
- I expect multiple community membership for many nodes — **Link clustering** (**Clique Percolation**)
- My network is HUGE and doesn’t have super-large hubs — **Louvain** (**Infomap, link clustering**)

My heuristic

- I don’t care too much and I just want to get rough clusters in my network — **Infomap** or **Louvain**
- I expect multiple community membership for many nodes — **Link clustering** (**Clique Percolation**)
- My network is HUGE and doesn’t have super-large hubs — **Louvain** (**Infomap, link clustering**)
- My network is HUGE and has lots of super-hubs — **Louvain** (**Infomap**)

My heuristic

- I don’t care too much and I just want to get rough clusters in my network — **Infomap** or **Louvain**
- I expect multiple community membership for many nodes — **Link clustering** (**Clique Percolation**)
- My network is HUGE and doesn’t have super-large hubs — **Louvain** (**Infomap, link clustering**)
- My network is HUGE and has lots of super-hubs — **Louvain** (**Infomap**)
- I’d like to see the detailed hierarchical structure — **Link clustering**

THANK YOU!

@yy

yyahn@indiana.edu